National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics
Speech and Language Processing Group

Diploma Thesis

Discovering Approaches for Social Video Question Answering using Deep Learning

CHRISTINA SARTZETAKI

Supervisor: Alexandros Potamianos
Associate Professor, NTUA

Athens, June 2022

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics
Speech and Language Processing Group

Diploma Thesis

Discovering Approaches for Social Video Question Answering using Deep Learning

CHRISTINA SARTZETAKI

Supervisor: Alexandros Potamianos
Associate Professor, NTUA

Approved by the examination committee on 14th June 2022.

(Signature)	(Signature)	(Signature)
....	....	....
Alexandros Potamianos	Constantinos Tzafestas	Athanasios Katsamanis
Associate Professor, NTUA	Associate Professor, NTUA	Principal Researcher, RC Athena

Athens, June 2022

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics
Speech and Language Processing Group

(Signature)

....
Christina Sartzetaki

Electrical & Computer Engineer, NTUA

The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the National Technical University of Athens.

To my mother,
Maria

Ευχαριστίες

Με την περάτωση αυτής της Διπλωματικής Εργασίας ολοκληρώνονται ταυτόχρονα οι προπτυχιακές σπουδές μου στο ΕΜΠ, το πρώτο μου ακαδημαϊκό επίτευγμα, καθώς και η εισαγωγή μου στον χώρο της Μηχανικής Μάθησης.

Θα ήθελα να ευχαριστήσω πρωτίστως τον Καθηγητή μου κ. Αλέξανδρο Ποταμιάνο, αρχικά για τις ενδιαφέρουσες διαλέξεις και την συμπερίληψη σύγχρονων ερευνητικών μεθόδων στα μαθήματά του, που με ενέπνευσαν να ασχοληθώ με το αντικείμενο, και εν συνεχεία για την διαρκή ενασχόλησή του με την ερευνητική μου προσπάθεια, καθώς και για τη μετάδοση πολύπλευρων γνώσεων μέσα από ιδέες, συζητήσεις, αλλά και κριτική κατά την επίβλεψη αυτής της εργασίας.

Ένα μεγάλο ευχαριστώ οφείλω στον Γιώργο Παρασκευόπουλο για την πολύτιμη βοήθεια και καθοδήγηση, αλλά και τις καθοριστικές ιδέες του καθ'όλη τη διάρκεια εκπόνησης αυτής της Διπλωματικής. Είμαι επιπλέον ευγνώμων για τους φίλους και συμφοιτητές μου, με τους οποίους ανταλλάσσαμε συνεχώς σκέψεις, ανησυχίες αλλά και υποστήριξη όλο αυτό το διάστημα.

Τέλος, θα ήθελα να ευχαριστήσω από τα βάθη της καρδιάς μου τους γονείς μου, Μαρία και Στέλιο, για όλα.

Χριστίνα Σαρτζετάκη

Αθήνα, Ιούνιος 2022

Περίληψη

Οι άνθρωποι είμαστε κοινωνικά πλάσματα, και η επιβίωση και ευημερία μας εξαρτάται από την αποτελεσματική επικοινωνία μας με τους άλλους. Αυτή επιτυγχάνεται μέσω της κατανόησης πληροφοριών από πολλαπλές αισθητηριακές πηγές καθώς και με τη χρήση λογικής για την εξαγωγή συμπερασμάτων, προκειμένου να ανταποκριθούμε ανάλογα. Το Social Video Question Answering είναι μια εφαρμογή Μηχανικής Μάθησης για τον έλεγχο των ικανοτήτων κοινωνικής συλλογιστικής ενός πράκτορα τεχνητής νοημοσύνης, που βασίζεται στο κατά πόσο μπορεί να απαντήσει σε ερωτήσεις πάνω σε ένα δεδομένο βίντεο. Μπορεί να απαιτεί περίπλοκους συνδυασμούς αναγνώρισης συναισθημάτων, γλωσσικής κατανόησης, και λογικής και συλλογιστικής σκέψης.

Σε αυτή τη Διπλωματική Εργασία, εστιάζουμε στον εντοπισμό διαφορετικών προσεγγίσεων για Social Video Question Answering με χρήση Βαθειάς Μάθησης, μέσα από την αξιοποίηση προηγούμενων εργασιών σε διαφορετικούς τομείς όπως η Όραση Υπολογιστών και η Επεξεργασία Φυσικής Γλώσσας. Κατά τη διάρκεια της έρευνάς μας ακολουθήσαμε δύο διαφορετικές προσεγγίσεις.

Στο πρώτο μέρος της εργασίας μας, αξιοποιούμε τις δυνατότητες συλλογιστικής πολλαπλών βημάτων του Compositional Attention Networks (MAC) και προτείνουμε μια πολυτροπική επέκταση (MAC-X). Το MAC-X βασίζεται σε ένα αναδρομικό κελί που εκτελεί επαναληπτική συγχώνευση μεσαίου επιπέδου τροπικοτήτων εισόδου (οπτική, ακουστική, κείμενο) σε πολλαπλά στάδια συλλογισμού, χρησιμοποιώντας έναν μηχανισμό χρονικής προσοχής. Στη συνέχεια συνδυάζουμε το MAC-X με LSTM για επεξεργασία χρονικής εισόδου σε μια αρχιτεκτονική από άκρο σε άκρο. Οι συγκριτικές μελέτες μας δείχνουν ότι η προτεινόμενη αρχιτεκτονική MAC-X μπορεί να αξιοποιήσει αποτελεσματικά τα πολυτροπικά στοιχεία εισόδου χρησιμοποιώντας μηχανισμούς συγχώνευσης μεσαίου επιπέδου. Εφαρμόζουμε το MAC-X στο σύνολο δεδομένων Social IQ και επιτυγχάνουμε απόλυτη βελτίωση 2,5% όσον αφορά τη δυαδική ακρίβεια σε σχέση με την τρέχουσα κατάσταση αιχμής.

Στο δεύτερο μέρος της εργασίας μας, ακολουθούμε την κατεύθυνση της απάντησης ερωτήσεων μέσα από περιγραφές βίντεο, που λαμβάνουμε μέσω της ενίσχυσης των διαλόγων με πληροφορίες από την ανίχνευση κοινωνικών ενδείξεων, συγκεκριμένα συναισθηματικές πληροφορίες για το βλέμμα. Αυτή είναι η πρώτη φορά, εξ όσων γνωρίζουμε, που προτείνεται ένα τέτοιο σύστημα εξαγωγής χαρακτηριστικών ειδικά σχεδιασμένο για κοινωνικά βίντεο. Πειραματιζόμαστε με διαφορετικές μεθόδους δημιουργίας περιγραφής φυσικής γλώσσας από μια ενδιάμεση δομή γραφήματος και παρέχουμε συγκριτικές μελέτες για διαφορετικά μοντέλα τύπου BERT και επίπεδα εκπαίδευσης. Εφαρμόζουμε τη μέθοδό μας στο σύνολο δεδομένων Social IQ και επιτυγχάνουμε σημαντικές βελτιώσεις σε σχέση με τη βασική απόδοση.

Λέξεις Κλειδιά

Βαθειά Μάθηση, Αυτόματη απάντηση ερωτήσεων σε βίντεο, Κοινωνική συλλογιστική, Αυτόματη απάντηση ερωτήσεων κοινωνικού περιεχομένου σε βίντεο, Compositional Attention Networks, MAC, Κοινωνικές ενδείξεις, Ανίχνευση βλέμματος, Ανίχνευση συναισθήματος, Επεξεργασία Φυσικής Γλώσσας, Transformers, BERT.

Abstract

Humans are social creatures; our survival and well-being depends on our effective communication with others. This is achieved through perceiving and understanding information from multiple sensory modalities as well as reasoning and arriving to conclusions, in order to respond accordingly. Social Video Question Answering is a Machine Learning task to test the social reasoning abilities of an AI agent, based on how accurately it can answer questions on a given video. It can require sophisticated combinations of emotion recognition, language understanding, cultural knowledge, logical and causal reasoning, on top of non-social layers of comprehension about physical events.

In this Diploma Thesis, we focus on discovering different approaches for Social Video Question Answering that leverage Deep Learning methods, through building on previous work in different fields such as Computer Vision and Natural Language Processing. We take two distinct approaches in the course of our research.

In the first part of our work, we propose a novel deep architecture for the task of reasoning about social interactions in videos. We leverage the multi-step reasoning capabilities of Compositional Attention Networks (MAC) [1], and propose a multimodal extension (MAC-X). MAC-X is based on a recurrent cell that performs iterative mid-level fusion of input modalities (visual, auditory, text) over multiple reasoning steps, by use of a temporal attention mechanism. We then combine MAC-X with LSTMs for temporal input processing in an end-to-end architecture. Our ablation studies show that the proposed MAC-X architecture can effectively leverage multimodal input cues using mid-level fusion mechanisms. We apply MAC-X to the task of Social Video Question Answering in the Social IQ dataset [2] and obtain a 2.5% absolute improvement in terms of binary accuracy over the current state-of-the-art.

In the second part of our work, we follow the direction of question answering on video captioning, which we obtain through augmentation of the dialogue transcripts with explicit social cues detection information, namely emotional eye-gaze information. This is the first time, to the best of our knowledge, that a feature extraction pipeline specifically designed for social video is proposed, standing in as a general framework for leveraging social information in video. We experiment with different natural language caption generation methods from an intermediate graph structure, and provide ablation studies for several BERT [3]-like language models and fine-tuning levels, as well as a hierarchical summary scheme based on question conditioning via extractive question answering. We apply our method to the Social IQ dataset [2] and obtain significant improvements over the baselines.

Keywords

Deep Learning, Video Question Answering, Social reasoning, Social Video Question Answering, Compositional Attention Networks, MAC, Social cues, Eye-gaze detection, Emotion recognition, Natural Language Processing, Transformers, BERT

Table of Contents

List of Figures

List of Tables

Εκτεταμένη Περίληψη

Εισαγωγή

Καθημερινά, τα ερεθίσματα που δεχόμαστε τόσο από τον φυσικό όσο και από τον ψηφιακό κόσμο είναι σε πολλαπλές μορφές (τροπικότητες) και μεγάλες ποσότητες. Είναι σημαντικό να αναπτύσσονται πράκτορες τεχνητής νοημοσύνης με σκοπό να βοηθούν τους ανθρώπους στην καθημερινή τους αλληλεπίδραση με αυτόν τον τεράστιο όγκο πολυτροπικών δεδομένων, μετριάζοντας τις ατέλειές τους, εμπλουτίζοντας τις εμπειρίες τους και ταυτόχρονα προστατεύοντάς τους. Για να δώσουμε ένα συγκεκριμένο παράδειγμα εφαρμογών στον πραγματικό κόσμο για καθέναν από αυτούς τους τρεις κινητήριους παράγοντες, μπορούμε να εξετάσουμε τα ακόλουθα. Πρώτον, συστήματα τεχνητής νοημοσύνης που είναι εκπαιδευμένα να κατανοούν το οπτικό περιβάλλον ή τις κοινωνικές αλληλεπιδράσεις γύρω τους μπορούν να παρέχουν ουσιαστική βοήθεια σε άτομα με προβλήματα όπως τύφλωση ή αυτισμό, παρέχοντάς τους πληροφορίες για το περιβάλλον τους που δεν μπορούν να αποκτήσουν μόνοι τους. Δεύτερον, από τις έξυπνες συστάσεις και την ανάκτηση πολυμεσικού περιεχομένου, έως τα συστήματα επαυξημένης πραγματικότητας για ψυχαγωγικούς και εκπαιδευτικούς σκοπούς, έχει δοθεί μεγάλη έμφαση τόσο από την ερευνητική κοινότητα όσο και από τη βιομηχανία στον εμπλουτισμό της καθημερινής ζωής. Τέλος, ένα σημαντικό πρόβλημα είναι ο εντοπισμός ρητορικής μίσους στα κοινωνικά δίκτυα [13], καθώς συχνά εκεί άνθρωποι στοχοποιούνται με βάση τη φυλή, το φύλο ή τον σεξουαλικό τους προσανατολισμό, με αποτέλεσμα να υποφέρουν από σοβαρά προβλήματα ψυχικής υγείας, αλλά και να στοχοποιούνται σε εγκλήματα μίσους στον πραγματικό κόσμο.

Η όραση και η γλώσσα είναι δύο από τις πιο θεμελιώδεις και αξιοσημείωτες ικανότητες του ανθρώπινου νου, αφού η όραση επιτρέπει τη δημιουργία νοητικών εννοιών που διαφορετικά δεν θα υπήρχαν, όπως το χρώμα και το φως, και η γλώσσα έχει τη δύναμη να μετατρέψει όλη την αισθητηριακή εμπειρία σε αυτές τις στοιχειώδεις νοητικές έννοιες και να τις συνδυάσει για να δημιουργήσει σύνθετες νέες ιδέες, διευκολύνοντας έτσι τη σκέψη κάνοντας «απεριόριστη χρήση πεπερασμένων μέσων» [14]. Η αλληλεπίδραση της γλώσσας με την όραση παρακινεί τους ερευνητές να εντοπίσουν τις σχέσεις ανάμεσα σε τροπικότητες, να τις συνδυάσουν και να τις αιτιολογήσουν με σκοπό την λήψη αποφάσεων. Μία σημαντική εφαρμογή στη μηχανική εκμάθηση όρασης-γλώσσας είναι αυτή του Visual Question Answering (VQA), που απαιτεί από τον πράκτορα AI να απαντήσει μια ερώτηση φυσικής γλώσσας με βάση μια εικόνα. Μια σημαντική πρόκληση είναι ότι λόγω στατιστικών της ιδιοτήτων η γλώσσα μπορεί να αποτελεί παράγοντα μεροληψίας και να είναι ένα πιο εύκολο σήμα για μάθηση από την εικόνα, με αποτέλεσμα μοντέλα όρασης-γλώσσας να αγνοούν εντελώς τις οπτικές πληροφορίες προς όφελος της εκμετάλλευσης της γλώσσας.

Οι άνθρωποι είμαστε κοινωνικά πλάσματα. Η επιβίωση και η ευημερία μας εξαρτάται από την αποτελεσματική επικοινωνία μας με τους άλλους. Αυτή επιτυγχάνεται μέσω της αντίληψης και της κατανόησης πληροφοριών από πολλαπλές αισθητηριακές πηγές καθώς και με τη χρήση λογικής για την εξαγωγή συμπερασμάτων, προκειμένου να ανταποκριθούμε ανάλογα. Πιο συγκεκριμένα, βασιζόμαστε στην ικανότητα κατανόησης της ψυχικής κατάστασης άλλων ανθρώπων (που περιλαμβάνει προθέσεις, κίνητρα, συναισθήματα), μέσω της επεξεργασίας πληροφοριών όπως το βλέμμα τους, η έκφραση του προσώπου, η γλώσσα του σώματος (στάση, χειρονομίες) και ο τόνος της φωνής τους. Μερικοί άνθρωποι, αν και πολύ έξυπνοι, όπως άτομα με διαταραχή αυτιστικού φάσματος (ΔΑΦ), δεν μπορούν να διακρίνουν αυτές τις ενδείξεις, καθώς λειτουργούν κυρίως με βάση λογικές και αντικειμενικές πληροφορίες, όπως συμβαίνει με τις περισσότερες εφαρμογές μηχανικής μάθησης. Η δημιουργία μιας μεθόδου απάντησης ερωτήσεων για αυτό το θέμα μπορεί να χρησιμεύσει τόσο ως τρόπος εκπαίδευσης ατόμων με ΔΑΦ ώστε να αναγνωρίζουν τέτοιες συμπεριφορές, καθώς και για μοντέλα μηχανικής μάθησης ως εφαρμογή απάντησης ερωτήσεων. Η Αυτόματη Απάντηση Ερωτήσεων Κοινωνικού Περιεχομένου σε Βίντεο (Social Video Question Answering) είναι μια εφαρμογή Μηχανικής Μάθησης για τον έλεγχο των ικανοτήτων κοινωνικής συλλογιστικής ενός πράκτορα τεχνητής νοημοσύνης (AI agent), που βασίζεται στο κατά πόσο μπορεί να απαντήσει σε ερωτήσεις πάνω σε ένα δεδομένο βίντεο. Μπορεί να απαιτεί περίπλοκους συνδυασμούς αναγνώρισης συναισθημάτων, γλωσσικής κατανόησης, λογικής και συλλογιστικής σκέψης, πέρα από μη κοινωνικά επίπεδα κατανόησης για φυσικά γεγονότα.

Έχοντας ουσιαστικά ένα πρόβλημα συλλογιστικής (reasoning), αντλούμε έμπνευση από ένα τμήμα της βιβλιογραφίας VQA που ονομάζεται νευροσυμβολικά μοντέλα και εστιάζει στην κατασκευή νευρωνικών μοντέλων ενώ ταυτόχρονα επιτρέπει σαφή συμβολικό συλλογισμό υψηλού επιπέδου. Από αυτά εστιάζουμε σε προσεγγίσεις πιο κοντά στη νευρωνική παρά στη συμβολική πλευρά, όπως το Learning from abstraction [15] και το Memory Attention Composition (MAC) Network [1]. Το Δίκτυο MAC επιχειρεί να συλλάβει τη λογική της σκέψης εκτός από την κατασκευή νευρωνικών αναπαραστάσεων από τα δεδομένα και δημιουργήθηκε για εφαρμογές που απαιτούν σκόπιμη συλλογιστική από γεγονότα σε συμπεράσματα λόγω του δομημένου και επαναληπτικού συστήματος συλλογισμού του.

Στο πρώτο μέρος αυτής της εργασίας, προτείνουμε μια προσέγγιση από άκρο σε άκρο που βασίζεται σε μια πολυτροπική επέκταση του MAC Network για την εφαρμογή του στο Social Video Question Answering, που ονομάζεται MAC-Extend (MAC-X). Αξιοποιούμε τις συλλογιστικές δυνατότητες του MAC και βασίζουμε το MAC-X σε ένα επαναλαμβανόμενο κελί που εκτελεί επαναληπτική συγχώνευση (fusion) μεσαίου επιπέδου τροπικοτήτων εισόδου (οπτική, ακουστική, κείμενο) σε πολλαπλά στάδια συλλογισμού, χρησιμοποιώντας έναν μηχανισμό χρονικής προσοχής. Στη συνέχεια συνδυάζουμε το MAC-X με τα LSTM για επεξεργασία χρονικής εισόδου σε μια αρχιτεκτονική από άκρο σε άκρο.

Ωστόσο, αυτή η προσέγγιση, όπως και οι περισσότερες εργασίες στη βιβλιογραφία, θεωρούν το βίντεο ως μια συνεχή πηγή που περιέχει εξίσου σημαντικά χαρακτηριστικά που πρέπει να εξεταστούν. Αν και αυτή η γενική προσέγγιση έχει νόημα στις περισσότερες εφαρμογές VQA όπου οι ερωτήσεις αναφέρονται στο περιβάλλον, τα αντικείμενα ή τις ενέργειες και τα γεγονότα, το Social VQA αφορά τους ανθρώπους και τις αλληλεπιδράσεις τους μέσω των οποίων ανταλλάσσουν τόσο λεκτικές όσο και μη λεκτικές πληροφορίες.

Στο δεύτερο μέρος αυτής της εργασίας, προτείνουμε μια προσέγγιση επαύξησης μέσω άμεσης ανίχνευσης κοινωνικών ενδείξεων από το βίντεο και σύνδεσής τους με τα άτομα που συμμετέχουν στην κοινωνική αλληλεπίδραση, η οποία μοντελοποιείται με χρήση του βλέμματος των ματιών για να σχηματίσει γραφήματα βλέμματος για κάθε σκηνή. Μέσα από αυτό, θέλουμε επίσης να διερευνήσουμε την υπόθεση ότι το βλέμμα μπορεί να συνοψίσει το κοινωνικό βίντεο. Για να επιλέξουμε πώς να επεξεργαστούμε αυτά τα γραφήματα βλέμματος με στόχο να απαντήσουμε σε ερωτήσεις φυσικής γλώσσας που απαιτούν κοινωνικό συλλογισμό, κινητοποιούμαστε από την ακόλουθη εργασία. Στο [16], οι συγγραφείς υποστηρίζουν ότι το NLP χρειάζεται κοινωνικό πλαίσιο για να επιτύχει πραγματικά, θεωρώντας το ως το τελευταίο στάδιο για να επιτευχθεί μια πραγματικά πλήρης κατανόηση του κόσμου μέσω της γλώσσας. Πιο συγκεκριμένα, η πρόοδος του NLP ορίζεται από την κατάκτηση διαφορετικών World Scopes (WS), το καθένα πιο γενικό από το προηγούμενο, ταξινομημένα ως Corpus, Internet, Perception (multimodal), Embodiment, και Social.

Αντλώντας έμπνευση από αυτήν την ανάλυση, πήραμε την κατεύθυνση εκπαίδευσης ενός μοντέλου καθαρά NLP μέσω της μετάφρασης πολυτροπικής εισόδου (WS3) που περιέχει κοινωνικές ενδείξεις (WS5) στη γλώσσα, δημιουργώντας επεξηγηματικά κείμενα βίντεο (video captions). Ένα πρόσθετο πλεονέκτημα της προσέγγισης σε σύγκριση με τα μοντέλα από άκρο σε άκρο είναι η πρόσθετη δυνατότητα επεξήγησης τόσο σε περιπτώσεις επιτυχίας όσο και σε περιπτώσεις αποτυχίας, καθώς τα ενδιάμεσα αποτελέσματα (video captions) μπορούν να μας βοηθήσουν να αποσυνδέσουμε τις αδυναμίες του τμήματος της κατανόησης της κοινωνικής σκηνής από τις αδυναμίες του τμήματος της επιλογής απάντησης. Αυτή η προσέγγιση παρακάμπτει επίσης το πρόβλημα της γλωσσικής μεροληψίας των πολυτροπικών μοντέλων, τα οποία τείνουν να επικεντρώνονται στις πληροφορίες κειμένου ενώ αγνοούν πληροφορίες από άλλες τροπικότητες.

Προτεινόμενες Προσεγγίσεις

Προσέγγιση από άκρο σε άκρο: MAC-X

Επισκόπηση

Σε αυτό το πρώτο μέρος της εργασίας μας, προτείνουμε μια πολυτροπική επέκταση του MAC Network για το Social-IQ, που ονομάζεται MAC-Extend (MAC-X). Τα κίνητρα αυτής της προσέγγισης είναι ότι το MAC: 1) προοριζόταν για εφαρμογές που απαιτούν συλλογιστική από γεγονότα σε συμπεράσματα λόγω του δομημένου και επαναληπτικού συστήματος συλλογισμού του και 2) αποτελείται από μονάδες και λειτουργίες γενικού σκοπού. Πιστεύουμε ότι αυτά τα χαρακτηριστικά το καθιστούν κατάλληλο για το Social-IQ και μια ισχυρή βάση για την εφαρμογή του Κοινωνικού Συλλογισμού καθώς και για οποιαδήποτε εφαρμογή συλλογισμού. Το μοντέλο μας βασίζεται στο δίκτυο MAC, μια αναδρομική αρχιτεκτονική μήκους p και διάστασης d που ορίζεται από το κελί Μνήμης, Προσοχής και Σύνθεσης (Memory, Attention and Composition / MAC) που εκτελεί ένα βήμα συλλογιστικής i με βάση την προσοχή δεδομένης μιας βάσης γνώσεων και ενός ερωτήματος. Το κελί MAC αποτελείται από τρεις λειτουργικές μονάδες, τη Μονάδα Ελέγχου, τη Μονάδα Ανάγνωσης και τη Μονάδα Εγγραφής. Αυτή η σειρά μονάδων διαβάζει από τα χαρακτηριστικά εισόδου με τρόπο που ελέγχεται από μέρος της ερώτησης και τη μνήμη από προηγούμενες αναγνώσεις, προχωρώντας στην ενσωμάτωσή τους στην τρέχουσα μνήμη. Ένα από τα πιο σημαντικά χαρακτηριστικά του κελιού είναι ο διαχωρισμός μεταξύ ελέγχου (c_i) και μνήμης (m_i) που επιβάλει, και ότι η αλληλεπίδραση μεταξύ της βάσης γνώσεων και του ερωτήματος διαμεσολαβείται μόνο μέσω κατανομών πιθανοτήτων. Στη διαδοχή των συνολικών p συνεχών επαναλήψεων, έχει την ικανότητα να αναπαριστά αυθαίρετα πολύπλοκα ακυκλικά γραφήματα συλλογιστικής με πιθανοτικό τρόπο [1].

Βασιζόμενο σε αυτές τις δομικές αρχές, το MAC-X εξάγει πληροφορίες από πολλαπλές πηγές, διαμορφώνει την προσοχή του με την πάροδο του χρόνου αντί του χώρου, εκτελεί μια συγχώνευση μεσαίου επιπέδου στις ενδιάμεσες αναπαραστάσεις των τροπικοτήτων και τελικά επιτρέπει την απάντηση ερωτήσεων πολλαπλής επιλογής σε πολυτροπικά δεδομένα. Μια επισκόπηση της αρχιτεκτονικής του μοντέλου για την εφαρμογή του Social Video QA φαίνεται στο σχήμα 1 και η αρχιτεκτονική του βελτιωμένου κελιού φαίνεται στο σχήμα 2. Στις επόμενες ενότητες, όλες οι εξισώσεις και τα σχήματα περιγράφονται για τη δυαδική περίπτωση για απλότητα και μπορούν να επεκταθούν απευθείας για την περίπτωση τεσσάρων επιλογών στην οποία αναφέρουμε επίσης αποτελέσματα.

Μονάδες Εισόδου

Όπως φαίνεται στο σχήμα 1, οι είσοδοι γλωσσικής τροπικότητας που αποτελούνται από την ερώτηση (Q), τη μεταγραφή διαλόγου (T) και τις σωστές και λανθασμένες απαντήσεις (A_1, A_2 αν-

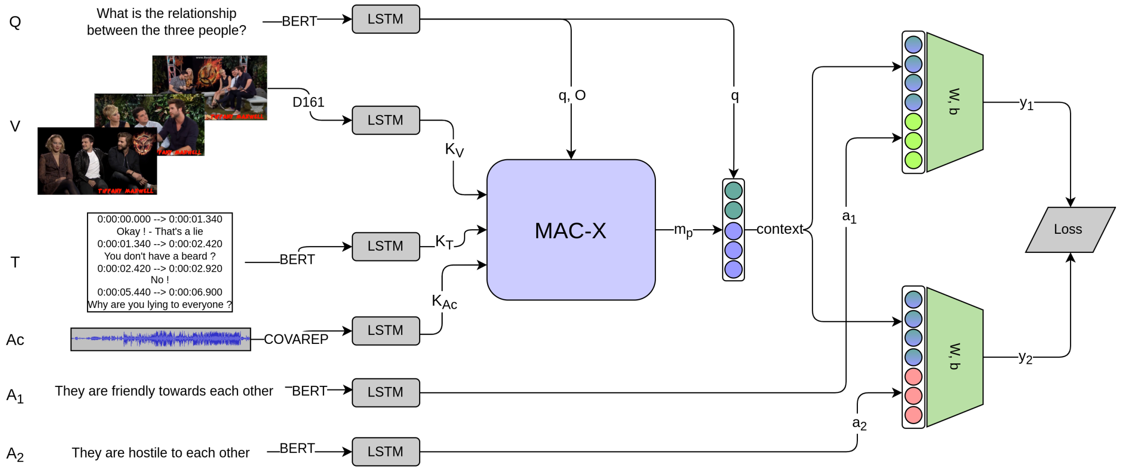


Figure 1. *Επισκόπηση της προτεινόμενης αρχιτεκτονικής από άκρο σε άκρο, με επίκεντρο το δίκτυο MAC-X: Στα αριστερά, η ερώτηση (Q), καρέ βίντεο (V), μεταγραφή διαλόγου (T), ακουστική είσοδος (Ac) καθώς και σωστές (A_1) και εσφαλμένες (A_2) απαντήσεις εμφανίζονται για τη δυαδική περίπτωση. Τα χαρακτηριστικά τους κωδικοποιούνται με LSTM, πριν από τη χρήση στο MAC-X ή σε τελική ταξινόμηση μαζί με την τελευταία μνήμη m_p. Δύο πανομοιότυποι ταξινομητές κάνουν τις προβλέψεις y_1, y_2 οι οποίες στη συνέχεια χρησιμοποιούνται για τον υπολογισμό του σφάλματος στην εξίσωση (7).*

τίστοιχα), κωδικοποιούνται αρχικά με χαρακτηριστικά BERT τελευταίας κρυφής κατάστασης, ενώ η τροπικότητα όρασης (V) με Densenet161 (D161) για κάθε καρέ (στο 1fps) και η ακουστική τροπικότητα (Ac) με COVAREP. Στη συνέχεια περνούν από αμφίδρομα LSTM των οποίων οι έξοδοι αποτελούν τις βάσεις γνώσης K_V, K_T και K_{Ac} για την οπτική, διαλογική και ακουστική είσοδο αντίστοιχα, και τις λέξεις με βάση τα συμφραζόμενα O για την ερώτηση. Οι κρυφές καταστάσεις q, a_1, και a_2 χρησιμοποιούνται ως διανυσματική αναπαράσταση για την ερώτηση και τις απαντήσεις αντίστοιχα. Η διάσταση εξόδου των LSTM είναι d, όπου d είναι η διάσταση του μοντέλου MAC. Κάθε μία από τις βάσεις γνώσης μπορεί να περιγραφεί ως $K_j^{L\times d} = \{k_t|_{t=1}^{L}\}$, όπου L είναι το μήκος ακολουθίας της τροπικότητας j στη χρονική διάσταση t.

Το κελί MAC-X

Μονάδα Ελέγχου

Η Μονάδα Ελέγχου (Εικόνα 2) παραμένει ίδια με την αρχική αρχιτεκτονική και μπορεί να συνοψιστεί ως

$$c_i = \sum_{s=1}^{S} \sigma(f_c(f_{cq}([c_{i-1}, f_q(q)]) \odot O_s)) \cdot O_s \tag{1}$$

όπου S είναι ο αριθμός των λέξεων με βάση τα συμφραζόμενα, σ η συνάρτηση softmax και $f_x = Wx+b$ είναι feedforward δίκτυα ενός επιπέδου. Στην παραπάνω εξίσωση, δίνεται προσοχή στις λέξεις O με βάση πληροφορίες από την ερώτηση q και το προηγούμενο στοιχείο ελέγχου c_{i-1}, προκειμένου να ενημερωθεί η τρέχουσα c_i. Αυτό το c_i καθορίζει με βάση ποιό μέρος της ερώτησης θέλουμε να εξάγουμε γνώση από τις τροπικότητες εισόδου στο τρέχον βήμα συλλογισμού.

Πολλαπλές Μονάδες Ανάγνωσης

Για την ανάγνωση από τις βάσεις γνώσης, προτείνεται μια απλή κλωνοποίηση της Μονάδας Ανάγνωσης για κάθε τροπικότητα, όπου η καθεμία λαμβάνει ένα αντίγραφο του προηγούμενου στοιχείου ελέγχου και μνήμης (βλ. σχήμα 2). Αυτή η προσέγγιση επιτρέπει στον έλεγχο c_i να αποδίδει προσοχή

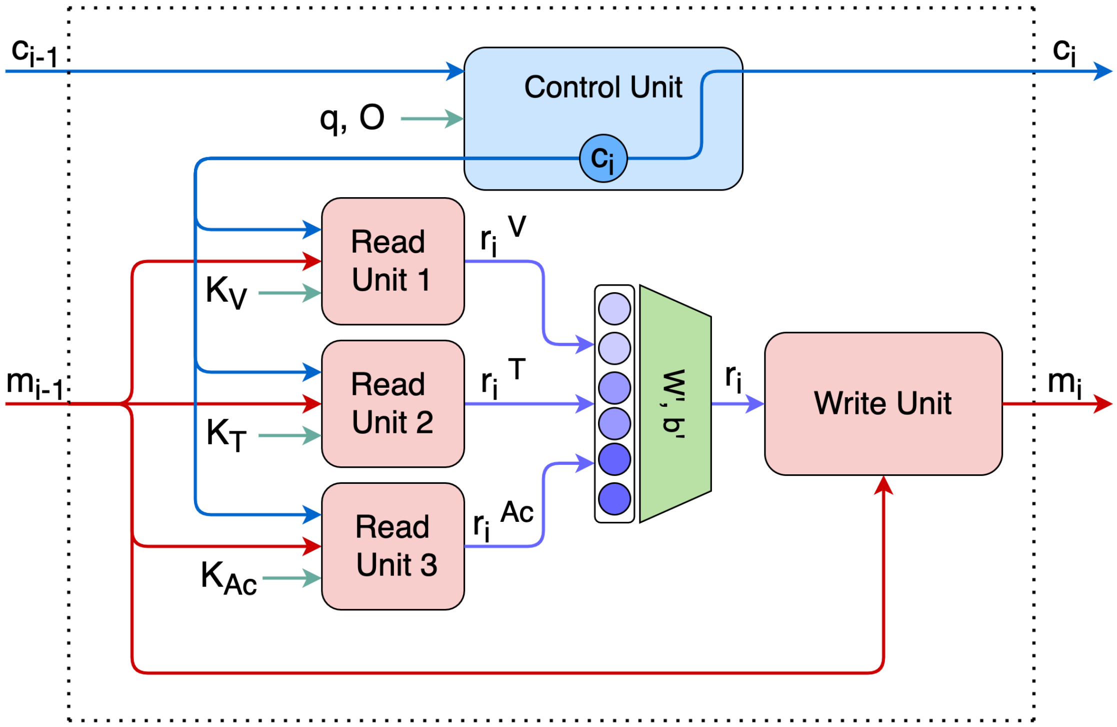


Figure 2. *Το αναδρομικό κελί MAC-X στο i οστό το βήμα συλλογισμού: Η πολυτροπική επέκταση του κελιού MAC εκδηλώνεται με την κλωνοποίηση της Μονάδας Ανάγνωσης και την επακόλουθη συγχώνευση των εξαγόμενων πληροφοριών των τροπικοτήτων r_i^j πριν από την ενσωμάτωση στη μνήμη m_i.*

ανεξάρτητα στις διαφορετικές τροπικότητες στο ίδιο βήμα συλλογισμού, ενώ ταυτόχρονα εξαρτάται από μια μνήμη που διατηρείται συλλογικά για όλες. Για παράδειγμα, οι προηγούμενες πληροφορίες από τις ηχητικές και οπτικές τροπικότητες θα μπορούσαν να είναι σημαντικές για τον προσδιορισμό της επόμενης πιο χρήσιμης πληροφορίας που θα ενσωματωθεί από τη μεταγραφή του διαλόγου. Η λειτουργία κάθε μονάδας ανάγνωσης j ορίζεται ως

$$I_{i,t}^{j} = f_{mk}([f_m(m_{i-1}) \odot f_k(k_t^j), k_t^j]) \tag{2}$$

$$r_i^j = \sum_{t=1}^{L} \sigma(f_r(c_i \odot I_{i,t}^j)) \cdot k_t^j \tag{3}$$

όπου $j = V, T, Ac$ είναι οι διαφορετικές τροπικότητες. Στην πρώτη από τις παραπάνω εξισώσεις, οι πληροφορίες $I_{i,t}^j$ συλλέγονται από τη βάση γνώσης της τροπικότητας j σε κάθε θέση t στη χρονική της ακολουθία. Αυτή η πληροφορία θεωρείται ότι σχετίζεται μόνο προαιρετικά με την προηγούμενη μνήμη m_{i-1} , και έτσι το αρχικό k_t^j συνδέεται επίσης στο διάνυσμα εισόδου της εξίσωσης (2). Στην εξίσωση (3), η προσοχή που βασίζεται στο τρέχον στοιχείο ελέγχου c_i εκτελείται στο $I_{i,t}^j$, για να δημιουργηθεί το τρέχον r_i^j για κάθε Μονάδα Ανάγνωσης.

Πολυτροπική Συγχώνευση

Για να πραγματοποιήσουμε μια συγχώνευση μεσαίου επιπέδου, συγχωνεύουμε τις τροπικότητες σε αυτό το στάδιο συνενώνοντας τα ενδιάμεσα αποτελέσματα εξαγόμενης γνώσης r_i^j για κάθε τροπικότητα j και περνώντας τα μέσα από ένα επίπεδο feedforward, δημιουργώντας αποτελεσματικά ένα ενιαίο επίπεδο κοινής αναπαράστασης r_i για όλες τις τροπικότητες. Αυτό φαίνεται στο σχήμα 2 και στην εξίσωση

$$r_i = W'[r_i^V, r_i^T, r_i^{Ac}] + b' \tag{4}$$

Η εφαρμογή της πολυτροπικής συγχώνευσης σε αυτό το πιο εσωτερικό στάδιο βρίσκεται σε αντίθεση με τις απλούστερες μεθόδους αργότερης συγχώνευσης, με την σύγκριση τους να συζητείται λεπτομερώς στην ενότητα «Αποτελέσματα και συζήτηση».

Μονάδα Εγγραφής

Η Μονάδα Εγγραφής (σχήμα 2) ενσωματώνει τις συνολικές πληροφορίες r_i από τις μονάδες ανάγνωσης στην προηγούμενη μνήμη m_{i-1} και έτσι αποκτά την τρέχουσα μνήμη m_i.

$$m_i = f_{mr}([m_{i-1}, r_i]) \tag{5}$$

Στην παρούσα εργασία παραλείπουμε τα προαιρετικά στοιχεία της Μονάδας Εγγραφής που προτείνονται στο [1] καθώς η χρήση τους δεν αποδείχθηκε σημαντικά ωφέλιμη.

Μονάδα Εξόδου

Μετά από p συνεχείς επαναλήψεις του κελιού MAC-X, όπως περιγράφεται στις προηγούμενες ενότητες, η τελική μνήμη m_p συνενώνεται με την αναπαράσταση της ερώτησης q για να δημιουργηθεί το πλαίσιο στο οποίο θα πρέπει να επιλεγεί η σωστή απάντηση (σχήμα 1). Αυτό συνενώνεται περαιτέρω με καθεμία από τις απαντήσεις a_1, a_2 και μεταβιβάζεται σε πανομοιότυπα feedforward δίκτυα δύο επιπέδων για ταξινόμηση, τα οποία εξάγουν τις προβλέψεις

$$y_1 = W[q, m_p, a_1] + b, \quad y_2 = W[q, m_p, a_2] + b \tag{6}$$

όπου y_1 and y_2 είναι οι προβλέψεις των σωστών και λανθασμένων απαντήσεων αντίστοιχα. Στη συνέχεια υπολογίζουμε το σύνθετο σφάλμα

$$\mathcal{L} = (\frac{1}{N}\sum_{i=1}^{N} y_1^i - 1)^2 + (\frac{1}{N}\sum_{i=1}^{N} y_2^i)^2 \tag{7}$$

όπου N είναι ο αριθμός των δειγμάτων σε μια παρτίδα. Σημειωτέον ότι πρόκειται για την ίδια μετρική σφάλματος που εμφανίζεται στον αρχικό κώδικα που διατέθηκε για το Social-IQ στο [2]. Η δυαδική ακρίβεια A2 διατυπώνεται ως

$$A2 = \frac{1}{M}\sum_{i=1}^{M}(y_1^i > y_2^i) \tag{8}$$

όπου M είναι ο συνολικός αριθμός δειγμάτων στο σύνολο για το οποίο υπολογίζεται η ακρίβεια.

Αποτελέσματα και συζήτηση

Στη συνέχεια παρουσιάζουμε τα αποτελέσματα για την προτεινόμενη αρχιτεκτονική και τις μεθόδους βασικής απόδοσης από αναπαραγωγή. Όλα τα αποτελέσματα υπολογίζονται κατά μέσο όρο από πέντε εκτελέσεις. Οι τροπικότητες εισαγωγής συμβολίζονται ως Q για την ερώτηση, A για τις απαντήσεις, V για τα οπτικά καρέ, T για τη μεταγραφή διαλόγου και Ac για την ακουστική είσοδο. Στον πίνακα 1 συγκρίνουμε το μοντέλο μας (MAC-X) με τα βασικά μοντέλα LSTM και TMFN με βάση τη δυαδική ακρίβεια (A2), σε μια μελέτη σύγκρισης για διαφορετικούς συνδυασμούς των τροπικοτήτων εισόδου. Κάθε συνδυασμός υποδηλώνεται από τις τροπικότητες που χρησιμοποιεί. Παρατηρείται ότι και στα δύο βασικά μοντέλα η πολυτροπικότητα δεν είναι απαραίτητα ευεργετική για την απόδοση, και μπορεί ακόμη και να την υποβαθμίσει σημαντικά. Αντίθετα, το MAC-X αποδίδει καλύτερα όταν χρησιμοποιούνται όλες οι τροπικότητες, σημειώνοντας μια απόλυτη βελτίωση ακρίβειας $0,25\%$ σε σχέση με τις αντίστοιχες εισόδους μεμονωμένης τροπικότητας, γεγονός που υποδεικνύει την ορθότητα των μεθόδων εξαγωγής γνώσης και συγχώνευσης. Ταυτόχρονα είναι πολύ αποτελεσματικό στις ρυθμίσεις μονοτροπικής εισαγωγής, ξεπερνώντας τόσο το βασικό μοντέλο LSTM όσο και το TMFN κατά τουλάχιστον πέντε ποσοστιαίες μονάδες. Όσον αφορά την παρατηρούμενη σημασία κάθε τροπικότητας, οι οπτικές και ακουστικές μορφές φαίνεται να αποδίδουν καλύτερα στα βασικά μοντέλα LSTM και TMFN αντίστοιχα, ενώ το MAC-X επωφελείται εξίσου από όλες τις τροπικότητες. Επιπλέον, δείχνουμε ότι η χρήση μόνο των τροπικοτήτων ερώτησης και απάντησης (ή ακόμα και μόνο της απάντησης) στο LSTM επιτυγχάνει απόδοση πολύ πάνω από την τυχαία, επιβεβαιώνοντας την ύπαρξη γλωσσικής μεροληψίας στο σύνολο επικύρωσης.

Τροπικότητες	**LSTM**	**TMFN**	**MAC-X**
A	63.22 (±0.41)	-	-
QA	64.51 (±0.58)	-	-
QAV	64.82 (±0.67)	65.67 (±0.38)	**71.01** (±0.24)
QAT	64.54 (±0.57)	65.51 (±0.43)	**70.97** (±0.44)
QAAc	64.17 (±0.32)	65.89 (±0.32)	**71.00** (±0.30)
QAVTAc	63.73 (±0.71)	65.62 (±0.55)	**71.25** (±0.15)

Table 1. *Μελέτη σύγκρισης των τροπικοτήτων εισόδου και των βασικών μοντέλων, αναφέροντας αποτελέσματα στο δυαδικό σύνολο επικύρωσης.*

Στον πίνακα 2 παρουσιάζουμε μια μελέτη σύγκρισης που δείχνει την αποτελεσματικότητα της μεθόδου συγχώνευσης μεσαίου επιπέδου, ξεπερνώντας την απόδοση αργότερης συγχώνευσης και στις δύο μετρικές. Στην πειραματική διαμόρφωση του δεύτερου είδους συγχώνευσης, κάθε τροπικότητα διέρχεται από ένα εντελώς ξεχωριστό δίκτυο MAC, των οποίων οι έξοδοι συγχωνεύονται σε αυτό το τελικό στάδιο με τον ίδιο τρόπο όπως στη συγχώνευση μεσαίου επιπέδου, πριν εισέλθουν στους τελικούς ταξινομητές. Αυτό δείχνει το πλεονέκτημα των τρόπων συγχώνευσης στο ενδιάμεσο στάδιο αναπαράστασης στα μοντέλα, όπου οι συλλογικές χρήσιμες πληροφορίες τους μπορούν να υποβληθούν σε περαιτέρω επεξεργασία από κοινού.

Μοντέλα	**A2**	**A4**
MAC με Αργή Συγχώνευση	70.59 (±0.62)	46.46 (±0.26)
MAC-X	**71.25** (±0.15)	**47.22** (±0.60)

Table 2. *Μελέτη σύγκρισης πάνω στο στάδιο της πολυτροπικής συγχώνευσης, αναφέροντας αποτελέσματα για το σύνολο επικύρωσης για όλες τις εισόδους τροπικοτήτων.*

Στον Πίνακα 3 συγκρίνουμε την απόδοση του προτεινόμενου μοντέλου με πέντε προηγούμενες μεθόδους τελευταίας τεχνολογίας, αναφέροντας αποτελέσματα και στις δύο μετρικές για το σύνολο επικύρωσης. Παρατηρούμε μια βελτίωση ακρίβειας $2,3-2,6\%$ από την προηγούμενη καλύτερη επίδοση στη δυαδική μετρική ακρίβειας (MCQA [17]), λαμβάνοντας υπόψη τη διακύμανση. Όσον αφορά την μετρική πολλαπλών επιλογών (A4), λαμβάνουμε συγκρίσιμα αποτελέσματα με το μοντέλο με την καλύτερη απόδοση TACO-Net [18]. Σημειωτέον ότι το TACO-Net μετρά ρητά τη συνέπεια μεταξύ κάθε απάντησης και τροπικότητας, συμβάλλοντας στην ευρωστία του μοντέλου στην περίπτωση πολλαπλών επιλογών. Συνολικά, μέσω της υλοποίησης και εφαρμογής του MAC-X, ορίσαμε μια νέα καλύτερη απόδοση για τη δυαδική μετρική απόδοσης του συνόλου δεδομένων Social-IQ.

Μοντέλα	**A2**	**A4**
TMFN [2]	65.62	36.24
Removing bias [19]	67.93	-
TACO-Net [18]	68.19	**49.08**
Perceptual score [20]	68.65	-
MCQA [17]	68.80	38.30
Ours (MAC-X)	**71.25** (±0.15)	47.22 (±0.60)

Table 3. *Σύγκριση επίδοσης με την τελευταία τεχνολογία στο σύνολο επικύρωσης του Social-IQ. Αναφέρουμε αποτελέσματα με μέση τιμή και τυπική απόκλιση σε 5 εκτελέσεις.*

Προσέγγιση Επαύξησης: Emogaze

Επισκόπηση

Οι περισσότερες προηγούμενες εργασίες θεωρούν το βίντεο ως μια συνεχή πηγή πάνω στην οποία μπορεί να υπολογιστεί κάποια κατανομή προσοχής. Αν και αυτή η γενική προσέγγιση έχει νόημα στις περισσότερες εφαρμογές VQA και Video QA, όπου οι ερωτήσεις αναφέρονται στο περιβάλλον, τα αντικείμενα ή τις ενέργειες και τα γεγονότα, το Social Video QA περιστρέφεται γύρω από τους ανθρώπους και τις αλληλεπιδράσεις τους. Εάν επρόκειτο να χωρίσουμε σε δομικές μονάδες τα κοινωνικά βίντεο, τότε αυτές θα ήταν τα άτομα που συμμετέχουν στις αλληλεπιδράσεις και οι πληροφορίες που ανταλλάσσουν, τόσο λεκτικές όσο και μη. Οι άνθρωποι επικοινωνούν μέσω γλώσσας και μη λεκτικών σημάτων όπως εκφράσεις προσώπου, βλέμματα, χειρονομίες και γλώσσα του σώματος.

Επιπλέον, αντλώντας έμπνευση από το [16], που δηλώνει ότι η κατανόηση της φυσικής γλώσσας θα ολοκληρωθεί μόνο με την ενσωμάτωση της πολυτροπικής και κοινωνικής σημασιολογίας, επιλέξαμε την εκπαίδευση ενός αμιγώς NLP μοντέλου με πολυτροπική είσοδο (WS3) που περιέχει κοινωνικές ενδείξεις (WS5), φιλτραρισμένες για να περιέχουν μόνο δεδομένα βλέμματος, δεδομένα συναισθημάτων, δεδομένα αναγνώρισης αντικειμένων και δεδομένα συνομιλίας, όλα μεταφρασμένα σε γλώσσα, στην ουσία εξάγοντας περιγραφές βίντεο.

Αυτό υποστηρίζεται επίσης από εργασίες για VQA που εκμεταλλεύονται τις ανώτερες συλλογιστικές ικανότητες των μοντέλων φυσικής γλώσσας σε σύγκριση με τα μοντέλα όρασης υπολογιστή που τείνουν να είναι πιο ενστικτώδη, και παρακάμπτουν το πρόβλημα γλωσσικής μεροληψίας πολυτροπικών μοντέλων που τείνουν να εστιάζουν στην πληροφορία κειμένου αφού είναι ευκολότερο να ληφθούν στατιστικές πληροφορίες από αυτό, ενώ αγνοούνται πληροφορίες από άλλες τροπικότητες. Ένα επιπλέον πλεονέκτημα αυτής της προσέγγισης είναι η αυξημένη δυνατότητα επεξήγησης τόσο σε περιπτώσεις επιτυχίας όσο και σε περιπτώσεις αποτυχίας, καθώς τα χαρακτηριστικά που ανιχνεύονται και βρίσκονται στις λεζάντες μπορούν να αντικατοπτρίζουν αυτό που εξάγει το σύστημα από τα καρέ, και τα ενδιάμεσα αποτελέσματα μπορούν να μας βοηθήσουν να αποσυνδέσουμε τις αδυναμίες του τμήματος της κατανόησης της κοινωνικής σκηνής από τις αδυναμίες του τμήματος επιλογής απάντησης.

Στην παρούσα εργασία, επιλέγουμε να επαυξήσουμε τη λεκτική κατανόηση των κοινωνικών αλληλεπιδράσεων με το συναισθηματικό περιεχόμενο του προσώπου που μεταφέρεται μέσω του βλέμματος των ματιών. Συγκεκριμένα, για κάθε ανιχνευμένο άτομο s σε ένα καρέ βρίσκουμε το άτομο t στο οποίο στοχεύει το βλέμμα του και καθορίζουμε το βλέμμα $g = (s, t)$. Αυτό το βλέμμα φέρει ένα συναισθηματικό και ένα προαιρετικό λεκτικό βάρος, εάν το άτομο είναι επίσης ο ομιλητής $W_g = (\text{face emotion}[, \text{uttered phrase}])$. Επιπλέον, με στόχο τη διατήρηση μόνο του υψηλότερου κοινωνικού περιεχομένου, τα βλέμματα φιλτράρονται κρατώντας μόνο τον ομιλητή και το άτομο που κοιτάζει, με αποτέλεσμα έναν γράφο $G_f = \{(s, t), (t, r)\}$ για το καρέ f. Όλοι οι συμμετέχοντες κόμβοι (άτομα) μπορούν να περιγραφούν με όρους οπτικής εμφάνισης, χωρισμένους σε χαρακτηριστικά και είδος αντικειμένου, π.χ. $s := \text{attr} + \text{obj}$. Με αυτή τη διατύπωση του προβλήματος, κατασκευά-

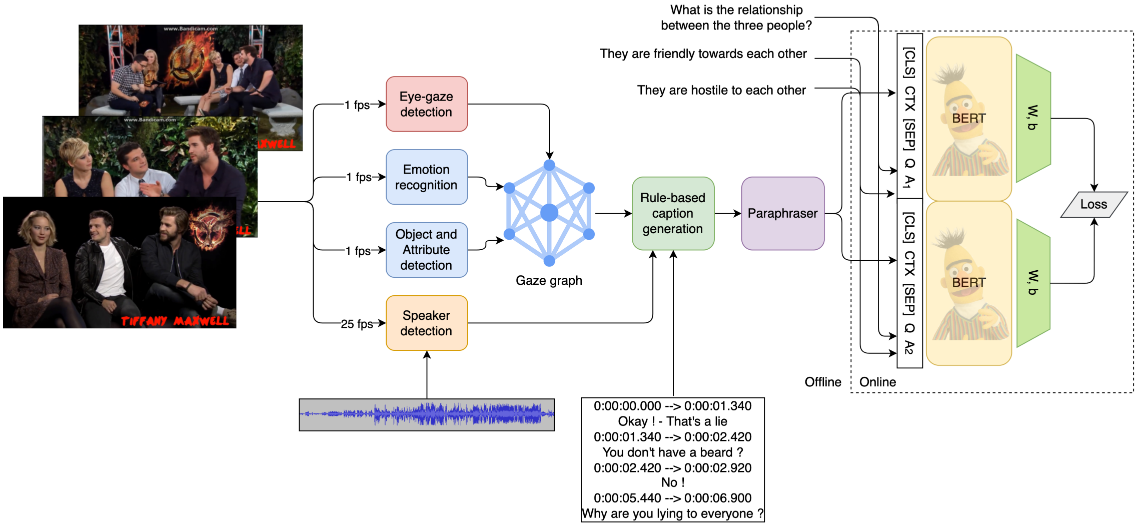


Figure 3. *Επισκόπηση της προτεινόμενης αρχιτεκτονικής επαύξησης και απάντησης ερωτήσεων πάνω σε κείμενο, που αποτελείται από ένα ασύγχρονο τμήμα παραγωγής περιγραφών βίντεο (αριστερά), ακολουθούμενο από ένα σύγχρονο τμήμα εκπαίδευσης (δεξιά). Στα αριστερά, οι τρείς ανιχνευτές του βλέμματος, συναισθήματος και αντικειμένων αντίστοιχα, συνδυάζονται σε μία ενδιάμεση δομή που μπορεί να ερμηνευθεί ως γράφος βλεμμάτων. Με την προσθήκη της μεταγραφής διαλόγου και της αναγνώρισης του ομιλητή, παράγεται μία περιγραφή βασισμένη σε κανόνες, προαιρετικά ακολουθούμενη από μοντέλο παράφρασης. Αυτές οι περιγραφές βίντεο στη συνέχεια χρησιμοποιούνται για απάντηση ερωτήσεων πολλαπλής επιλογής σε προεκπαιδευμένα μοντέλα τύπου BERT, πριν τους τελικούς ταξινομητές.*

ζουμε περιγραφές φυσικής γλώσσας μέσω ενός συνόλου κανόνων, εκτελώντας έτσι μια μετατροπή γράφου σε κείμενο. Αυτές οι περιγραφές επεξεργάζονται προαιρετικά από ένα δίκτυο παράφρασης, για να παραχθεί περισσότερο φυσικό και με περισσότερες παραλλαγές περιεχόμενο. Αυτές οι λεζάντες βίντεο με επίκεντρο τον άνθρωπο χρησιμοποιούνται ως περιβάλλον εισαγωγής σε μοντέλα γλώσσας τύπου BERT για την εκτέλεση απαντήσεων σε ερωτήσεις πολλαπλών επιλογών.

Επιπλέον η παροχή ορισμένων παραλλαγών σε μια περιγραφή βασισμένη σε κανόνες θα την κάνει πιο φυσική και επομένως πιο κοντά στα δεδομένα προεκπαίδευσης των transformer-based μοντέλων. Σε αυτή την εργασία υλοποιούμε την παράφραση μέσω backtranslation. Η βασική ιδέα του backtranslation είναι ότι, μέσω της μετάφρασης μιας πρότασης σε άλλη γλώσσα και στη συνέχεια της μετάφρασης πίσω στην αρχική (με δύο αντίστοιχα προεκπαιδευμένα μοντέλα μετάφρασης, π.χ. transformers), η τελική έξοδος θα αποτελεί μια μικρή παραλλαγή της εισόδου, αλλά δεν θα έχει χάσει το νόημά της. Αυτό δεν συμβαίνει με άλλα μοντέλα που έχουν εκπαιδευτεί ειδικά για την εφαρμογή της παράφρασης, όπως το T5, καθώς αυτά τα μοντέλα τείνουν να αφαιρούν χρήσιμες πληροφορίες και να προσθέτουν μη ρεαλιστικές λεπτομέρειες.

Προτεινόμενα μοντέλα

Η βασική ιδέα είναι ο πειραματισμός με τη χρήση χαρακτηριστικών από μεγάλα προεκπαιδευμένα μοντέλα γλώσσας σαν το BERT ή η επιλεκτική εκπαίδευση ορισμένων από τα τελευταία επίπεδά τους, για την εφαρμογή απάντησης σε ερωτήσεις πολλαπλής επιλογής. Αυτό οφείλεται στο ότι το επαυξημένο σύνολο δεδομένων που δημιουργήσαμε είναι μικρό (μόνο 1015 μοναδικές περιγραφές βίντεο) και μπορεί να παρατηρηθεί ότι έχει πολύ διαφορετική κατανομή από το σώμα στο οποίο έχει προεκπαιδευτεί το BERT. Πειραματιζόμαστε με δύο μοντέλα, το βασική έκδοση BERT και την μεγάλη έκδοση RoBERTa, προηγουμένως εκπαιδευμένο σε ένα άλλο σύνολο δεδομένων απάντησης ερωτήσεων πολλαπλής επιλογής κειμένου, το RACE [21]. Το τελευταίο επιλέγεται με στόχο να ξεκινήσει με καλύτερες αρχικοποιήσεις βαρών για την ίδια εφαρμογή, και επομένως να περιέχει το σενάριο μεταφοράς εκμάθησης σε διαφορετικούς μόνο τομείς αντί για διαφορετικό τομέα και εφαρμογή. Επιπλέον, πειραματιζόμαστε με ένα πιο επιθετικό σχήμα εκπαίδευσης για το μικρότερο μοντέλο DistilBERT .

Ένα άλλο ζήτημα κατά τη χρήση τέτοιων γλωσσικών μοντέλων για την ερμηνεία μεγάλων όγκων κειμένου είναι ότι το μέγιστο μήκος εισαγωγής είναι περιορισμένο, π.χ. σε 512 ή 1024 στοιχεία ανάλογα με το μοντέλο. Για να το αντιμετωπίσουμε αυτό, πειραματιστήκαμε με τη σύνοψη μέσω εξαγωγικής απάντησης ερωτήσεων, χρησιμοποιώντας BERT εκπαιδευμένο στο εξαγωγικό σύνολο δεδομένων ερωταπαντήσεων SQUAD [22]. Η ιδέα είναι να εξάγουμε χρήσιμη πληροφορία από την είσοδο που εξαρτάται από την ερώτηση και να προχωρήσουμε στην απάντηση της ερώτησης πολλαπλής επιλογής βάσει αυτής της μικρότερης πληροφορίας. Ωστόσο, η εξαγωγική απάντηση ερωτήσεων έχει επίσης ένα μέγιστο μήκος εισόδου, το οποίο οδηγεί στην ιδέα της εκτέλεσης αυτών των περιλήψεων με ιεραρχικό τρόπο σε μικρότερα μέρη του κειμένου και στη συνέχεια στο αποτέλεσμα της συνένωσης των αποσπασμάτων. Αυτά τα μικρότερα μέρη μπορούν να ληφθούν είτε χρησιμοποιώντας τμηματοποίηση σκηνής (από την έξοδο SyncNet που ανιχνεύει την αλλαγή λήψης κάμερας) είτε κόβοντας σε ίσα μέρη.

Ο τρόπος με τον οποίο εισάγονται τα μέρη του συστήματος απαντήσης ερωτήσεων σε ένα μοντέλο τύπου BERT μπορεί να περιγραφεί ως εξής, για τη δυαδική εφαρμογή:

$$\begin{aligned} E_1 &= bert([CLS] + CTX + [SEP] + Q + A_1) \\ E_2 &= bert([CLS] + CTX + [SEP] + Q + A_2) \end{aligned} \tag{9}$$

όπου το CTX είναι η περιγραφή του βίντεο που δημιουργήσαμε, τα E_1 και E_2 είναι τα χαρακτηριστικα CLS τελικού επιπέδου και το $+$ υποδηλώνει συνένωση. Οι προβλέψεις εξόδου $Y_1, Y_2 = f(E_1), f(E_2)$ όπου το f είναι ένας γραμμικός ταξινομητής τροφοδοτούνται σε τυπική απώλεια διασταυρούμενης εντροπίας μαζί με την ετικέτα σωστής απάντησης.

Αποτελέσματα και συζήτηση

Κάποια πρώτα αποτελέσματα που πήραμε με αυτή την αρχιτεκτονική παραγωγής κοινωνικών περιγραφών βίντεο είναι δοκιμάζοντας τα προεκπαιδευμένα μοντέλα BERT base και RoBERTa επιπλέον fine-tuned στο σύνολο δεδομένων απάντησης ερωτήσεων κατανόησης πάνω σε μεγάλα κείμενα, το RACE.

Παρατηρούμε ότι το BERT base, καθως δεν ειναι fine-tuned για την εφαρμογή της απάντησης ερωτήσεων πολλαπλής επιλογής πάνω σε μεγάλα κείμενα, δεν μπορει να εκμεταλλευτει επιτυχως τις περιγραφες βίντεο, ενώ όμως ευνοείται από τη χρήση ενός aggregated emotion tag από όλο το video μαζί με την ερώτηση και απάντηση, δείχνοντας τη συσχέτιση των ερωτήσεων με το συναίσθημα. Από την άλλη, το RoBERTa RACE, αποδίδει καλύτερα όταν χρησιμοποιούνται οι επαυξημένες περιγραφές μας, και συγκεκριμένα ακόμα καλύτερα με το επιπλέον βήμα της παράφρασης.

Επιπλέον να σημειωθεί ότι εδώ χρησιμοποιήσαμε τα μοντέλα παγωμένα, ενώ το fine-tuning δεν βοήθησε. Τα πειράματα και οι αναλύσεις μας στο δεύτερο μέρος της εργασίας ενώ δεν είναι στο ίδιο επίπεδο ωριμότητας με του πρώτου μέρους, θεωρούμε ότι δίνουν κάποια ενθαρρυντικά αποτελέσματα, ιδίως όσον αφορά τα ενδιάμεσα ποιοτικά αποτελέσματα που παίρνουμε από τα gaze graphs και τις περιγραφές.

Είσοδος	BERT	RoBERTa RACE
QA	74.37 (±0.06)	79.05 (±0.05)
QA+transcript	74.34 (±0.05)	78.50 (±0.06)
QA+emo	**74.87** (±0.06)	79.24 (±0.05)
QA+rule-based ctx	73.12 (±0.05)	79.43 (±0.06)
QA+paraphrased ctx	74.21 (±0.07)	**79.56** (±0.06)

Table 4. *Συγκριτική μελέτη εκπαίδευσης ταξινομητή πάνω σε χαρακτηριστικά τελευταίου επιπέδου με βάση διαφορετικά επίπεδα επαύξησης εισόδου. Αναφέρουμε αποτελέσματα με μέση τιμή και τυπική απόκλιση σε 5 εκτελέσεις.*

Συμπεράσματα

Σε αυτήν την εργασία, διερευνούμε δύο πολύ διαφορετικές προσεγγίσεις για την εφαργμογή Social Video Question Answering, συγκεκριμένα για το σύνολο δεδομένων Social-IQ. Αυτό έγινε για να διερευνηθούν ειδικά οι δυνατότητες που λείπουν συχνά από τα συστήματα μηχανικής εκμάθησης, αλλά είναι πολύ απαραίτητες ειδικά στο Social Video QA, όπως οι λειτουργίες ρητού συλλογισμού και ανίχνευσης κοινωνικών ενδείξεων.

Στην πρώτη προσέγγιση, ακολουθούμε ένα σχήμα εκπαίδευσης από άκρο σε άκρο χρησιμοποιώντας προεκπαιδευμένα χαρακτηρστικά CNN και χαρακτηριστικά ήχου, και εκτελούμε τη συγχώνευση τροπικοτήτων μέσω μιας επέκτασης του δικτύου MAC που ονομάζουμε MAC-X. Πιο συγκεκριμένα παρουσιάζουμε το MAC-X, μια πολυτροπική επέκταση του δικτύου MAC ικανή να χειρίζεται σύνθετες εφαρμογές συλλογισμού πολλαπλών επιλογών και πολλαπλών τροπικοτήτων, όπως το Social-IQ, όπου το αξιολογούμε και λαμβάνουμε αποτελέσματα τα οποία ξεπερνούν τις υπάρχουσες πιο αποδοτικές μεθόδους. Καταλήγουμε στο συμπέρασμα ότι οι δομικές αρχές καθώς και ο συλλογισμός σύνθεσης μπορούν να αποδειχθούν χρήσιμα για το Social Video Question Answering, στο οποίο - εξ όσων γνωρίζουμε - αυτή η κατεύθυνση εφαρμόζεται για πρώτη φορά. Μπορούμε περαιτέρω να επιβεβαιώσουμε από τις συγκριτιές μας μελέτες ότι το MAC-X μπορεί να επωφεληθεί αποτελεσματικά από όλες τις τροπικότητες και ότι η συγχώνευση μεσαίου επιπέδου αποδίδει σημαντικά καλύτερα από την αργότερη συγχώνευση.

Στη δεύτερη προσέγγιση, ακολουθούμε την κατεύθυνση εκτέλεσης του VQA σε περιγραφές βίντεο, τις οποίες λαμβάνουμε μέσω της επαύξησης των μεταγραφών διαλόγων με πληροφορίες συναισθήματος στο βλέμμα καθώς και άλλα χαρακτηριστικά της εικόνας. Τα συναισθηματικά βλέμματα συνδέουν τους ανθρώπους που εμπλέκονται σε μια κοινωνική σκηνή, τα οποία φιλτράρονται μέσω ανίχνευσης ομιλητών, η οποία τους συνδέει επίσης με τη μεταγραφή διαλόγου. Είναι η πρώτη φορά που - από όσο γνωρίζουμε - προτείνεται ένα τέτοιο σύστημα εξαγωγής χαρακτηριστικών ειδικά σχεδιασμένο για κοινωνικά βίντεο. Θεωρούμε ότι παρέχει ένα γενικό πλαίσιο για τη επεξεργασία των κοινωνικών πληροφοριών σε βίντεο. Επιπλέον, μέσω της επαύξησης των μεταγραφών των διαλόγων και της δημιουργίας περιγραφών βίντεο, παρέχουμε επίσης μια βάση στην οποία μπορούμε να στηρίξουμε πιο εξελιγμένες μεθόδους δημιουργίας περιγραφών σε κοινωνικό βίντεο. Τέλος, παρέχουμε συγκριτικές μελέτες για πολλά μοντέλα γλώσσας τύπου BERT και επίπεδα εκπαίδευσης, καθώς και ένα ιεραρχικό σχήμα περίληψης που βασίζεται σε σύστημα εξαγωγικής απάντησης ερωτήσεων.

Chapter 1

Introduction

1.1 Motivation

Everyday, the sensory input one receives both from the physical and the digital world comes in multiple modalities and overwhelming quantity. It is crucial that AI agents are developed in such a direction that they aid humans in their everyday interaction with this huge amount of multimodal data, by alleviating their shortcomings, enhancing their experience, and at the same time protecting them from harm. To give a concrete example of a real-world application for each of these three motivating factors, we can consider the following. First, AI systems trained to understand the visual environment or the social interactions around them can provide meaningful assistance to people with physical or mental disabilities like blindness or autism, by giving them information about the world that they cannot get on their own. Second, from smart media recommendations and retrieval in applications like YouTube and Netflix, to augmented reality systems for entertainment and educational purposes, there has been great focus from both the research community and the industry towards enriching daily life. Finally, an important problem is hateful media detection on the internet [13], as people are targeted based on their race, gender, or sexual orientation, leading to serious mental health problems and even real-world hate crimes.

Vision and language are two of the most fundamental and most remarkable capabilities of the human mind, since vision enables the creation of mental concepts that would not otherwise exist such as colour and light, and language has the power to distill all sensory experience to these elementary mental concepts and combine them to create complex new ideas, thus facilitating thinking through making an “infinite use of finite means” [14]. The interaction of language with vision motivates researchers to identify the relations between the modalities, combine and reason about them for decision making. A prominent task in vision-and-language machine learning is the task of Visual Question Answering (VQA) [23], which requires the AI agent to answer a natural language question based on an input image. An important challenge is that language can inadvertently impose strong priors, which give it a tendency to be an easier signal to learn from than the visual modality, resulting in vision-and-language models that completely disregard visual information in favor of exploiting language biases.

Humans are social creatures; our survival and well-being depends on our effective communication with others. This is achieved through perceiving and understanding information from multiple sensory modalities as well as reasoning and arriving to conclusions, in order to respond accordingly. More precisely, we rely on the ability to understand other peoples’ mental states (which include intentions, motivations, feelings), through processing information like their eye gaze, facial expression, body language (posture, gestures), and tone of voice. Some people, albeit very intelligent like people with ASD, cannot discern those cues as they operate mainly on logical and factual information, as is the case with most machine learning tasks. The invention of a question answering task for this matter serves both as a way to train people with ASD to recognise such behaviours, as well as machine learning models when framed similarly to other question answering tasks. Social Video Question Answering is a task to test the social reasoning abilities of an agent, based on how accurately they can answer questions on a given video. It can require sophisticated combinations of emotion recognition, language understanding, cultural knowledge, logical and causal reasoning, on top of non-social layers of comprehension about physical events.

Having a problem at hand that is essentially a reasoning task, we draw inspiration from a portion of VQA literature called neurosymbolic models that focuses on building neural models while at the same time enabling explicit high level symbolic reasoning. From these we focus on approaches closer to the neural rather than symbolic side, such as Learning from abstraction [15] and Memory Attention Composition (MAC) Network [1]. Memory Attention Composition (MAC) Network attempts to capture the logic of thought in addition to constructing neural representations from the data and was intended for tasks that require deliberate reasoning from facts to conclusions on account of its structured and iterative reasoning.

In the first part of this work, we propose an end-to-end approach based on a multimodal extension of the MAC Network [1] for Social Video Question Answering, called MAC-Extend (MAC-X). We leverage MAC's reasoning capabilities and base MAC-X on a recurrent cell that performs iterative mid-level fusion of input modalities (visual, auditory, text) over multiple reasoning steps, by use of a temporal attention mechanism. We then combine MAC-X with LSTMs for temporal input processing in an end-to-end architecture.

However, this approach as well as most work in literature regards video as a continuous source containing equally important attributes to consider attending to. Although this general approach makes sense in most VQA tasks where questions refer to the environment, objects, or actions and events, Social Video QA revolves around people and their interactions through which they exchange both verbal and non-verbal information.

In the second part of this work, we propose an augmentation approach of explicitly extracting social cues from video and connecting them to the people participating in the social interaction, which is modeled through eye-gaze to form scene-specific gaze graphs. Through this, we also wish to explore the hypothesis that eye-gaze can summarize social video.

To choose how to process these gaze graphs for the goal of answering natural language questions that require social reasoning we consider the following motivating factor. In [16] authors make the case that NLP needs social context to truly succeed, marking that as the last stage to obtain a truly complete understanding of the world through language. More specifically, the NLP progress is defined by the conquering of different World Scopes (WS), each one more general than the last, ordered as Corpus, Internet, Perception (multimodal), Embodiment, and Social.

Drawing inspiration from this analysis, we take the direction of training a purely NLP model through translating multimodal (WS3) input that contains social (WS5) clues into language, effectively performing captioning. An additional benefit compared to end-to-end models is the added explainability both in success and in failure cases, as the intermediate results (video captions) can help us decouple the inabilities of the social scene understanding part from the inabilities of the answer inference part. This approach also bypasses the language bias problem of multimodal models, which tend to focus on the textual information while ignoring information from other modalities.

1.2 Contributions

Our main contributions are the following:

- We present MAC-X, a multimodal extension of the MAC Network [1], featuring temporal attention, a mid level fusion mechanism, and multiple-choice Video Question Answering capabilities.
- Through using MAC, we obtain a model for social reasoning that uses explicit reasoning steps, and apply it to the challenging Social-IQ dataset [2], analysing its performance through ablation studies and comparison to prior state-of-the-art methods.
- Our ablation studies show that the proposed MAC-X architecture can effectively leverage multimodal input cues using mid-level fusion mechanisms, and we obtain a 2.5% absolute improvement in terms of binary accuracy over the current state-of-the-art.
- We uncover an important aspect of language bias in the Social-IQ dataset, which exists in analogy to most VQA datasets and was hidden behind a miscalculation in the precomputed embeddings.
- We introduce a novel pipeline for Social Video Question Answering based on explicitly detecting social cues and connecting them via eye gaze. Through this, we provide a framework for leveraging social information in video, which can be adapted, extended, and used with multiple different machine learning architectures for question answering.
- We propose translating multimodal detections in natural language, to use the resulting video captions as context for question answering, and provide a baseline for future social captioning approaches.
- We perform ablations between different pre-trained language models, and observe that augmentation through social cues enhances the pure dialogue understanding by a significant amount.

1.3 Thesis Outline

In Chapter 2, Machine Learning Background, we provide the theoretical foundations for the reader to familiarize themselves with the field of Machine Learning, with a focus on Supervised Learning techniques and the more recent advances of Deep Learning, as well as Transfer Learning, to which most of the recent achievements in Artificial Intelligence can be attributed.

In Chapter 3, Multimodal Machine Learning, we explore and analyse the theory, as well as the previous work in the field of Multimodal Machine Learning, in which this Thesis belongs. In particular, we dive more into the Vision and Language modalities, and the task of Question Answering, both in images and in video.

In Chapter 4, Social Video Question Answering, we define the problem that this Thesis addresses, which is Social Video Question Answering, and review the available datasets as well as the previous work, focusing on the Social-IQ dataset. Additionally, we provide different setups for the baselines of our experiments, comparing different approaches and analysing characteristics of the dataset.

In Chapter 5, Proposed Approaches, we present our two different approaches to the task, which are applied to the Social-IQ dataset. The first one, which we will refer to as the end-to-end approach, offers an extension of the MAC network and is compared to multiple baselines as well as the state-of-the-art. The second one, offers an augmentation approach to the Social-IQ data, which consists of incorporating social information from the video in textual descriptions, and is analysed in multiple setups comparing to the unaugmented data.

In Chapter 6, Conclusions, we discuss our research efforts and draw conclusions on their limitations, as well as future work that could build and enhance them. We end with a dedicated section considering the ethical implications of applying this research in real world use case scenarios.

Chapter 2

Machine Learning Background

2.1 Introduction

Machine Learning is a branch of Artificial Intelligence (AI) which focuses on the use of data to algorithmically imitate the way that humans learn, gradually improving performance on tasks that require some level of intelligence to complete. It was defined in the late 1950s by AI pioneer Arthur Samuel as "the field of study that gives computers the ability to learn without explicitly being programmed". In traditional programming, computers are given detailed instructions to follow which, for tasks like recognizing different people or objects in images, verges on the impossible. Machine learning takes the approach of letting computers learn to program themselves through experience (i.e. observing a set of example data) with the help of statistical models and optimization algorithms. This set of example data is called the training data, and the process of learning from this data is called training. The more data, the better the program. The set of parameters that can be refined through this training, along with their underlying architecture is often described as a machine learning model. Human programmers choose a machine learning model to use, supply the data, and optionally tweak the model's external parameters to help push it toward more accurate results.

Earlier efforts in artificial intelligence focused on expert knowledge systems which, based on logical inference rules, derived new fragments of knowledge or reasoned over statements. This is also referred to as symbolic AI and is characterized by serious limitations such as the difficulty of formally describing all possible knowledge based on a given task. Machine learning approaches such as neural networks were initially disregarded due to infeasibility concerns which, however, were refuted by the technological developments around storage and processing power, enabling the success of modern artificial intelligence with real life applications practically everywhere. In fact, most current advances in AI involve machine learning, causing the terms to be used interchangeably. Deep Learning is a branch of machine learning that focuses on deep neural network architectures, leveraging huge amounts of data to learn what are called meaningful representations. The main difference from shallow architectures and other machine learning algorithms is that those often require more structured data to learn, thus depending on human intervention to determine a set of hand-crafted features, while deep learning automates much of this feature extraction process, enabling the use of larger data sets. In this way, deep learning can also be thought of as scalable machine learning. From this point on we will refer to "non-deep" machine learning methods with the term "classical machine learning" for clarity.

Types of Data

Machine Learning can be divided in fields depending on the source of input data, and whether a single or multiple sources are used. Data sources are often referred to as "modalities", and can either come from measurements of quantities in tabular format, or in signal format. Such signals can include 1D timeseries such as stocks data and brain signals, as well as 2D signals such as images and video, and many ideas in machine learning are evolved from concepts in the field of signal processing. In the case that a machine learning model uses, or a task requires data from a single modality, the respective model or task is called unimodal, and in the case multiple modalities are required they are called multimodal. Some important areas that emerged from different types of data are Natural Language Processing (NLP), Computer Vision (CV), Speech Processing, and Recommender Systems, to mention a few.

Types of Learning

There are two main approaches in using the data machine learning is based on. The first one involves labeling a set of data with ground truth information and using algorithms that optimize performance through minimizing the total loss (error) of predictions with respect to those ground truth labels, pushing the model to a better internal set of parameters that describes the data better. The ability of the model to perform well in projecting accurate representations of data outside the distribution of this labeled training data is called "generalizability". This first approach involving labeled data (supervision) is called Supervised Learning.

The second approach, called Unsupervised Learning, is based on learning from the internal structure of the data through finding similarities by which they can be grouped, or projecting them into a space where they can represent different classes or clusters.

We will delve more into these two approaches, as well as their hybrid combinations in the next sections. In our discussion of supervised and unsupervised learning next, we will explore different models from the viewpoint of classical machine learning. Most algorithms in the next sections are what is called parametric, meaning an assumption on the distribution of the data is made in order to describe it using a finite set of parameters. We will identify instances of the non-parametric models as we come across them. The following discussion in the next sections largely follows Bishop's [24] book.

2.2 Unsupervised Learning

Unsupervised learning models are utilized for two main tasks — clustering and dimensionality reduction.

2.2.1 Clustering

Clustering is the problem of grouping together different data points based on certain factors of similarity that they share in some multidimensional space. According to [24] we can think of a cluster as a group of data points whose inter-point distances are small compared with the distances to points outside of the cluster.

K-Means

K-means clustering is one of the simplest and most popular clustering algorithms. In k-means there is a predefined number of clusters k, each corresponding to a centroid. These centroids are points in the space of the data, which are randomly initialized. The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.

This is described by the objective function $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\mathbf{x}_n - \boldsymbol{\mu}_k||^2$, $r_{nk} \in \{0, 1\}$ which represents the sum of the squares of the distances of each data point $\mathbf{x}_n$ to each centroid $\boldsymbol{\mu}_k$. Our goal is to find values for the r_{nk} and the $\boldsymbol{\mu}_k$ so as to minimize J. This is achieved through an iterative process with two steps per iteration corresponding to optimizations with respect to r_{nk} and $\boldsymbol{\mu}_k$ respectively, which is repeated until convergence. These two stages correspond to the expectation (E) and maximization (M) steps of the EM algorithm which we explore later.

The "means" in k-means is derived from the fact that at the M-step when setting the derivate with respect to $\boldsymbol{\mu}_k$ to zero and solving for $\boldsymbol{\mu}_k$ we get the mean of all the data points assigned to the cluster k. At each iteration, each data point is assigned to the cluster that minimizes the sum, and then the centroids are recalculated as the mean of the data points assigned in each cluster.

Gaussian Mixture Models

K-means is a hard clustering method, which means that each point is associated to one and only one cluster. In that approach, there is no probability measure to describe the degree to which a data point is associated with a specific cluster. In probabilistic (or soft) clustering, data points are clustered based on the likelihood that they belong to a particular distribution.

The Gaussian mixture distribution can be written as a linear superposition of Gaussians in the form $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Each gaussian is identified by k $\in \{1, \ldots, K\}$, where K is the number of clusters of our dataset and π_k is the mixing coefficient for each Gaussian, the sum of which should be 1.

In case the mean and variance are not known, the Expectation-Maximization (EM) algorithm is commonly used to estimate the assignment probabilities of a given data point to a particular data cluster.

Expectation Maximization

The expectation maximization (EM) algorithm is used to perform maximum likelihood estimation in the presence of latent (missing or unobserved) variables. It achieves this by first estimating the values for the latent variables, then optimizing the model, and repeating these two steps until convergence.

The goal is to maximize the log likelihood $lnP(D|\theta)$, $D = D_{good} \cup D_{bad}$ through the maximizing the expectation $\mathbb{E}_{D_{bad}}[lnP(D|\theta) \mid D_{good}, \theta^{(0)}]$. The Expectation step comprises of computing the expectation in this last expression, and the Maximization of updating the parameter $\theta^{(i)}$ at the ith iteration. In the EM algorithm the two steps of Expectation (E) and Maximization (M) are repeated iteratively until convergence is reached for the estimation of the latent variable $\theta^{(i)}$.

This missing values paradigm can be applied to the case where the actual missing information lies in the assigned cluster of the data, and through assuming a Gaussian or Gaussian Mixture distribution, EM can be used to predict their latent parameters.

2.2.2 Dimensionality Reduction

When our input features exceed a threshold of dimensionality, we often want to utilize methods to reduce these dimensions. This is mainly an effort to combat a phenomenon known as the curse of dimensionality, which is actually an umbrella term for many phenomena that occur when dealing with data in much higher-dimensional spaces than the 3d physical world. The most common of these problems is that as dimensionality increases, the volume of the space increases so fast that the available data become sparse. This has two main consequences, firstly that in order to ensure that the machine learning algorithm has "seen" enough examples of each combination of values during training, the amount of data needed needs to grow exponentially with the dimensionality. A standard rule of thumb is that there should be at least 5 training examples for each dimension in the feature representation, or the model will start overfitting the training data and fail in out-of-sample generalization. Secondly, clustering relies on distance measures such as the Euclidean distance to quantify the similarity between observations. If the distances are all approximately equal, then all the observations appear equally alike (as well as equally different), and no meaningful clusters can be formed.

The general idea of dimensionality reduction is to locate underlying trends in the features, combinations of which can describe the entirety of the samples. These underlying trends hidden in the features are often called latent features, and every other feature is re-written in terms of their exposures to these latent features.

PCA

We aim to describe our input features as linear combinations of a reduced number of components such that the distances, or variance, between the resulting samples is maximized. The input features are often highly correlated, so the greatest reduction in features will be achieved if each component is chosen to be uncorrelated to the others. These components are called the principal components, and this method the Principal Components Analysis (PCA), which is a non-parametric method. In essence, PCA creates a set of principal components ranked by variance which are uncorrelated, and end up low in number as lower ranked components are thrown away.

2.3 Supervised Learning

Supervised learning models are utilized for two main tasks, regression and classification, which we will analyse next. The goal of all supervised learning algorithms is to correctly model a mapping function f that projects an input feature space X to an output label space Y.

2.3.1 Regression

Regression is used to identify the relationship between a dependent variable (output label) and one or more independent variables (input features) and is typically leveraged to make predictions about future outcomes. The label here is a value in a continuous space, which the model is called to predict.

Linear Regression

As the name suggests, linear regression models have the goal of finding the optimal line that best describes the data points, a process described as "fitting“ the data. The key property of linear regression is that it is a linear function of its parameters w. In its simplest form, it is also a linear function with respect to the input variables x, but a much more useful class of functions can be obtained through applying a fixed set of non-linear functions to the input, known as basis functions.

$$y(\mathbf{x}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) \tag{2.1}$$

where ϕ_j are the fixed nonlinear basis functions ($\phi_0 = 1$) and M the total number of parameters in the model.

The error of estimating the curve for N data points can be described by

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}[t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x_n})]^2 \tag{2.2}$$

also known as the sum of squares error (loss) function, with respect to the ground truth values t_n. Another metric called mean squared error (MSE) refers to the unbiased estimate of error variance, which is the sum of squares divided by the number of degrees of freedom (usually N).

The optimal parameters $\mathbf{w}$ are those that minimize this function, which is called the least squares solution. To find the least squares solution for all the samples at once would be too costly in the case of large datasets. Sequential algorithms, also known as on-line algorithms, consider the data points one at a time or in small batches, and update the parameters after each batch.

As known from basic calculus, the values minimizing a differentiable function can be calculated through setting its gradient equal to zero and solving for the desired variables. We can obtain a sequential learning algorithm by moving in the direction of $-\nabla E$ one batch at a time, which is what is done in the technique of stochastic gradient descent, summarized in the update equation

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E \tag{2.3}$$

where η is a parameter called learning rate. This type of parameter that is external to the model's parameters is called a hyperparameter. The initial model parameters are set to a (usually) random starting vector $\mathbf{w}^{(0)}$.

If too many parameters M are used (and therefore number of basis functions), the curve that will be estimated from the training data will not leave any room for variability in the test data. This results in a generalization problem called overfitting, where the model ends up being too case-specific to the data it was trained on.

To overcome overfitting, we can add a regularization term to the error function, called a regularizer.

$$E_{total}(\mathbf{w}) = E(\mathbf{w}) + \lambda R(\mathbf{w}) \tag{2.4}$$

One of the simplest regularizers is weight decay, given by the equation

$$R(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} \tag{2.5}$$

It is called weight decay because it urges the weight values to decay close to zero.

A more general regularizer is given by the equation

$$R(\mathbf{w}) = \frac{1}{2}\sum_{j=1}^{M}|w_j|^q \tag{2.6}$$

which for q=1 gives the lasso regularizer (also called L1) and for q=2 the ridge regularizer (L2). Of these two, both minimize loss through minimizing the less important features w_j, but lasso can actually render $w_j = 0$ because even small values of w_j are enough to cancel out the error. This is useful in the case a kind of feature selection is desired.

Even though the family of linear models in which linear regression belongs to has significant limitations which we explore later, they form the foundation for many more sophisticated models, and ultimately all of neural networks.

2.3.2 Classification

Classification, as the term indicates, is the task of assigning a class label to each data sample, given a set of pre-defined classes. Each sample in the training set is paired with a ground truth label out of these classes. In the case that there are only two classes, it is called a binary classification task, and multi-class classification otherwise. It is similar to clustering in unsupervised learning, but with the difference of making an assumption on what the groupings are that we want to recognise in our data. A further categorization that can be made is between generative and discriminative classification models. In generative models, the distribution of the data is explicitly modelled, whereas discriminative models focus on optimizing an objective function to best discern between the classes, thus implicitly modeling the data in the underlying parameters.

To measure the success of a classification model we employ some evaluation metrics as well, apart from error metrics. To describe the most important of those briefly (for the binary task), given the TP (true positive), FP (false positive), TN (true negative) and FN (false negative) predictions: (A) Accuracy = (TP+TN)/(TP+FP+FN+TN), is the most intuitive and most used and it is simply a ratio of correctly predicted samples to the total samples, (B) Precision = TP/(TP+FP), is the ratio of correctly predicted positive samples to the total predicted positive samples, (C) Recall = TP/(TP+FN), is the ratio of correctly predicted positive samples to the total positive samples, and lastly (D) F1 score = 2*(Recall*Precision)/(Recall+Precision), which provides a better accuracy measure for imbalanced class distributions.

Bayes Classifier

Naive Bayes is generative classification approach that adopts the principle of class conditional independence from the Bayes Theorem.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{2.7}$$

This means that the presence of one feature does not impact the presence of another in the probability of a given outcome, and each predictor has an equal effect on that result.

K-Nearest-Neighbours

K-nearest neighbor, also known as KNN, is a discriminative non-parametric algorithm that classifies data points based on their proximity and association to their surrounding samples (neighbours). This method makes the assumption that similar data points can be found near each other. As a result, it aims to calculate the distance between data points, usually through Euclidean distance, and then it assigns a category based on the most frequent category (ground truth label) in the area.

The Perceptron

The Perceptron is the simplest type of neural network. It is also called a single layer neural network. It consists of a single neuron that models a linear combination of the input features, each x_i multiplied by a weight parameter w_i, which is then mapped to the output, which is 0 or 1 depending on whether a threshold b called bias is met. The input features are also often described as the input layer, although they are not modeling any transformation function (but the identity), and this is not added in the total count of layers in the neural network.

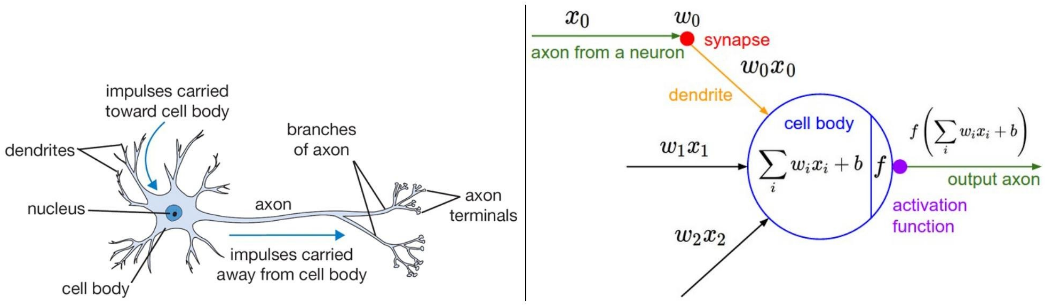


Figure 2.1. *The analogy of the artificial neuron (right) to the biological neuron (left). Source: cs231n.github.io*

These parameters w_i, b are updated through an iterative algorithm, with the goal of mapping the most inputs to the correct output. This is called the Perceptron Learning Algorithm, and it could be seen as the predecessor of the Gradient Descent which is the algorithm on which modern neural networks are build on. Starting by initializing the weights randomly, it updates them by iteratively going through every sample in the training set and for each misclassified sample, pushes the weights to the opposite direction of the false prediction by the amount of each sample. This is repeated until convergence. We can define the this algorithm mathematically as

$$\begin{aligned}\hat{y}_n &= f(\mathbf{w}^{(\tau)}\mathbf{x}_n)\\ \mathbf{w}^{(\tau+1)} &= \mathbf{w}^{(\tau)} + \eta(y_n - \hat{y}_n)\mathbf{x}_n\end{aligned} \tag{2.8}$$

which is repeated for all examples n in the training set.

The perceptron algorithm can be said to use a 0-1 loss function, where for each incorrect prediction you incur a penalty of 1 and for each correct prediction no penalty. If we were to assign classes in a soft manner, we could apply a transformation function like the logistic before doing the final classification, which we will introduce in the next section. This is called an activation function of the neuron, which in the perceptron is simply the linear (identity) function.

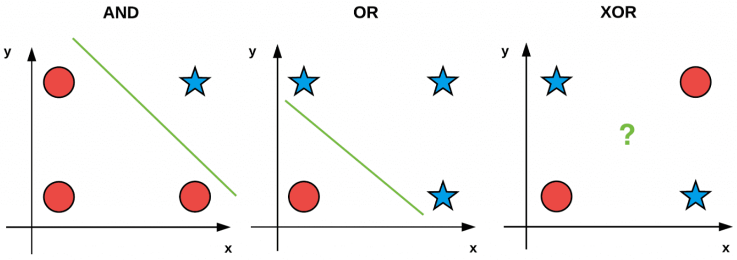


Figure 2.2. *The XOR analogy to the non-linearly separable data classification problem. A single neuron perceptron fails to solve this. Source: pyimagesearch.com*

A single neuron cannot possibly model non-linearly separable data, as it can only model a line function. To combat this, more layers need to be introduced, composing the multi-layer perceptron which is the first neural network and which we will introduce in section 2.3.3.

Logistic Regression

While linear regression is leveraged when dependent variables are continuous, logistic regression is selected when the dependent variable is categorical. Instead of fitting a linear function to the training data, now we use what is called a logistic function that models the probability that a sample falls in a category (class). For the binary classification case, this can be expressed as

$$p(\mathcal{C}_1|\boldsymbol{\phi}) = y(\boldsymbol{\phi}) = \sigma(\mathbf{w}^T\boldsymbol{\phi}) \tag{2.9}$$

with $p(\mathcal{C}_2|\boldsymbol{\phi}) = 1 - p(\mathcal{C}_1|\boldsymbol{\phi})$ and where $\boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{x})$ and

$$\sigma(a) = \frac{1}{1+e^{-a}} \tag{2.10}$$

is the logistic function. The resulting decision boundaries will be linear in the feature space ϕ of the fixed non-linear basis functions, and these correspond to non-linear decision boundaries in the original x space, which means that linearly separable classes in ϕ need not be linearly separable in the original observation space x.

We can write the likelihood function as $p(\mathbf{t}|w) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}$ for which we can define the error (loss) function from the negative logarithm

$$E(\mathbf{w}) = -lnp(\mathbf{t}|w) = -\sum_{n=1}^{N}[t_n lny_n + (1-t_n)ln(1-y_n)] \tag{2.11}$$

where $y_n = y(\phi_n)$ is called the cross entropy loss function.

Logistic regression is a discriminative model, and it leverages the optimization function of maximum likelihood, in the form of cross entropy loss to find the most accurate classification of the data. MSE loss does not apply in the case of logistic regression as using MSE means that we assume that the underlying data has been generated from a normal distribution, when actually categorical variables fall into the case of a Bernoulli distribution.

We can define a sequential algorithm with stochastic gradient descent like in linear regression.

In the case of multi-class classification, the logistic function turns into the softmax function

$$p(\mathcal{C}_k|\boldsymbol{\phi}) = y_k(\boldsymbol{\phi}) = softmax(a_k) \tag{2.12}$$

where

$$softmax(a_k) = \frac{e^{a_k}}{\sum_m e^{a_m}} \tag{2.13}$$

and the likelihood and cross entropy functions are generalized.

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} lny_{nk} \tag{2.14}$$

Support Vector Machines

Support Vector Machines are a type of discriminative classification algorithm. We will define an SVM classifier for the binary classification problem. In Support Vector Machines (SVMs), data points from different classes that lie the closest to each other define a set of support vectors. If we visualized the lines defined from each class's support vectors as the side-walk on the two sides of a street, the goal of the SVM is to make this street as wide as possible, and draw the line discriminative function right in the middle.

However this only describes the case of data that are linearly separable. In the case of non-linearly separable data, the SVM employs what is called the kernel function, which transforms the data in a higher dimension such that they are linearly separable. A commonly used kernel function is the RBF. To reduce the computational cost of this transformation, the kernel trick is used, which avoids calculating the kernel function for a single data point and instead depends on the product of the kernel functions for two different data points, this way computing a distance score instead.

2.3.3 Neural Networks

In Linear and Logistic Regression we described models for regression and classification that composed of linear combinations of fixed basis functions. However, due to the curse of dimensionality, if the basis functions are fixed before the training data set is observed, their number needs to grow exponentially with the dimensionality D of the input space. There are two main properties of real data sets that are leveraged by neural networks in order to solve this problem. First, due to often highly correlated input features, the data vectors can lie close to a non-linear manifold whose intrinsic dimensionality is smaller than that of the original space. Second, target variables may depend on just a small number of possible directions within that data manifold. Neural network models, through using adaptive basis functions, can make the regions of input space over which the basis functions vary correspond to the data manifold, as well as select the directions to which they respond.

The way these adaptive basis functions are formed, is that each basis function is a nonlinear function of a linear combination of the inputs, where the coefficients in the linear combination are the adaptive parameters. This requires a hierarchy of layers of parameters, which can be represented in the form of a network diagram, where the nodes (corresponding to parameters) are called neurons. In this section we will discuss the simplest neural network form that consists of only two layers of neurons.

$$y_k(\mathbf{x}) = f(\sum_{j=0}^{M} w_{kj}^{(2)} h(\sum_{i=0}^{D} w_{ji}^{(1)} x_i)) \tag{2.15}$$

The first neural network called the Multi Layer Perceptron can be also seen as a descendant of the perceptron. The perceptron was unable to handle non-linearly separable data. To solve this problem, the MLP introduces a second layer of neurons, apart from the single neuron output layer we saw in the perceptron, called the hidden layer. For example for the XOR problem, an MLP with a single hidden layer can be designed such that there are an OR and NAND hidden neurons, and an output AND neuron, composing the desired function.

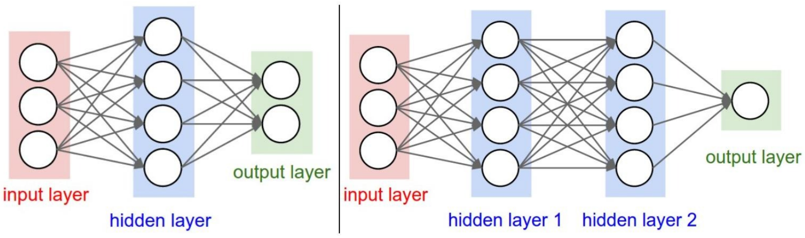


Figure 2.3. *Neural Network with single hidden layer (left) as well as two hidden layers (right). It can also be viewed as a Multilayer Perceptron, which succeeds in classifying non-linearly separable data. Source: cs231n.github.io*

In order to add the hidden layer, another necessary component for the Multi Layer Perceptron is a non-linear activation function. Without a non-linear transformation, the stacking of two neuron layers is simply a linear transformation of the input, and so equivalent to a single layer.

Neural network training is composed of the forward and the backward pass. The process of evaluating the final output y can be interpreted as a forward propagation of information through the network. As for the backward pass, in order to compute the gradients required for gradient descent, we arrive at the algorithm called backpropagation. Backpropagation is a way to calculate the gradients of the loss with respect to each of the model weights, through following the chain rule starting from the output layer all the way to the first layer until all components are known.

$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial a_j}\frac{\partial a_j}{\partial w_i} = \frac{\partial L}{\partial y_j}\frac{\partial y_j}{\partial a_j}\frac{\partial a_j}{\partial w_i} = \frac{\partial L}{\partial a}\frac{\partial a}{\partial y_j}\frac{\partial y_j}{\partial a_j}\frac{\partial a_j}{\partial w_i} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial a}\frac{\partial a}{\partial y_j}\frac{\partial y_j}{\partial a_j}\frac{\partial a_j}{\partial w_i} \tag{2.16}$$

In order to do this, we need a loss function that is differentiable, as opposed to the one used by the perceptron algorithm. As we saw in the logistic regression section, a good loss function for the classification task is cross entropy loss.

Additionally, neural network research has, over the years, explored other optimizers apart from stochastic gradient descent (SGD), other activation functions apart from the logistic, and other regularization methods apart from loss terms. More specifically, the Adam optimization algorithm (Kingma et al.) provides an extension to SGD, which computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients, and has been proven to be more efficient. In addition, Loshchilov et al. demonstrated that L2 regularization is significantly less effective for adaptive algorithms than for SGD, and proposed an improved version of Adam called AdamW. As for the activation functions, the logistic function suffers from the problem of vanishing gradients and thus cannot be used in networks with many layers, a problem which the Rectified Linear Unit (ReLU) overcomes by utilizing a (mostly) linear function (gradients remain proportional to the activations), given by the equation $a(x) = max\{0, x\}$. This is actually a non-linear function as negative values always output zero, allowing complex relationships in the data to be learned. As for the fact that it is not differentiable at zero, this is not a problem in practice, and the gradient is assumed to be zero there. Additionally it is less computationally expensive and can achieve representational sparsity. Finally, as regards different types of regularization, a dropout layer can be introduced which randomly sets input units to 0 with a given probability, helping to prevent overfitting, and the technique of early stopping is employed to perform model selection on a validation set according to some metric, and stop the training before overfitting occurs.

Neural networks are said to be universal approximators, meaning that a two-layer (single hidden layer) network with linear outputs can uniformly approximate any continuous function on a compact input domain, provided the network has a sufficiently large number of hidden units.

2.4 Deep Learning

In this section we will talk about the advances that achieved the breakthrough in leveraging huge amounts of data and left behind the era of expert knowledge and feature engineering. The area which we call deep learning was enabled by the creation of deep neural networks.

2.4.1 Feed-forward Neural Networks

In the Deep Learning world where there are many different kinds of architectures for neural networks, the vanilla neural network introduced in the previous section is most often called a feed-forward neural network (FFNN), or a fully connected neural network, to differentiate especially from cases where neural networks are either recurrent or sparsely connected, which we will examine in the next sections. In general, feed-forward neural networks are good in the case that the input features are independent with each other: not parts of a sequence, not points in a grid, because of the permutation invariance of the inputs in the neuron operation, i.e. relative position does not matter.

2.4.2 Convolutional Neural Networks

What would happen if we were to flatten all pixels of an image and give it as input to a FFNN? Some problems with this setup are that (A) it would require too many parameters (B) it would not be able to factor in pixel relative positions, as the order in a FFNN is permutable - which means that the network would not be able to recognize patterns. When thinking of a way to factor in those relative positions, there is an important property that we want to take under consideration. Vision, and more specifically the task of visual recognition, is governed by positional invariance - meaning the same object is recognised regardless on where it is placed on the visual field. It is intuitively obvious that it would be desirable to maintain this property in the neural network. The modeling of such a property in a neural network's design is called an inductive bias.

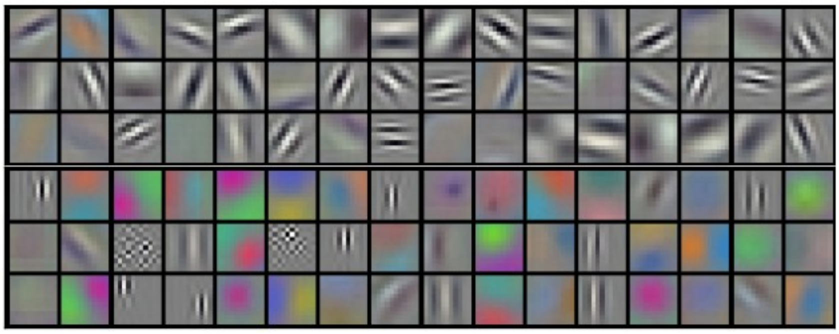


Figure 2.4. *96 filters learned by the first convolutional layer of AlexNet. Source: [4]*

The operation that manages to solve all of the above requirements at once, is convolution. The first paper to introduce Convolutional Neural Networks (CNNs) was [25]. This operation can be described as

$$(I * K)(i, j) = \sum_m \sum_n I(i + m, j + n)K(m, n) \tag{2.17}$$

where I is the image and K is a filter (kernel). The output of the convolution is often called a feature map. Strictly speaking, the above equation corresponds to cross-correlation, but it is equivalent to convolution in the context of CNNs.

When applying convolution we slide the filter through the image, each time shifting by a step-amount and restricting the receptive field to a small patch. This is what makes the CNN a sparse network, which only calculates weights while seeing a portion of the input at a time.

From signal processing we recall that we use convolution to describe frequency filtering in the time domain (when applying the fourier transform, convolution is turned into multiplication). What we want is to filter the input image with a range of filters - if this were traditional computer vision these would be pre-defined / engineered to produce specific results such as edge detection with the Laplacian of Gaussian (LoG) filter. Since in machine learning we want everything to be learned straight from the data, these filters will be learnable as well - so they make up some of the network's parameters. What filters essentially do, is to recognize the same pattern regardless of its absolute position.

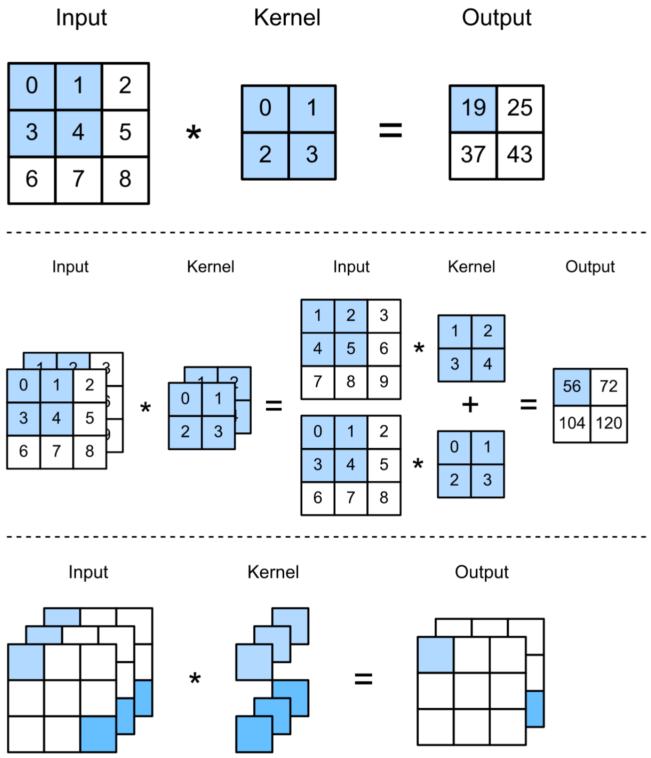


Figure 2.5. *Top: convolution with a single kernel on a single channel input. Middle: convolution with a single (3D) kernel on multiple channel input, resulting in single channel output. Bottom: convolution with multiple (3D) kernels on multiple channel input, resulting in multiple channel output. Source: d2l.ai*

We pass an image through multiple learnable filters - each of which extracts different kind of information - this information is much lower in dimension and higher in information density than the pure image pixels. To learn features at different scales, we can stack one convolutional layer after another - this way the filters in the first layers learn to extract more low-level and general patterns, such as edges, and the last more high-level, task-specific patterns, which are often less interpretable. Between convolutional layers, it is important to apply a non-linear activation function, just like between regular fully connected layers, to enable learning more complex patterns by stacking layers. Additionally, it is desired to reduce the dimensionality and increase the number of filters after each convolution layer, to enable learning a wider range of more high level (abstract) patterns as the layers progress. To do this, a pooling layer (most often max) is applied, which downsamples the input volume spatially and independently in each depth slice. This also leads to a reduction in the model parameters as we move to deeper layers which learn more details, which can also reduce overfitting. The features after the final layer can then be passed into a regular FFNN, followed by softmax for classification.

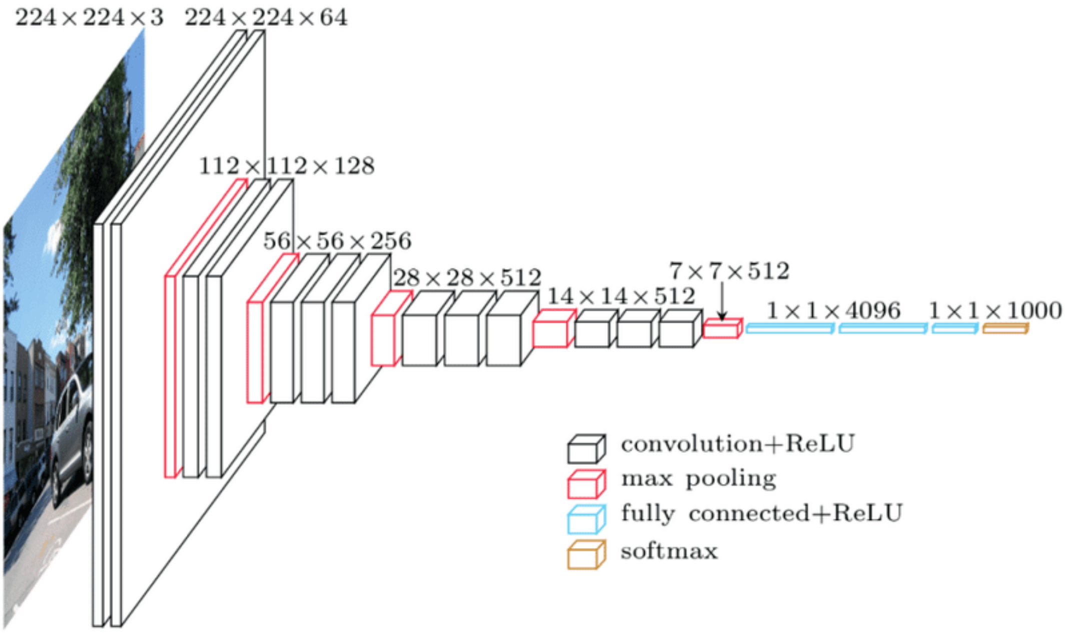


Figure 2.6. *VGGNet architecture: notice how the resolution of the output images decreases as the number of filters increases. Source: [5]*

2.4.3 Recurrent Neural Networks

Often we want to model sequential data, meaning variable length sequences such as text, speech, or any time series data. In addition, for this type of data we may want to make predictions based on context - meaning the inputs are not independent, and their order and proximity matters. Regular FFNNs have fixed sized inputs, and lack the notion of order (you can permute a sum anyway you want).

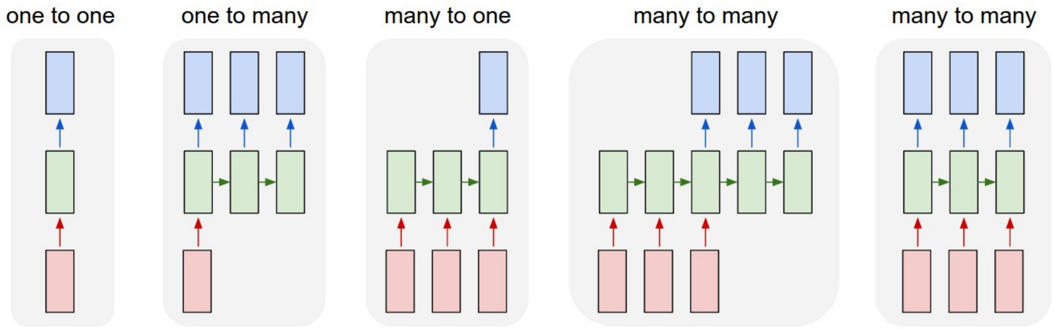


Figure 2.7. *The different sequential applications approached through the introduction of RNNs. Source: karpathy.github.io*

We need to allow previous outputs to be used as inputs (by making the neural network recurrent) in order to keep contextual information, called a hidden state. This allows for the model weights to be shared across time, and it can be described by the following equations

$$\begin{aligned} h_t &= \sigma(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \\ y_t &= \sigma(W_{hy}h_t + b_y) \end{aligned} \tag{2.18}$$

where h_t is the hidden state at timestep t and y_t the output.

Recurrent Neural Networks (RNNs) [26] can be used to encode sequences into a single vector (final hidden state), or to output another sequence. In general, sequence tasks can be divided in the following groups (figure 2.7); one to one (no sequence, e.g. image classification), one to many (e.g. image captioning), many to one (sequence classification), many to many (two kinds, generating future sequences or transforming the one to another - e.g. text generation and machine translation respectively).

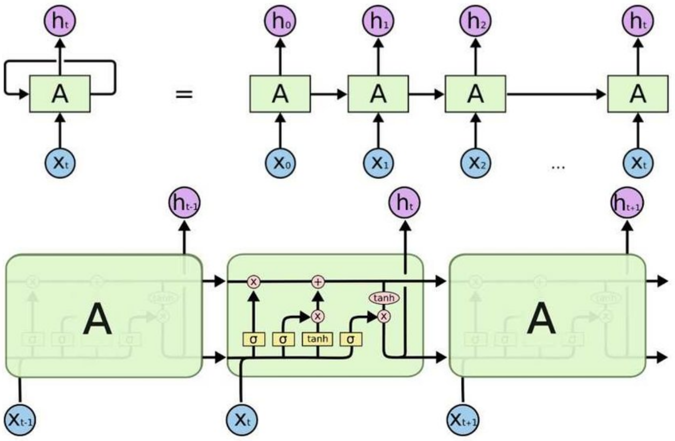


Figure 2.8. *RRN and LSTM comparison. Source colah.github.io*

Despite its effective modeling of sequential tasks, the vanilla RNN suffers from the problem of exploding/vanishing gradients. This is due to backpropagation through the (unfolded) feedback loop, which reinforces extreme values through repeated multiplication as sequences get longer. To solve this problem, another recurrent network was introduced, called the Long Short Term Memory network (LSTM) [27]. The key to the LSTMs is the cell state, which runs straight down the entire chain, with only some minor linear interactions. This enables information to flow through unchanged, while allowing to remove or add information only as regulated through gates. An LSTM has three of these gates, to protect and control the cell state.

$$\begin{aligned}
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\
\hat{c}_t &= tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
c_t &= c_{t-1} * f_t + i_t * \hat{c}_t \\
h_t &= o_t * tanh(c_t)
\end{aligned} \tag{2.19}$$

The forget layer f_t, learns what we want to forget from the previous hidden state and current input, which the forget gate applies it to the cell state through multiplication. Then, new information is added by the input gate, which is learned again through the input layer i_t. Finally, the output gate is used to produce the new hidden state, which is just like in the vanilla RNN except it is also conditioned on the current cell state we just produced. The LSTM models long term dependencies better than RNNs - but not so good still, as we will see more of that next.

2.4.4 Attention Mechanisms

In applications, recurrent neural networks often only use the final hidden state to model a sequence, to be used in a successive network. However this creates a bottleneck for very long sequences and fails to model long term dependencies, even with the optimizations we saw in the LSTM. This became especially evident in the task of neural machine translation, where the goal is to translate a sentence from one language to another. Because of the way the last hidden state is aggregated, the system pays more attention to the last parts of the sequence, corresponding to the current step at the translation. There is no way to factor in other positions, or give more importance to some of the input words compared to others while translating the sentence. For example, translations between languages like french and english inherently require to take into account words in different positions than the one at the current step - as the syntax imposes different order in many cases.

Attention mechanisms [28] are inspired by human attention and memory. Memory can be described as attention through time, so if a network with better memory is desired, it would make sense to explicitly learn on which inputs to factor in more - taking them all into consideration (with the restraint that the factors sum to 1). These factors are called attention weights, and we say that the model pays more attention on a sequential input step that has a bigger weight assigned to it. These weights are learned through a regular FFNN and they can be different for each output of a seq2seq model.

Let h_i, s_t be the hidden states of the input and output RNNs respectively. The input context vector to the output RNN will be defined as:

$$\begin{aligned} c_t &= \sum_i \alpha_{t,i} h_i \\ \alpha_{t,i} &= softmax(score(s_t, h_i)) \end{aligned} \tag{2.20}$$

where $score(s_t, h_i)$ is a learnable alignment scoring function. Some scoring functions proposed for attention can be seen in the table.

Additive [28]	$v_a tanh(W_a[s_t, h_i])$
Bilinear [29]	$s_t^T W_a h_i$
Dot-Product [29]	$s_t^T h_i$
Scaled Dot-Product [6]	$\frac{s_t^T h_i}{\sqrt{d}}$

Table 2.1. *Different attention scoring functions*

The specification of "soft attention" means that the function varies smoothly over its domain and, as a result, it is differentiable.

Attention was soon applied in many areas outside machine translation, including computer vision and multimodal translation tasks (e.g. image captioning) [30], where there is need for attention between image regions and words in a sentence. Vision research was directed away from grid-based features and bottom up processing and more towards region-based and top-down [31].

Note that very deep neural networks already learn a form of implicit attention, but attention is such a useful and important property that we want to insert it explicitly as an inductive bias to the model. This very insightful idea has proved groundbreaking in modern machine learning, as we will also see next.

2.4.5 Transformers

However, even with the use of attention in networks such as the LSTM, the modelling of long-term dependencies in the recurrent setting still leaves room for improvement. Another downside of recurrent networks with attention is that they are very computationally expensive, due to back-propagating for every recurrent step and input/output score. One of the main novelties introduced in the Transformer [6] architecture is to replace recurrency with positional encodings on fixed size, feed-forward input sequences, and compute attention between different parts of this same vector, thus bypassing both the vanishing gradient and the computation cost problems caused by back-propagation in recurrent networks. In addition, the Transformer lends itself to parallelization, due to the multiple heads mechanism it employs which we will explore shortly.

Self-attention via Scaled Dot-Product

Self-attention, also known as intra-attention, is an attention mechanism relating different positions of a single sequence in order to compute a representation of the same sequence. Borrowing terminology from information retrieval and databases, we define the Scaled Dot-Product attention as the alignment scoring (via dot-product) on a series of keys (K) by a series of queries (Q), followed by softmax and application of the resulting weights on a series of values (V) to compute the output context representation. These keys, queries, and values are learnable linear transformations of the input vectors X - it is in this parametrization of the K, Q, V that the attention distribution is learned.

$$\begin{aligned} &K = W_K X, \quad Q = W_Q X, \quad V = W_V X \\ &C = softmax(\frac{QK^T}{\sqrt{d_K}})V \end{aligned} \tag{2.21}$$

In the context of neural machine translation, each row in the query matrix corresponds to each next-to-be translated word, for which we score the alignment to each of the rows in the key matrix. Through the parametrization of the Q and K, the highest scoring value will not necessarily be assigned to the row corresponding to the same input in the sequence, but will rather be distributed to the rows in V in the way that helps towards the most accurate translation. Note that the division of the dot-product by the square root of the keys' dimension (scaling) is performed for stabilizing the gradients in the case that the dot-product becomes too large in larger dimensions d_K.

Multiple attention heads

In the attention mechanism of the Transformer, not only one set of (W_K, W_Q, W_V) matrices are learned, but a multitude of h such matrices, specifically 8 in the original paper. These are called “attention heads” and the resulting context matrices from computing self-attention with each of them are concatenated and linearly transformed to form a single context matrix.

$$\begin{aligned} &C_{multihead} = \text{Concat}(C_1, ..., C_h)W_O \\ &\text{where } C_i = softmax(\frac{Q_i K_i^T}{\sqrt{d_K}})V_i \end{aligned} \tag{2.22}$$

This idea of learning multiple attention distributions intuitively attributes to the notion that attention in the human and artificial neural network is meant to have specializations; for example we can have a separate attention mechanism for figuring out the right pronoun to use to address someone based on a sentence, and a separate for knowing how that sentence impacted us emotionally. This can be seen as giving the model the ability to jointly attend to information from different representation subspaces at different positions [6]. From another perspective, multi-head attention is also beneficial because it is a form of model ensembling, which also helps performance by itself.

Positional encoding

We still haven't addressed how the order of the words is modelled in the transformer - how did it manage to dispense with the recurrent architecture? The information of absolute and relative positions of the tokens must be encoded in their vector representations. This can be done via summation of a position-representative vector to each input embedding, called the positional encoding. This vectors' function over the inputs must both be periodic and have variable frequency, to model both relative and absolute distance. The following sinusoidal function covers these requirements, and additionally gives the advantage of being able to scale to unseen lengths of sequences.

$$\begin{aligned} PE(pos, 2i) &= sin(pos/10000^{2i/d}) \\ PE(pos, 2i+1) &= cos(pos/10000^{2i/d}) \end{aligned} \tag{2.23}$$

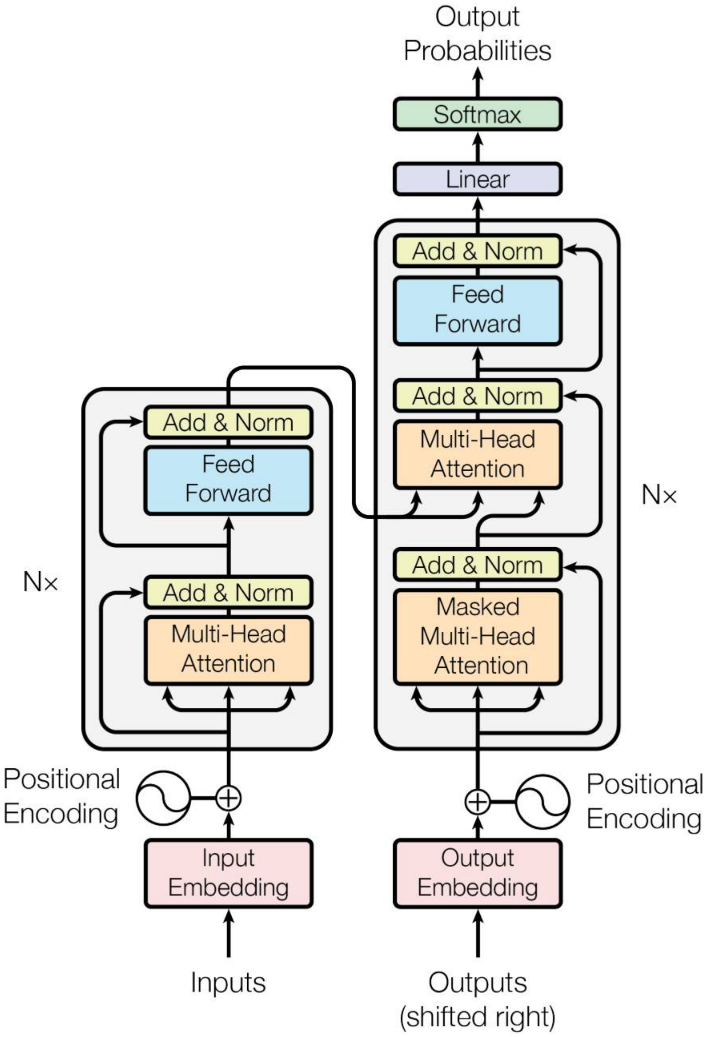


Figure 2.9. *The Transformer architecture. Source: [6]*

Encoder-Decoder architecture

Preceding both the encoder and the decoder side of the model, inputs are initially embedded via a trainable embedding layer, before being summed with the positional encoding.

The encoder produces an attention-based representation for each input token in the sequence. It consists of 6 identical blocks of two submodules, the multi-head self-attention layer and a point-wise fully connected feed forward network, where point-wise means that the linear transformations of the elements in the sequence share the same weights, similar to a convolutional filter of size 1. Additionally, each submodule contains layer normalization and a residual connection which can help preserve useful information from previous layers.

The function of the decoder is to retrieve information from the encoded representation, and based on that, output a new sequence. The architecture is almost exactly the same as the encoder, with two key differences. First, there is an additional multi-head attention submodule in each repeating block called the "Encoder-Decoder Attention" layer, which functions the same as the multi-head self-attention, except that it takes the Keys and Values matrices from the output of the encoder stack, and the Queries from the layer below it. Second, the multi-head self-attention layer is only allowed to attend to earlier positions in the output sequence, through masking future positions before the softmax step (-inf before the softmax, so that it becomes a zero after).

The output layer is a simple fully connected neural network that projects the vector produced by the stack of decoders into the logits vector, which is much larger in size - the same as the size of the model's vocabulary. After a softmax layer, the word with the highest probability is chosen as the output for the current time step. It is interesting to notice that the overall operation of the transformer layers does a form of squeezing and expanding in the dimensions, as the creation of the QKV matrices squeezes, and the output linear expands. This is often performed when learning intermediate representations in neural networks in general, for example CNNs also reduce dimensionality dramatically, only to expand to a million-class logit vector.

2.5 Transfer Learning

To describe the subject of transfer learning, we first need to introduce the concepts of domain and task. Plainly speaking, the task is the objective a model aims to perform, e.g. recognize objects in images, and the domain is where the data is coming from, e.g. daylight images from security cameras. More formally, a domain D consists of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$ over that feature space, where $X = \{x_1, x_2, ..., x_n | x_i \in \mathcal{X}\}$. Given a domain $D = \{\mathcal{X}, P(X)\}$, a task T consists of a label space Y and a conditional probability distribution $P(Y|X)$ to be learned from the training data. Ideally, if we want to train a model to perform a task on some target task and domain, it is natural that we would want to have training data from the same domain. If we want to use supervised learning, the training data for this task also has to be labeled. This stands in contrast with the abundance of diverse target domains that require machine learning solutions, for which it is near impossible to have enough data, let alone labeled data. Note that the domain can change very easily if even a fraction of the data collection conditions change, e.g. time of day, location, traffic.

Therefore, in reality, most real world ML use cases are only enabled if it is possible to make predictions on the target tasks, having done most of the training on a different (source) domain and task. In the general case, without any modifications most models will fail to do this successfully. How little these conditions need to change for a machine learning model to be thrown off, is in direct relation to its generalizability capacity. Transfer learning is the research subject that studies how to transfer as much common knowledge as possible between different domains and tasks, i.e. to use whatever might be useful to the target domain or task and throw away domain or task specific biases. Given a source domain D_S, a source task T_S, a target domain D_T, and a target task T_T, the objective of transfer learning is to learn the target conditional probability distribution $P(Y_T|X_T)$ in D_T with the information gained from D_S and T_S where $D_S \neq D_T$ or $T_S \neq T_T$. Since both the domain and task are described by tuples of two elements, there are two ways in which each of the above inequalities can manifest, namely the domains having completely different feature spaces, or just different distributions over that space (e.g. in text, different languages vs. just different topics), and respectively the tasks having completely different labels or just different distributions of labels in the data.

In most cases, a limited number of labeled target samples, which is much smaller than the number of labeled source examples are assumed to be available. The above conditions lead to the idea that, given a large amount of labeled data in a general source domain and task, we could first train models there, and then use the resulting weights to initialize models for training in smaller, more specific datasets. These steps are called pre-training (in the source dataset) and fine-tuning (in the target dataset). When fine-tuning, it is common practice to disable weight updates on some layers, a process called "freezing", so that the model does not forget useful information from the source dataset, as well as to use much smaller learning rates for the same reason. In addition, it is also common to use the pre-trained model as a feature extractor, for input features in another model. If the target task in which we fine-tune the model is different from the source task, it is also called a "downstream task", and the source task a "pre-training task"

Assuming models that are in the form of repeated identical blocks, as it was the case with CNNs and Transformers, if the source dataset is large and general enough and the model respectively deep enough, the features learned by the first layers of blocks tend to be more general, and the layers closer to the final prediction obtain more task-specific customizable representations. For this reason, the first layers are the ones that we freeze, so as not to forget the useful, general representations, and the last layers the ones that are fine-tuned.

2.5.1 The Imagenet Revolution

In order to motivate the application of transfer learning in the field of computer vision (CV) we must understand what accounts for the outstanding success of large convolutional neural networks on ImageNet [32]. Imagenet is the first large scale visual database for object recognition including more than 14 million hand-annotated images of more than 20,000 categories with a typical category consisting of several hundred images. Models trained on ImageNet seem to capture relevant and general features about the structure and composition of animals and objects. As a result, the ImageNet task seems to be a good proxy for general computer vision problems, as the same knowledge required for it is also applicable to a wide range of other tasks. Useful representations that capture broad information about how an image is built and the combinations of edges and shapes it includes are usually stored in one of the final convolutional layers or early fully-connected layers in large convolutional neural networks trained on ImageNet. A brief overview of the different CNN architectures that have succeeded in this task is given as follows. **AlexNet** [4] was the first convolutional neural network to do well on Imagenet, due to its bigger depth of 8 layers, utilization of GPU processors, and use of the ReLU non-linear activation function. **VGGNet** [5] increased the number of layers to 16 and 19 layer variants, and utilized the smallest possible receptive fields (3x3 convolutional filter) which was what enabled it to have more weight layers. **ResNet** [33], is much deeper and includes several variations in the order of hundreds of layers. This was made possible by utilizing skip connections or residual connections to jump over some layers, in this way preserving base information in a very deep network and preventing phenomena such as the vanishing gradient. ResNet is also the first CNN to use the technique of Batch Normalization. Finally, in **DenseNet** [34], extending the idea in ResNet, each layer obtains additional inputs from all preceding layers and passes on its own feature-maps to all subsequent layers. Since each layer receives feature maps from all preceding layers, the network can be more compact thus achieving higher computational and memory efficiency.

A slightly different task than object recognition is that of object detection which additionally requires to localize the predicted classes in the image and often recognize multiple different objects in a single image. An important class of models for this task are the region based CNNs (R-CNNs) which include a series of improvements, namely the Fast R-CNN [35] and the Faster R-CNN [36]. The vanilla **R-CNN** [37] performs forward propagation by using a CNN per region proposal from the input image, extracting features to be used for predicting its class and bounding box. Although the R-CNN model uses pretrained CNNs to extract image features, for thousands of region proposals from a single input image the computing load makes it infeasible to use in real-world applications. However, since these regions usually have overlaps, independent feature extractions lead to much repeated computation, and so the **Fast R-CNN** [35] performs a single CNN propagation on the entire image. To improve its accuracy the Fast R-CNN model usually has to generate a lot of region proposals in selective search, which **Faster RCNN** [36] combats through replacing the selective search with a region proposal network which is jointly trained with the rest of the model including its loss over proposals in the overall objective function.

Due to the scarcity and general nature of large pre-training datasets such as Imagenet, as well as the huge depth of the models required for state-of-the-art performances, very few people train an entire CNN from scratch. Utilizing transfer learning, it is common to pretrain a CNN on a very large dataset, and then use it either as an initialization or a fixed feature extractor for the task of interest. This has revolutionized the ease at which the community can leverage the capabilities of these deep networks.

2.5.2 The BERT Family

NLP background

In order for machine learning models to process words, these first need to be represented numerically in vectors that can be used in the models' calculations. These vectors are called word embeddings. In the beginning, these had the simplest form possible, meaning a one-hot vector the size of the known vocabulary, with the value 1 at the index corresponding to the current word and all the others 0. Alternatively, in a process still similar to one-hot embeddings, the mathematical significance of words in documents can be reflected as well, through TF-IDF (Term Frequency - Inverse Document Frequency) embeddings. In those, the value of the current word corresponds to its significance in the document (instead of 1), which is computed as the product of the Term Frequency and the Inverse Document Frequency metrics. The former corresponds simply to the frequency of a word in the current document, i.e. the ratio of the word's instances over the total words in the document, while the latter is the logarithm of the ratio of the total number of documents to the number of documents in which the word occurs, and measures how rare the word is.

However, apart from being inefficient due to their extreme sparsity, these sparse representations are also inaccurate as they fail to model the relationships between words. For example, the words "flower" and "plant" will correspond to different indices which will make their vectors' cosine similarity equal to zero, which conceptually is not the case. In Word2Vec [38], word representations manage to capture semantic as well as syntactic relationships, for example that "king" is to "man" what "queen" is to "woman", as well as that the relationship between "had" and "has" is the same as between "was" and "is". There were two major pretraining tasks employed in those language models for computing word representations, continuous-bag-of-words (CBOW) and skipgram. In a CBOW model, a target word is predicted given its surrounding context in the form of a "bag of words" where order does not matter. The skipgram model has the reverse goal, which is to predict the context of a word based solely on its representation. This is the first attempt in self-supervised learning, which we will delve more into in the next section.

After the Word2Vec and other similar approaches such as Glove vectors, the need arose to have different representations for the same word used in different context. For example, the word "arm" has different meaning when referring to the part of the body, and different when referring to weapons or equipment. ELMo [39] uses bi-directional LSTMs to create contextualized word-embeddings, by taking the entire sequence into consideration before assigning an embedding to the word. The LSTMs are trained on predicting the next word in a sequence of words, a tasked called Language Modelling (LM).

However, to model long term dependencies better, an attention mechanism would need to be inserted, which as we saw is costly in recurrent language models. Even the simple LSTMs' computational inefficiency is enough to discourage pre-training in huge enough datasets (the level of Imagenet or more, such as training on the whole internet text) to achieve successful transfer learning. This gave rise to the idea of solving language modeling the same way as seq2seq problems, through the idea of self-attention. In addition, the task of next word prediction (classic LM) reminds us of the transformer decoder, which is exactly what the openAI[1] transformers in the GPT [40, 41] series consist of - stacked decoder layers (only without the encoder-decoder attention). One can notice that in this setup, the only difference the decoder has from the encoder is the masking of the future tokens, as well as the fact that using stacked decoder layers seems to have one important drawback compared to ELMo - the context used is uni-directional, again due to the masking of future tokens. This gave rise to the idea of BERT [3], which we explore in depth shortly.

Self-supervised learning

Even with the advantages of transfer learning, labeling huge amounts of data for the pre-training task is still an unresolved issue which has a human bottleneck - while at the same time unlimited unlabelled data is being generated constantly. However, unsupervised learning is often difficult and less effective than supervised learning. Self-supervised learning is a learning scheme that empowers us to exploit a variety of labels that come with the data for free, by framing a supervised learning task in a way that uses a portion of the data as labels to be predicted from the rest of the data. This self-supervised task can sometimes be unrealistic but is chosen in such a way that it drives the model to the desired direction. The final performance of this invented task is of no importance, rather we are interested in the learned intermediate representation with the expectation that it can carry good structural biases and can be beneficial to a variety of practical downstream tasks. One of the first uses of self-supervised learning was in language modeling.

[1] https://openai.com/

BERT (Bidirectional Encoder Representations from Transformers)

BERT [3] uses the encoder side of the transformer, of which the main advantage compared to using the decoder side with the classic LM task (like in GPT [40]), is that the constraint for using only past inputs is lifted and the representations' context is bidirectional. It is not unexpected that a representation that learns the context around a word rather than just before the word is able to capture its meaning more accurately, both syntactically and semantically.

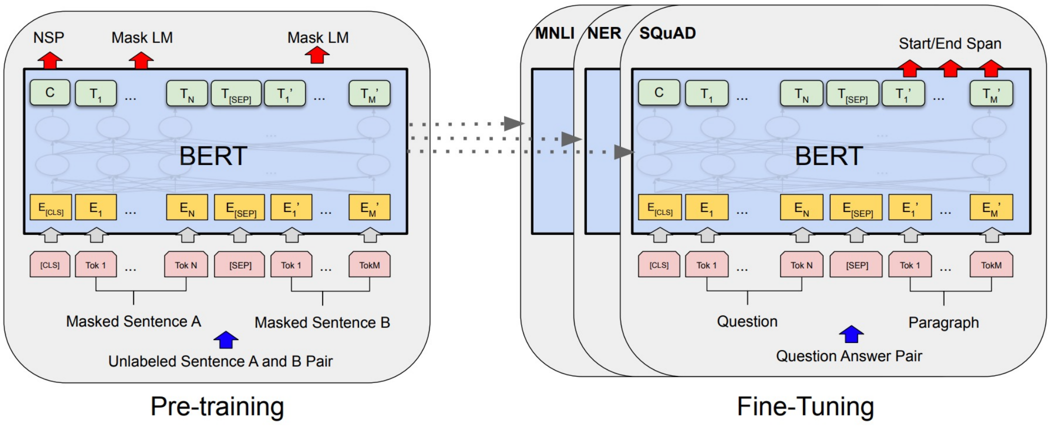


Figure 2.10. *Comparison of BERT pre-training versus fine-tuning configurations. The downstream task shown on the right is the task of extractive question answering, where the model outputs a start/end span prediction. Source: [3]*

This is achieved by setting a bidirectional LM task, instead of the classic LM, called Masked Language Modeling (MLM), which is what was called the "cloze" task in earlier literature. The idea is to randomly mask 15% of tokens in each sequence, by replacing them with a spacial token [MASK], to be used for prediction. The output size is only 15% of the input size. However, this token would never be encountered in fine-tuning which would make the distributions of the tasks too different. To overcome this a trick is employed, namely, of the chosen random tokens to mask

- 80% are replaced with [MASK]
- 10% are replaced with a random word
- 10% are kept the same

Furthermore, an additional task is introduced to promote sentence-level understanding, as many downstream tasks involve the understanding of relationships between sentences. The objective of this task is to perform binary classification on whether one sentence is the next sentence of another, and it is called the Next Sentence Prediction (NSP) task. Sentence pairs (A, B) are sampled so that:

- 50% of the time, B follows A
- 50% of the time, B does not follow A

The two sentences A and B are separated by a special token [SEP], and concatenated as input to the model. Before sentence A, a special [CLS] token is inserted, which is later used as input to the final classifier for prediction. The objective of this classification token is to learn the aggregated representation of the inputs that is useful for classification.

BERT's total pre-training loss is the sum of the mean MLM likelihood and mean NSP likelihood. BERT also introduces some novelties in the way the input word embeddings are formed, compared to the transformer. More specifically, they are formulated as the sum of three parts:

1. WordPiece tokenization embeddings: Instead of using tokenization at word level, words are further divided into smaller sub-word units in order to handle words with common root or rare words more effectively.

2. Segment embeddings: If the input contains two sentences, embeddings that denote on which sentence a token belongs in are added.

3. Position embeddings: Positional embeddings are learned rather than hard-coded as it was in the vanilla transformer.

As for BERT's downstream tasks (i.e. the tasks that the model weights learned in the pre-training tasks of MLM and NSP can be fine-tuned for), these correspond to a very wide range, including but not limited to (1) entailment, (2) extractive question answering, and (3) text classification or sequence tagging. At the output, the token representations are fed into an output layer for token level tasks, such as sequence tagging or extractive question answering, and the [CLS] representation is fed into an output layer for classification, such as entailment or sentiment analysis [3], as well as multiple-choice question answering.

Succeeding the release and widespread adoption of BERT, a multitude of variations and improvements over its vanilla architecture were proposed, such as RoBERTa [42] and DistilBERT [43], which are two of the most popular. RoBERTa [42] (stands for Robustly optimized BERT approach), is a retraining of BERT with optimized training methodology and much more data and compute power. RoBERTa's optimizations include the removal of the Next Sentence Prediction (NSP) pre-training task and the introduction of dynamic masking which makes the masked token change during the training epochs. In addition it recommends the use of larger batch sizes. In an opposite direction of the optimizations in RoBERTa, it is often needed to reduce computational costs and use a smaller network to approximate the performance. DistilBERT [43] learns a distilled (approximate) version of BERT, retaining 97% performance but using only half the number of parameters. It employs a technique called distillation, which approximates a trained large neural network's output distributions by a smaller network, through posterior approximation and the use of Kulback Leiber divergence in the optimization function. In addition, it omits token-type embeddings, pooler layers and half of the layers from the vanilla BERT.

Chapter 3

Multimodal Machine Learning

3.1 Introduction

Human perception and understanding of the world involves multiple sensory modalities such as vision, hearing, smell, taste, and touch. The word "modality" refers to the source of a signal, whether this is a measurement from the physical or the digital world, and machine learning research problems that are referred to as "multimodal" require the use of multiple such sources. For example, an inherent multimodal source is video, as it includes image, sound, and often text.

Thus, artificial intelligence systems need to be able to process and capture interactions between these modalities in order to gain an in-depth understanding of their environment. From early research on audio-visual speech recognition to the more recent explosion of interest in vision and language models, multi-modal machine learning is a vibrant multi-disciplinary field of increasing importance and with extraordinary potential [7]. The goal of multimodal machine learning is to design models that can process and relate information from multiple modalities, often through learning a common joint representation.

It is common that multimodal machine learning involves multiple disciplines of machine learning such as Computer Vision and Natural Language Processing, as each of those deal with a different modality; in this case vision and language, a combination which we will specifically analyse in depth in this chapter. An important theme in multimodal machine learning revolves around the fact that different modalities are also characterized by different statistical properties. In this case, images are mainly represented in the scale of pixels which contain dense and diverse information while text is represented through much sparser word embeddings.

The main challenges posed in the multi-modal setting, and consequently the direction of the methods approaching them, can be divided in five categories following Baltrusaitis et al.'s [7] taxonomy: representation learning, fusion, alignment, translation, and co-learning. In this chapter we will focus on techniques in the fusion, alignment, and translation domains, as well as focus on deep multimodal learning, leaving aside earlier classical machine learning multimodal approaches. It has to be noted that in deep learning the representation learning technique is a subset of fusion (early and mid-level approaches), where the main goal is to compute meaningful joint representations.

3.2 Techniques

3.2.1 Fusion

The goal of multimodal fusion is to predict the task's target variables through joining information from multiple modalities, which may differ in their predictive power. Fusion methods are often divided in early (i.e. at the input feature level), late (i.e. after the uni-modal predictions), and mid-level (i.e. at the intermediate representation level). Another fusion type mentioned in literature is hybrid fusion, which combines outputs from early fusion with uni-modal outputs in a late fusion manner [44].

Early fusion often only requires the concatenation of the input features, followed by the training of a single model, making the pipeline easier compared to late and hybrid approaches. It also consists the simplest kind of joint representation learning, as it can learn to exploit the correlation and interactions between low level features of each modality. In contrast, late fusion fuses the uni-modal predictions through methods such as averaging [45] and voting-schemes [46]. An advantage compared to early fusion is that it enables modelling each modality separately, which is often better due to the statistical differences in their content, at the cost, however, of ignoring the low level interactions between modalities.

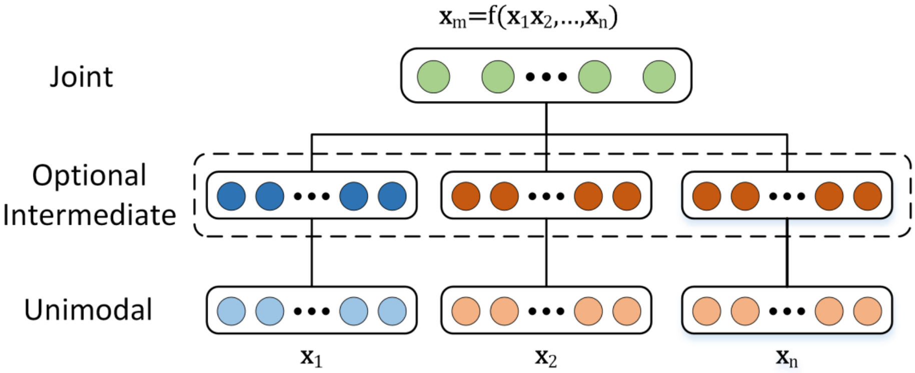


Figure 3.1. *Joint representation learning through early or optionally mid-level fusion. Source: [7]*

The problem of learning multimodal joint representations, also described as the projection of uni-modal representations together into a multimodal space, is most generally modelled by the case of mid-level fusion, and is expressed by figure 3.1, and the equation $\mathbf{x}_m = f(\mathbf{x}_1, ..., \mathbf{x}_n)$, where the function f is often modelled by a fully connected neural network (FFNN) layer. More specifically, when creating multimodal joint representations with mid-level fusion in neural networks, each modality is first processed separately with one or more layers, followed by a layer that projects the modalities into a joint space [23, 47, 48, 49]. The joint multimodal representation can then be passed through another network or used directly for prediction.

3.2.2 Alignment

The goal of alignment is to identify the direct relations between parts of the input from multiple modalities, for example the areas of an image corresponding to a text caption's words. This is often achieved through employing similarity measures. There are cases where the multimodal downstream task also corresponds to the alignment problem and the supervision is provided exactly for those targets, in which case we call that a problem of explicit alignment. In most cases however, multimodal alignment will be used as an intermediate or latent step for another task, where there are no supervised alignment examples and is referred to as an implicit alignment problem.

Translation, which we will delve more into in the next section, can also be formulated to contain alignment as an intermediate step, where in the case of an encoder-decoder model this is needed in order for the encoder module to not be required to summarize the whole input into a single vector representation. This implicit alignment is often achieved through employing attention mechanisms, which compute a soft score of importance on the input, and allow the decoder to focus on more important tokens, as we saw in the dedicated section of the previous chapter. Implicit alignment through attention is also commonly applied to the question answering task, either uni-modal or multi-modal, as it enables the alignment of the question words with part of an information source such as text, image, or video.

3.2.3 Translation

Translation deals with the problem of mapping data from one modality to another, which are heterogeneous to each other as well as often have an ambiguous, or subjective relationship. For the example of video captioning, there are several correct ways to describe the events and content in a video, which can differ in terms of perspective and focus - e.g. they can be more object or human centered. Multimodal translation is a long studied problem, with early work in speech synthesis [50], visual speech generation [51], video description [52], and cross-modal retrieval [53]. An especially popular problem is visual scene description, which includes image [54] and video captioning [55], where not only detecting the salient parts in a visual scene is required, but also to generate linguistically correct and comprehensive sentences describing them.

Translation techniques can be categorized in the example-based and generative types. The former use a static dictionary for the connection between modalities, while the latter train a model to predict the translation output. Generative models are much more difficult to build since they require an added capacity to generate sequences successfully, with temporal and structural consistency. An example of a very prominent recent generative multimodal translation model is the DALL-E series [56, 57]. On the other hand, example-based algorithms are constrained by their training data and dictionary, but are much easier to design and apply. They can be further categorized in retrieval-based and combination-based algorithms where the first directly use the retrieved translation without further processing, and the second rely on more complex rules to create translations based on several retrieved instances.

3.3 Vision and Language

Vision and language are two of the most fundamental and most remarkable capabilities of the human mind, since vision enables the creation of mental concepts that would not otherwise exist such as colour and light, and language has the power to distill all sensory experience to these elementary mental concepts and combine them to create complex new ideas, thus facilitating thinking through making an “infinite use of finite means” [14]. Humans routinely perform tasks through the interactions between vision and language, supporting the uniquely human capacity to talk about what they see or hallucinate a picture on a natural-language description.

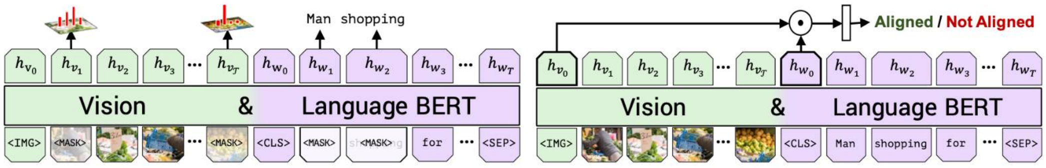
The interaction of language with vision motivates researchers to identify the relations between the modalities, combine and reason about them for decision making. Language can inadvertently impose strong priors since it is easier to get statistical information from (due to the sparse granularity of the linguistic domain), which give it a tendency to be an easier signal to learn from than the visual modality, resulting in vision-and-language models that completely disregard visual information in favor of exploiting language biases, leading to an exaggerated sense of their capabilities and no real understanding of the visual content. There is no clear solution to this caveat, and it has been a direction for research in multiple works, while several datasets are constructed in such a way that models that rely mostly on uni-modal signals will perform poorly [13, 58, 59].

Vision and language tasks can be categorized in three major areas. (A) Generation tasks, for example in image captioning text descriptions are generated for a given visual input, and in text-image generation visual output is generated from a textual input. (B) Classification tasks, for example in multiple-choice Visual Question Answering the correct answer to a question is chosen given a visual input, and in Visual Entailment statements regarding a visual input are classified as correct or incorrect. (C) Retrieval tasks, for example in image retrieval images are retrieved based on a textual description. These tasks challenge systems to understand a wide range of detailed semantics of an image, including objects, attributes, spatial relationships, actions and intentions, and how all of these concepts are referred to and grounded in natural language.

Vision and language is a domain with many end tasks, such as visual question answering, captioning, image retrieval, visual grounding, and more. This makes it ideal to apply the logic of transfer learning with general self-supervised pre-training tasks, and multiple downstream tasks like we saw in BERT. This gave rise to a new family of BERT-based models, that enhances the language model setup with the visual modality, through approaches that include image object region features from CNN backbones together with the word tokens and invent masking tasks for those as well. In this next section we proceed to describe these general purpose vision and language models, and introspect into their inner workings, through our own analyses and implementations.

3.3.1 Vision and Language in the BERT family

In the original transformer configuration for language, the concept of self-attention is introduced. Here, this is extended for vision as well as cross-modality relations, giving rise to two self-attention mechanisms and a cross-attention mechanism. This cross-attention configuration is actually a way to perform the technique of modality alignment that we saw earlier, as it scores an explicit weight for each combination of modality tokens. In addition, these models compute joint representations for the modalities, through aggregating useful information during the unsupervised pretraining tasks (such as masking). Models can be roughly categorized into single-stream and two-stream architectures, where in the first visual and text features are processed by a single transformer model, while in the two stream by separate ones. In the first case, cross attention is implicit and hidden inside the token's self-attention, while in the second case it consists of an explicit separate attention module.


Man shopping
h_{v_0} h_{v_1} h_{v_2} h_{v_3} ··· h_{v_T} h_{w_0} h_{w_1} h_{w_2} h_{w_3} ··· h_{w_T}
Vision & Language BERT
<IMG> <MASK> ··· <MASK> <CLS> <MASK> <MASK> for ··· <SEP>
Aligned / Not Aligned
h_{v_0} h_{v_1} h_{v_2} h_{v_3} ··· h_{v_T} h_{w_0} h_{w_1} h_{w_2} h_{w_3} ··· h_{w_T}
Vision & Language BERT
<IMG> ··· <CLS> Man shopping for ··· <SEP>

(a) Masked multi-modal learning

(b) Multi-modal alignment prediction

Figure 3.2. *The different multimodal pretraining tasks in the ViLBERT model. Source: [8]*

First, some important single-stream models are the following. In VisualBERT [60], image features extracted from object proposals by Faster-RCNN are viewed as unordered input tokens and fed into a BERT-like transformer encoder together with the text in a single stream. Multiple layers process the text and image inputs together, and the joined representations are trained with two pre-training tasks, Masked Language Modeling with the image without masking image region features, and Sentence-Image Prediction. In UNITER [61], a single-stream model with more pretraining tasks is introduced, namely Masked Language Modeling (MLM), Masked Region

Modeling (MRM) with three variants, Image-Text Matching (ITM), and Word-Region Alignment (WRA). In contrast to other similar methods that use random masking, authors propose conditional masking on pre-training tasks. In Unified Vision-Language Pre-Training for Image Captioning and VQA [62] authors propose a shared transformer network for both encoding and decoding, in contrast to most existing methods that contain separate encoder and decoder models. It is trained with two unsupervised pre-training tasks that differ only in the context on which the prediction is conditioned on, Bidirectional and Sequence-to-sequence (seq2seq) Masked Vision-Language Prediction.

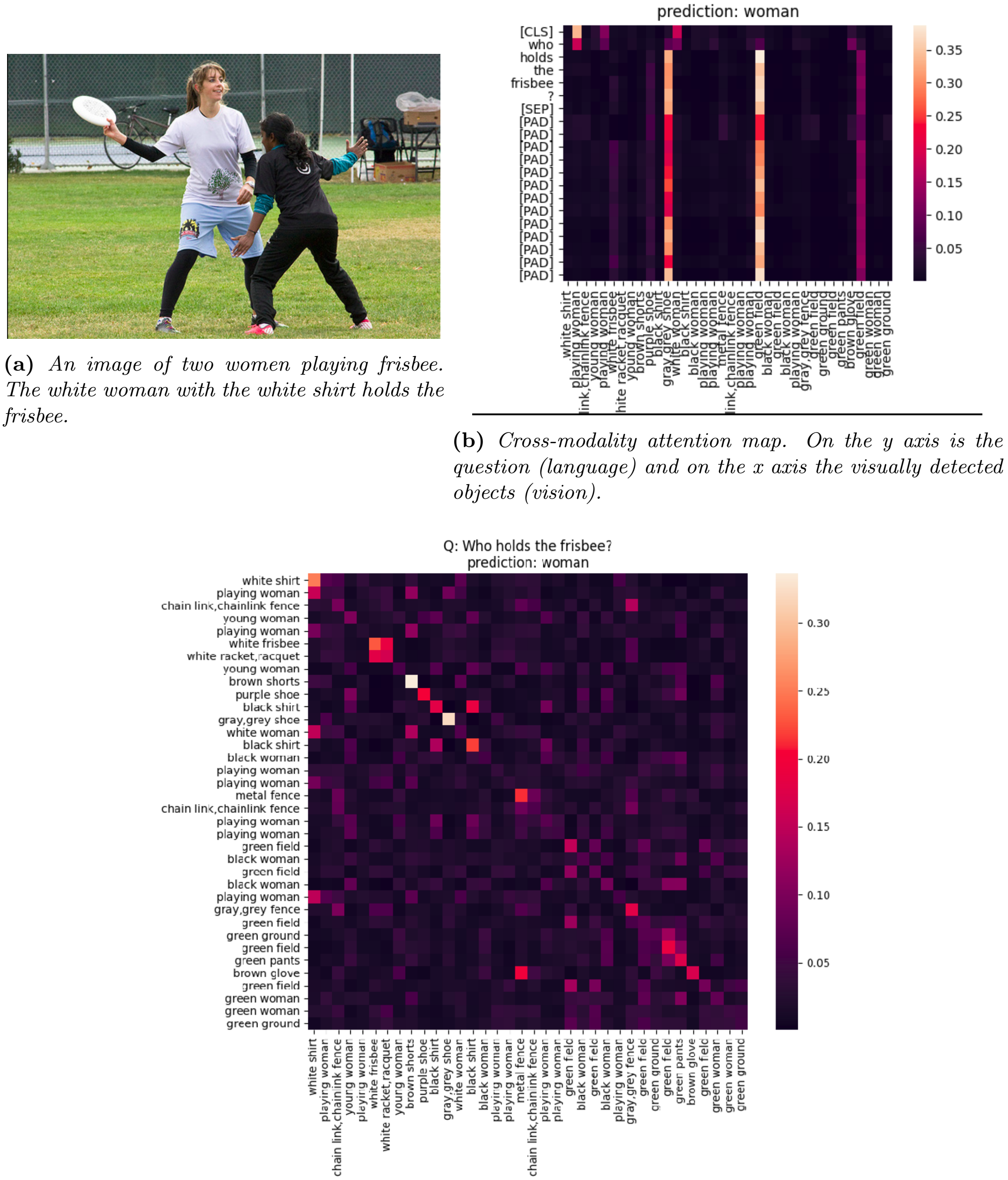
We move on to describe the two-stream models next. In ViLBERT [8], a two stream self-supervised attention-based model is proposed, extending BERT's architecture to visual-linguistic tasks. It consists of two parallel BERT-style models operating over image regions and text segments. This structure can accommodate the differing processing needs of each modality and provides interaction between modalities at varying representation depths. Analogous to BERT, for pretraining, ViLBERT uses Masked Multi-modal Modelling and Multi-modal Alignment Prediction. LXMERT [63] is a model almost identical to ViLBERT in the architecture, featuring slightly enhanced pre-training tasks, namely Masked Object Prediction (both feature regression and label classification) and Visual Question Answering, apart from Masked Language Modeling and Cross-Modality Matching. In ERNIE-ViL [64], authors propose a novel pre-training task for self-supervised vision-language representation learning, that utilizes scene graphs extracted from the text modality to incorporate alignments of detailed semantics into the joint representation. This task is composed of Object, Attribute and Relationship Prediction, that correspond to masked nodes on the scene graph. That way, it forces the model to extract object/attribute/relationship information from the visual modality, through concentrating on semantic words rather than common words like in MLM.

3.3.2 Attention maps: introspection and analysis

To better understand the functionality of the V-L BERT models' attention mechanisms and measure up their interpretability capacity, in this section we show plots of attention maps of the LXMERT [63] model finetuned for the GQA [9] visual question answering dataset, which we have produced ourselves through running inference and visualizing the results. We focus on the visual self-attention maps and the modalities' cross-attention maps.

For example, in Figure 3.3, we have an image of two women playing a game with a frisbee, in which the shorter brown woman in the black shirt tries to block the taller white woman in the white shirt. In the GQA dataset, this image is accompanied by the question "Who holds the frisbee?". In the top right, we visualize the cross-attention map where larger weight values are given to the connection of "who" and "CLS" to the (visually) detected "playing woman" and "white woman", as well as to the connection of the whole question to "grey shoe", "green field", "purple shoe" and "white frisbee". We can see that the model has gained some understanding of which of the two women the question is referring to, as well as an understanding of the environment and background. In the bottom figure we show the visual self-attention map, where the interesting connections can be shown if we filter out the main diagonal of the matrix map. The main of these here are in the 1st column, where "playing woman" and "white woman" are connected to "white shirt", 3rd column, where "chainlink fence" is connected to "metal fence" and "grey fence", and 4th column where "young woman" is connected to "black woman", "black shirt" and "playing woman". We can see that the model has gained some understanding of which woman is wearing which shirt.

In the example above, we observe that the question's answer is primarily mirrored in the



(a) *An image of two women playing frisbee. The white woman with the white shirt holds the frisbee.*

(b) *Cross-modality attention map. On the y axis is the question (language) and on the x axis the visually detected objects (vision).*

(c) *Vision self-attention map. Important connections lie outside of the main diagonal.*

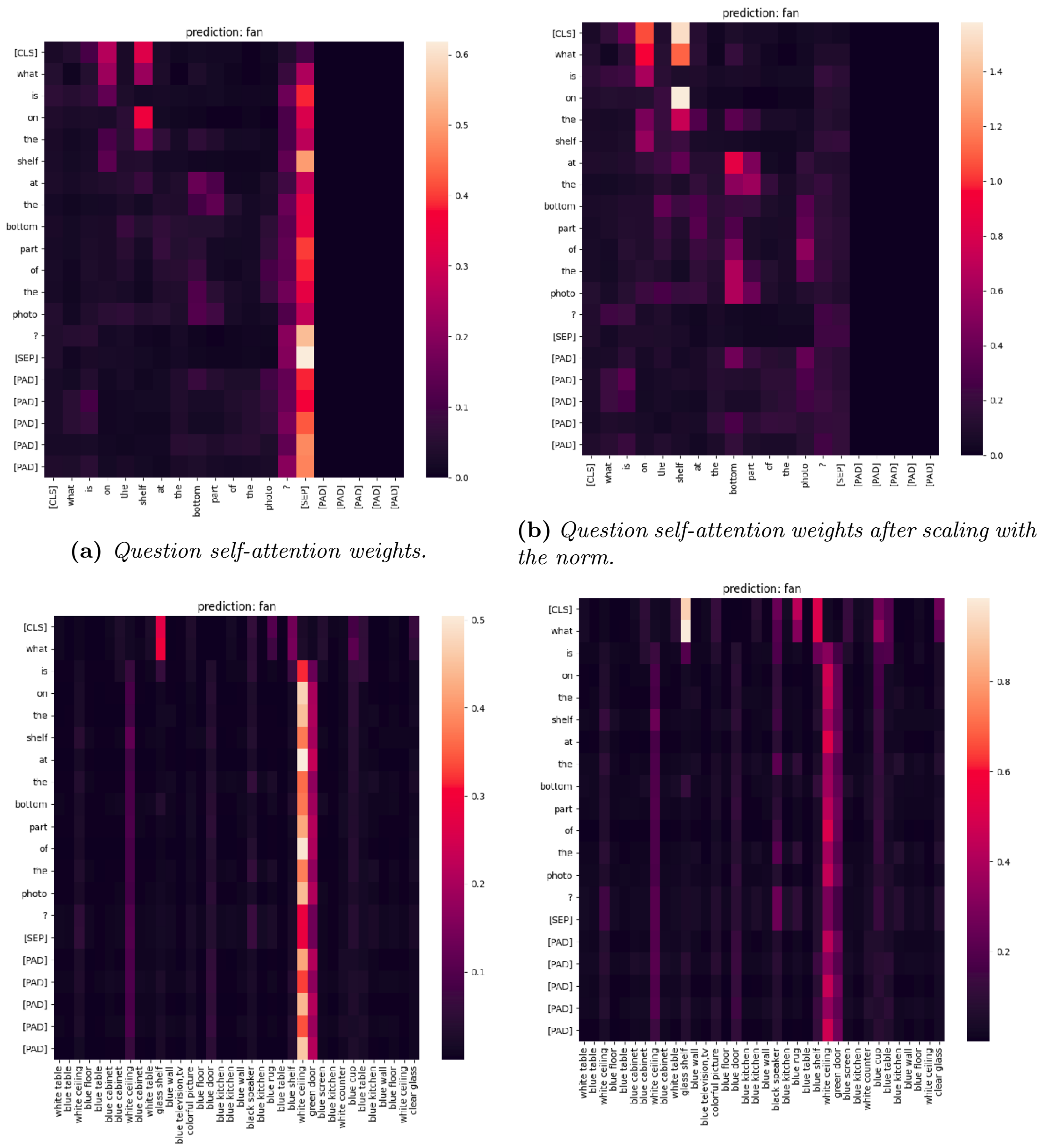
Figure 3.3. *In the attention maps, larger attention weights are denoted by lighter colours. The visualizations are produced by the author of this thesis. Image source: GQA dataset [9]*

attention given by the "CLS" token, which is the token trained for classification in the original BERT as well as V-L BERT models. In fact, 69.36% of those GQA test set samples which contained the final prediction in the detected objects, showed attention in the object-CLS connection. As far as the visual self-attention is concerned, in the previous example we noticed that we have to filter out the main diagonal to extract useful pairs. Indeed, by using a small threshold for the minimum attention weight and cutting out self-attended values, pairs like the following emerge.

```
Q: What color is the mountain peak? A: black
```

```
Visual self-attention pairs:
    (white clouds, blue sky), (large water, brown water), (green water, rocky rocks),
    (green water, green shore), (rocky rocks, black rocks)
```

In the example we showcased in Figure 3.3, one can also notice a phenomenon where, apart from "CLS" which usually contains the answer, very large attention weights are given on single (general) object across all tokens of the question (even the special padding and separator tokens), which we have to visually filter out when interpreting the map, to notice some of the smaller weight connections that are actually more useful (for example "white woman" has actually lower attention weight values than "green field"). The question here is how could we enforce this mathematically. This could be useful either simply for better visualizations or to even use the maps themselves as input to another model.



(a) *Question self-attention weights.*

(b) *Question self-attention weights after scaling with the norm.*

(c) *Cross-attention weights.*

(d) *Cross-attention weights after scaling with the norm*

Figure 3.4. *The effect of applying the norm-based analysis proposed in [10], both in the language self-attention (top) as well as in an extension for the cross-attention maps (bottom). The cross-modal extension implementation, as well as all visualizations are produced by the author of this thesis.*

In Attention is not only a weight [10], authors propose using the product of the attention weight α with the transformed input vector's norm $||f(\mathbf{x})||$, instead of just the weight, when analysing and interpreting attention maps, which they call "norm-based analysis". This is proposed for the original BERT language model, and can be described by the following equation.

$$\text{Attention Measure} = ||\alpha f(\mathbf{x})||, \quad f(\mathbf{x}) = (\mathbf{x}\mathbf{W}^V + \mathbf{b}^V)\mathbf{W}^O \tag{3.1}$$

Following this approach, we can easily extend the implementation for cross-attention in LXMERT, to combat the respective problem we previously described. In Figure 3.4 we show an example of the question "What is on the shelf at the bottom part of the photo?". We can see in the top row the effects of norm-based analysis in the original language self-attention formulation, where it is very effective in eliminating a useless distribution of weights in the separator token, and turning focus to more important syntactic and semantic connections like "on" with "shelf", "at" with "bottom", and "of" with "photo". Similarly, in the bottom row we can observe the effect of the norm for our implemented LXMERT cross-attention extension, where more focus is given to objects that can potentially answer the question such as between "what" and "glass shelf", "black speaker", and "blue rug", rather than "white ceiling" which is irrelevant to the question.

3.4 Visual Question Answering

A prominent task in vision-and-language machine learning that has received the attention of both the computer vision and natural language processing research communities, is the task of Visual Question Answering. It requires the AI agent to answer a natural language question based on an input image, either from a set of given answers (multiple-choice) or open-ended. It is sometimes described by authors as a visual Turing test [65, 66, 67, 68] as it could be used to check if a computer can trick a human into thinking it's human, based on its visual and language understanding skills. First introduced in the paper [23] along with the first dataset VQA-v1, it has inspired the creation of multiple datasets each focusing on a different aspect of the task. Some important examples of these are VQA-v2 [58] and VQA-CP [59], where visual understanding is promoted through making the same question have a different answer on different images, and through explicitly changing prior distributions respectively, GQA [9], where questions promote compositional reasoning and semantic understanding, and CLEVR [69] a diagnostic synthetic dataset that tests a range of visual reasoning abilities.

A direction that has proven very successful in the VQA literature is combining modules of memory and attention. In Dynamic Memory Networks for Visual and Textual Question Answering [70], Dynamic Memory Network (DMN) which was previously successful in the Text QA task, is extended for application in VQA. In Multimodal Residual Learning for Visual QA [71] the idea of deep residual learning is extended for joint representations in vision and language attention networks. In Dual Attention Networks for Multimodal Reasoning and Matching [72] the proposed architecture is composed of separate multi-modal reasoning and matching modules. In the reasoning model visual and textual attentions interact via collaborative inference, while the matching model estimates similarity between images and sentences using the two attention mechanisms. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering [31] proposed a combined bottom-up and top-down attention mechanism that enables attention to be calculated at the level of salient image regions. In Deep Modular Co-Attention Networks for Visual Question Answering [73] authors propose processing the images and questions through multiple layers of self-attention and cross-attention.

Another approach in recent VQA research is neurosymbolic models, which attempt to strike a balance between deep neural networks and older symbolic-AI approaches of functional programs and first order logic to get the best of both worlds. The method presented in The Neuro-Symbolic Concept Learner [74] translates sentences into symbolic programs which are executed on an object-based scene representation and makes use of curriculum learning, while in Neural Module Networks [75] authors propose partly differentiable models that use strong supervision to translate questions to functional programs and compose a question-specific neural network from a set of specialized modules. Moving towards a more neural approach, the method proposed in Learning by Abstraction: The Neural State Machine [15] predicts a probabilistic graph for the image semantics and performs sequential reasoning over the abstract latent space of that graph, iteratively traversing its nodes to answer a given question. On the one hand, the dependence of neurosymbolic models like [75] on externally provided functional programs and requirement of complex reinforcement learning training schemes undermines their robustness and generalization capacities. On the other hand, most deep neural network models are ultimately correlation engines that contain weak inductive bias and learn all structure from the amount of data they are trained on, often causing them to fail at generalization and few shot learning. In contrast to full neural, full symbolic, and other neurosymbolic models, Memory Attention Composition (MAC) Network [1] attempts to capture the logic of thought in addition to constructing neural representations from the data, exploiting the core ideas of attention that underlie neural models, but also providing an architecture suited for soft symbolic reasoning.

It has to be noted here that all of the vision and language models described in the previous section can also be fine-tuned for the downstream task of visual question answering, and achieve state of the art performances, being some of the most competitive models in the task.

Another interesting approach is based on the superior reasoning and common sense abilities of language models, and instead of fighting the tendency of multimodal models to depend on language, it translates everything to this easier modality. In Tell-and-Answer: Towards Explainable Visual Question Answering using Attributes and Captions [76] authors use pre-trained attribute detectors and image captioning models to extract attributes and generate descriptions for the image, and use them in place of the image to infer an answer to the question. The following are some benefits from this decomposition: (1) the attributes and captions can reflect what the system infers from the image, providing some justifications for the predicted answer (2) these intermediate results can help us identify the limitations of both the image understanding part and the answer inference part when the predicted answer is incorrect. This adds explainability in both success and failure cases. In Generating Question Relevant Captions to Aid Visual Question Answering [77] the proposed model generates an image caption to help answer the question, with the novelty being that this caption (what it focuses on) is also dependent on the question. In Image Captioning for Effective Use of Language Models in Knowledge-Based Visual Question Answering [78] authors propose to use a unimodal (text-only) procedure based on off-the-shelf captioning models for the image and pretrained language models for question answering, specifically for tasks that require external knowledge. This is based on the intuition that pretrained language models have been shown to include world knowledge much more than models trained in other modalities. In their analysis they show that increasing the language model's size notably improves its performance and that even though automatic captions often fail to capture relevant information (to the question) in the images, this is balanced by the better inference ability of the text-only language models.

3.5 Video Question Answering

This same task can be formulated with video content. What changes is that the input now has a temporal dimension, and may include audio and transcript of dialogue, which add to the difficulty of the multimodal setting but also offer opportunities to exploit more of the data in the real world. Important to consider is how the content of the questions is affected, as they may now deal with events (visual or audio), and spoken facts. Instead of just testing object recognition capabilities, the task can also require action and gesture recognition, conversation and story line understanding, as well as speech characteristics such as prosody, timbre and pitch. Some important datasets are TGIF-QA [79] and AGQA [80], adding actions to the VQA and compositional GQA settings respectively, TVQA [81] which incorporates data from tv-series with questions that refer to the dialog and story line in addition to temporal video events, and Social-IQ [2] that introduces the task of Social Video Question Answering through requiring the model to learn complex relations between the characters, an emotional understanding, as well as common-sense facts.

In Motion-Appearance Co-Memory Networks for Video Question Answering [82] DMN is enhanced with new mechanisms for Video QA, including a co-memory attention mechanism that utilizes motion and appearance cues. In Beyond RNNs: Positional Self-Attention with Co-Attention for Video Question Answering [83] authors replace the RNNs commonly used in Video QA with positional self-attention. Heterogeneous Memory Enhanced Multimodal Attention Model for Video Question Answering [84] is composed of three components, a heterogeneous memory to learn global context information from video, a question memory to understand the semantics of the question, and a multimodal fusion layer to performs multi-step reasoning by attending to relevant visual and textual hints with self-updated attention. Multiple cycles of reasoning can be made iteratively to improve the final representation of the QA pair. Progressive Attention Memory Network for Movie Story Question Answering [85] involves three main features, first an attention mechanism that utilizes cues from both question and answer to progressively discard irrelevant temporal parts in memory, second a fusion mechanism that dynamically calculates the contribution of each modality for answering the question, and third a belief correction answering scheme that corrects the prediction score on each candidate answer. In Dense-Caption Matching and Frame-Selection Gating for Temporal Localization in VideoQA [86] image captions are utilized to identify objects and salient regions and actions, in explicit textual format to allow easier matching for answering questions, in addition to the video features. The model formulation includes dual-level attention (word/object and frame level) and multi-head self/cross-attention for different sources (video and dense captions). Finally, the frame selection problem is cast as a multi-label classification task and two new loss functions are introduced for supervision with human frame selection annotation.

Similar to visual question answering, a line of work inspired from neurosymbolic approaches is also evident. Hierarchical Conditional Relation Networks for Video Question Answering [87] proposes a general-purpose reusable neural unit called Conditional Relation Network (CRN) that serves as a building block to construct more sophisticated structures for representation and reasoning over video. It is used in a CRN hierarchy whose branches represent sub-videos or clips, all sharing the same question as the contextual condition. Neural Reasoning, Fast and Slow, for Video Question Answering [88] introduce a dual process neural architecture for Video QA where they use MAC as System 2, which obtains input from a temporal attention mechanism across space-time (System 1). This System 1 is almost identical to the unit they propose in HCRN [87]. Their results showcase MAC's potential for use in more complex settings of video and multimodal input.

VideoBERT [89] builds upon the BERT model to learn bidirectional joint distributions over sequences of visual and linguistic tokens, derived from vector quantization of video data and off-the-shelf speech recognition outputs, respectively. For the pretraining task, for text-only and video-only the standard masking objectives are used, and for text-video, a linguistic-visual alignment classification objective is used. Learning Video Representations using Contrastive Bidirectional Transformer [90] builds on the fact that in VideoBERT, vector quantization (VQ) loses critical fine-grained information, while in models such as vilbert pre-trained visual encoders are needed to measure visual similarity. Authors propose a way to train bidirectional transformer models on sequences of real valued vectors, using noise contrastive estimation (NCE), without the use of pretrained models. HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training [91] utilizes large scale pretraining of two different transformers for cross-modal and temporal modeling respectively, with the help of the novel Video-Subtitle Matching (VSM) and Frame Order Modeling (FOM) pretraining tasks in addition to Masked Language Modeling(MLM) and Masked Frame Modeling (MFM). In Less is More: CLIPBERT for Video-and-Language Learning via Sparse Sampling [66], sparsely sampled sub-clips from the input video are used in a ResNet backbone prior to temporal pooling and addition of a 2D positional embedding for joint encoding with text input in BERT, an architecture that is trained end-to-end in cross-modal pretraining and downstream task finetuning, including Video QA. Predictions are derived for each sub-clip and later aggregated through a consensus function (e.g. mean pooling) to obtain the task-specific loss.

In MMFT-BERT [92], authors propose a network comprising of three different pretrained BERT instances for question and answer, object detection labels, and subtitles respectively, followed by a BERT-like transformer for fusion, and trained end to end with separate and joint loss functions. They also conduct extensive ablation studies, many of which are also followed in this work. In a similar manner, BERT Representations for Video Question Answering [93] explores a similar simplified setting with vector summation for modality fusion, different separation token ablations, and different input pruning strategies including TF-IDF. Focusing on that, it can be observed that in datasets such are TVQA where temporal localization is provided, problems with subtitle content length can be easily avoided, but in datasets where there is no such supervision one must either strategically prune or extract a meaningful summarization. The latter is what [94] and [95] proposing, and the direction partly taken in aspects of our work. In Knowledge-Based Video Question Answering with Unsupervised Scene Descriptions [94], authors propose extracting rich and diverse information by processing scene dialogues, generating unsupervised video scene descriptions, and obtaining external knowledge via weak supervision. The information generated by each of the above parts is encoded with a Transformer and encodings are fused through a modality weighting mechanism. In On the hidden treasure of dialog in Video Question Answering [95] authors treat dialog as a noisy source to be converted into text description via dialog summarization, much like recent methods treat video. The input of each modality is encoded by transformers separately, followed by a simple fusion method using soft temporal attention.

Chapter 4

Social Video Question Answering

4.1 Introduction

Humans are social creatures; our survival and well-being depends on our effective communication with others. This is achieved through perceiving and understanding information from multiple sensory modalities as well as reasoning and arriving to conclusions, in order to respond accordingly. More precisely, we rely on the ability to understand other peoples' mental states and make forecasts about their behaviour, which, in the view of evolution, has proven critical in determining potential threats or advantageous opportunities as well as to form and maintain relationships in order to fulfill safety and basic physiologic needs [96].

Most humans seem to effortlessly understand other people's mental states, (which include intentions, motivations, feelings) without their having to directly communicate them, or can even discern them even in the case where what they show is the exact opposite, as is the case with sarcasm. For example, someone may express their dissatisfaction by saying "I'm having so much fun", and others will be able to tell what they truly feel, by factors like their eye gaze, facial expression, body language (posture, gestures), and tone of voice. These are called non-verbal cues, and facilitate what is called non-verbal communication. In general, the term social cues refers to both verbal and non-verbal signals, which guide conversations and other social interactions by influencing our impressions of and responses to others [96].

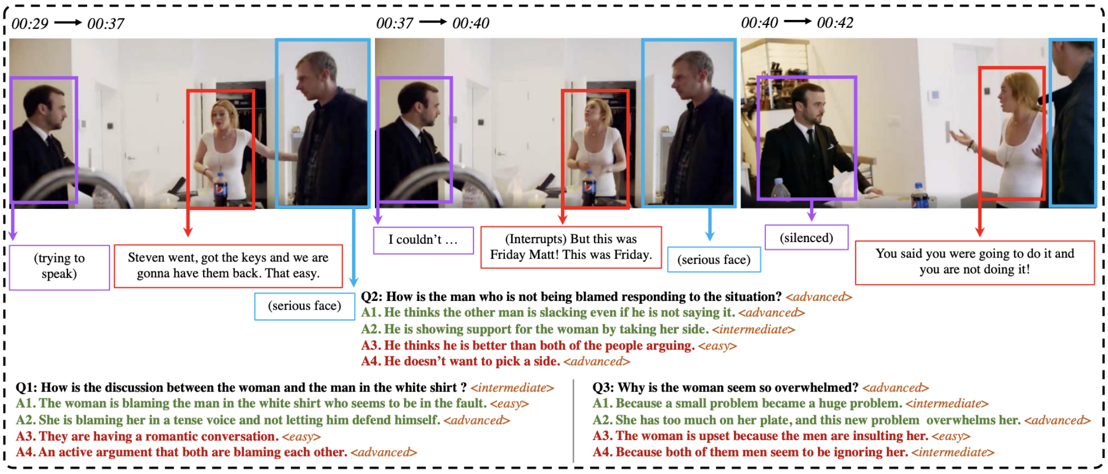
The ability to perceive social cues and form conclusions about other people's mental states is often referred to as theory of mind (ToM) or mentalization and is already evident from ca. 18 months of age [97]. Some people, albeit very intelligent in other matters like people with ASD, cannot discern those cues as they operate mainly on logical and factual information, as is the case with most machine learning tasks. The invention of a question answering task for this matter serves both as a way to train people with ASD to recognise such behaviours, as well as machine learning models when framed similarly to other question answering tasks.

Social Video Question Answering is a task to test the social reasoning abilities of an agent, based on how accurately they can answer questions on a given video. It incorporates just one manifestation of the human social cognition abilities, but it is very easy to formulate for artificial agents in a supervised machine learning scenario. Compared to regular Video Question Answering, it consists of social and theory-of-mind-related (as opposed to more factual) questions [11], with the focus being on people, rather than objects/environment, and on interactions, rather than actions. It can require sophisticated combinations of emotion recognition, language understanding, cultural knowledge, logical and causal reasoning, on top of non-social layers of comprehension about physical events. It also provides a valuable methodology both for studying social reasoning in humans (e.g. with ASD), and developing AI agents with social reasoning skills. Implicitly, people are doing social-VQA-like reasoning every time they watch videos. Given the amount of time that children now spend watching videos each day, such video watching is an major part of social learning experiences for modern humans.

Social reasoning in artificial agents also has a strong connection with the goal of artificial general intelligence. In [16] authors explain that NLP needs social context to truly succeed, marking that as the last stage to obtain a truly complete understanding of the world through language. More specifically, the NLP progress is defined by the conquering of different World Scopes, each one more general than the last, ordered as Corpus, Internet, Perception (multimodal), Embodiment, and Social.

4.2 Datasets

Social-IQ [2] is an unconstrained benchmark that introduced the task of Social Video Question Answering. It consists of human-centered videos in the wild (YouTube) featuring real-world interactions along with social and theory-of-mind-related multiple choice questions that probe social judgment, motivations and behaviors, mental states, and more. It is currently the only large scale Social VQA dataset.


00:29 ⟶ 00:37
00:37 ⟶ 00:40
00:40 ⟶ 00:42
(trying to speak)
Steven went, got the keys and we are gonna have them back. That easy.
(serious face)
I couldn't …
(Interrupts) But this was Friday Matt! This was Friday.
(serious face)
(silenced)
You said you were going to do it and you are not doing it!

Q2: How is the man who is not being blamed responding to the situation? *<advanced>*
A1. He thinks the other man is slacking even if he is not saying it. *<advanced>*
A2. He is showing support for the woman by taking her side. *<intermediate>*
A3. He thinks he is better than both of the people arguing. *<easy>*
A4. He doesn't want to pick a side. *<advanced>*

Q1: How is the discussion between the woman and the man in the white shirt ? *<intermediate>*
A1. The woman is blaming the man in the white shirt who seems to be in the fault. *<easy>*
A2. She is blaming her in a tense voice and not letting him defend himself. *<advanced>*
A3. They are having a romantic conversation. *<easy>*
A4. An active argument that both are blaming each other. *<advanced>*

Q3: Why is the woman seem so overwhelmed? *<advanced>*
A1. Because a small problem became a huge problem. *<intermediate>*
A2. She has too much on her plate, and this new problem overwhelms her. *<advanced>*
A3. The woman is upset because the men are insulting her. *<easy>*
A4. Because both of them men seem to be ignoring her. *<intermediate>*

Figure 4.1. *Example from the Social-IQ dataset, along with annotations of bounding boxes and non-verbal cues which are not included in the dataset. Source: [2]*

TinySocial [11] dataset is a small Social VQA dataset that includes social VQA samples from popular television and movie clips (on the order of 100 video clips, with 6-12 multiple choice questions per clip), and is intended primarily for human consumption but can also serve as a useful test for artificial social reasoning agents.

There are also some video QA datasets that include a few social questions along with a vast majority of factual questions, such as TVQA [81], PororoQA [98] and MovieQA [99], as well as datasets that are not video QA but probe social reasoning, such as visual commonsense reasoning [100], which evaluates commonsense reasoning in VQA format, violin [101], which includes social reasoning skills in an entailment format, the Theory of Mind Task dataset [102], which contains short textual stories and theory-of-mind questions, and the Motivations dataset [103] which includes images of people labeled with their likely motivations.

4.2.1 Social-IQ

Analysis

The Social-IQ dataset consists of videos in the wild (from YouTube), that are primarily TV shows, interviews, vlogging content, or TV series. This is a parameter that makes the dataset especially hard, as the videos are as close as possible to what one would randomly watch on the internet. There is also some variability in the active characters per video, meaning the characters that participate in the social situation and are not bystanders. In Figure 4.2, it is observed that about 70% of the videos contain 2 or 3 active characters and a 20% having 4 or more characters. Video duration is typically 1 minute, in contrast to other video QA datasets where videos are much longer (1 hour, MovieQA [99]) or shorter (a few seconds, TGIF-QA [79]).

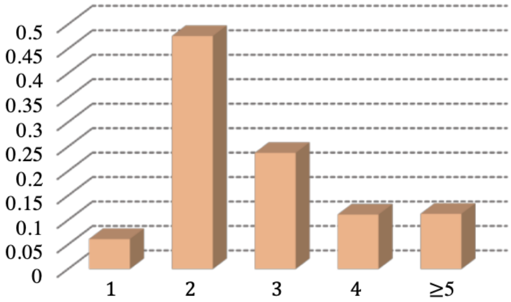


Figure 4.2. *Distribution of active characters per video. Source: [2]*

The social questions in the dataset mainly refer to people as is expected for this type of human-centered tasks, more specifically by their gender, as can be seen in Figure 4.3. Characterizations like "people", "person", and "audience" are also evident. In a secondary level, one can see that there are often affective and theory of mind related words such as "thinks", "feels", "wants", "happy", "sad", "angry", as well as visual attribute descriptions such as colors, clothes, hair types, and activity specifications such as "speaking" and "talking".

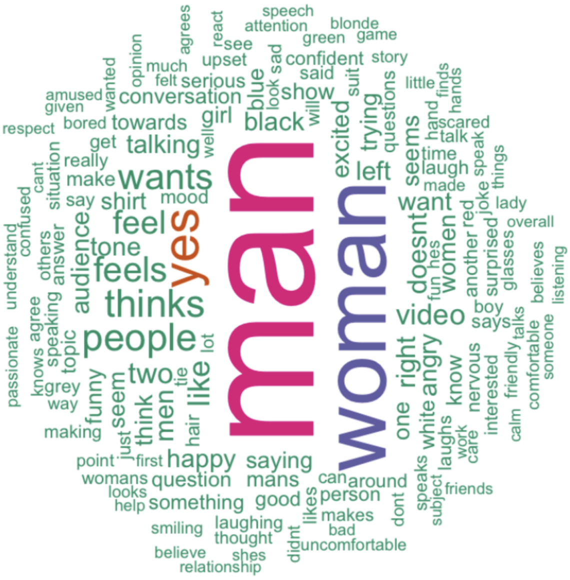


Figure 4.3. *Word frequency cloud on the Social-IQ questions and answers. Source: [11]*

Following the analysis of [11], as shown in Figure 4.4 there are several types of clues relevant for answering the question, which are mostly the words being said, the prosody of what is said (tone, volume, pitch, etc), facial expressions, and body language (gestures, posture), rather than physical actions, objects and events or the environment that the scene takes place. This is in consistency with the description of the social video question answering task.

Authors in [11] also inspect the kind of answering process that is required for the questions, like inferring emotions, relationships between people (either a surface relationship such as siblings, or a functional relationship such as the conflict-starter), inferring the motivation or beliefs of people (labeled as reasoning), or understanding the difference between surface meaning and deeper meaning, like in the case of sarcasm. Here, it is found that the answering process is required to be based mostly on reasoning and emotion recognition (Figure 4.5).

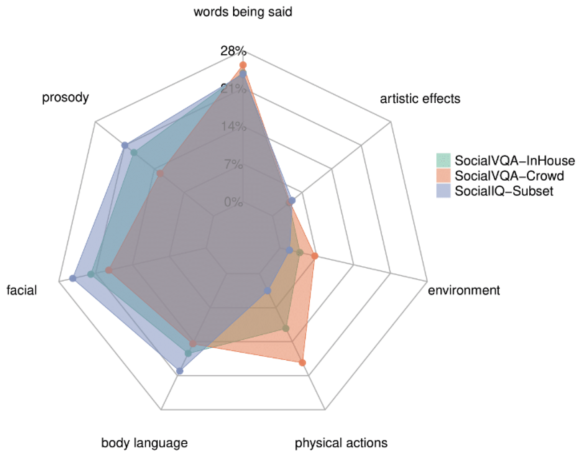


Figure 4.4. *Types of clues relevant for answering the question (human raters). Source: [11]*

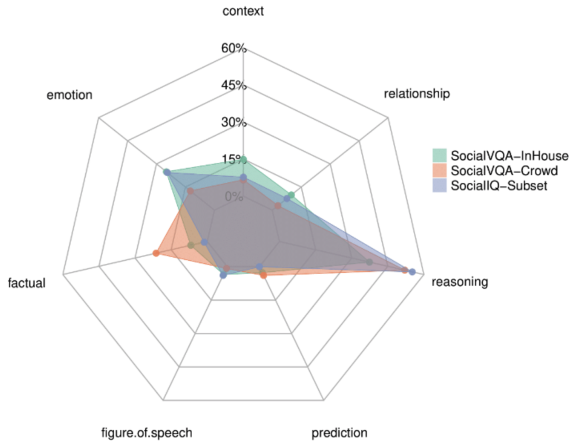


Figure 4.5. *Type of knowledge / reasoning process the question draws upon (human raters). Source: [11]*

Evaluation

The Social IQ dataset (public release) contains 1015 videos, with six questions corresponding to each video and each question having four correct and three incorrect candidate answers. The training set contains 888 videos and the validation set 127 (87% - 13% split). In all experiments the above validation set is used for evaluation and comparison of the models, as the private test set is reserved by the authors for future challenges.

The dataset metrics are binary (A2) and four-way (A4) accuracy for the binary and multiple choice tasks respectively, following the original formulation presented in [2]. For the binary task (A2) we take all 12 combinations of correct and incorrect answers for a question, resulting in a dataset of 73,080 total samples where the goal is to select the correct answer between the two. For the multiple choice task (A4) we take all four combinations of one correct and three incorrect answers for a question, resulting in a total of 24, 360 samples where the goal is to select the single correct answer from four choices. Note, the performance of random choice is 50% for A2 and 25% for A4.

It has to be noted here, that although the dataset statistics mentioned above are the ones reported by the authors, this is not really the case in the dataset. There's a minority of the questions with either fewer or more than 7 answers, and some with different correct/incorrect ratio. This affects the multiple choice task's formulation, as there is ambiguity as to how these cases should be handled. It also has to be mentioned that there are 184 samples with missing transcripts, and the way this is handled in this work is by substituting transcript with the empty string for those samples.

4.3 Previous Work

For the task of Social Video Question Answering, the methods previously explored on Social-IQ typically make use of attention and fusion mechanisms, and can be summarized as follows. First, Tensor-MFN (TMFN) [104] is a baseline created by performing architecture and hyperparameter search on TFN and MFN models and combining them into a joint model. More specifically, TMFN uses Tensor Fusion for multimodal fusion in the recurrent stages of MFN. MCQA [17] consists of two main components, "Multimodal Fusion and Alignment" which is the input fusion and alignment, and "Multimodal Context-Query Alignment" which is the cross-alignment of joint context with query, where the query corresponds to the question and answers. The chosen method for fusion is (pair-wise) co-attention performed on BiLSTM encodings (for capturing context). The RNN-based model TACO-Net in [18] is also based on two basic parts. First, temporal attention for the multimodal aligned features (BiLSTM encoded) and keyword highlighting for the questions and answers, and second, a consistency measurement module that computes cosine similarity between Q&A and multimodal data, and between modalities. This is then fed into a multi-step reasoning module that outputs joint consistency measurement scores for all answers. In [19] the authors suggest a regularization term to balance the clues between modalities, with the goal of maximizing it while minimizing loss. This term is based on the functional entropy which they estimate via bounding it with the functional Fisher information using the log-Sobolev inequality.

4.4 Baseline Models

In this section we will describe several variations for the formulation of the baseline models. In all of these, the language modality inputs (Q: question, A: answer, T: dialogue transcript) are encoded with BERT embeddings, and the visual modality (V: video frames) with Densenet161 features for each frame. Frames are always down-sampled at 1fps, and the transcript is truncated at 512 tokens since this is BERT's maximum input size.

As for the baseline architecture, both a simple classifier for the embeddings, and further LSTM processing is explored. The latter follows the baseline models used by the authors in [2] and involves further encoding the BERT last hidden state and frame feature sequences with LSTMs, before feeding them to a linear classifier. For the former, either BERT pooler output or averaging of the last hidden state is used to aggregate the sequences, in both cases followed by a linear classifier.

In the experimental setup for running the baseline models two distinct approaches can be seen, one based on the reproduction of the pipeline in the published code of [2], and the other based on more intuitive decisions for the loss function as well as other factors. We will describe these with regard to the binary task for simplicity.

4.4.1 Reproduction and analysis

The first approach follows the code released by the dataset's authors, where MSE loss is used to train the model. More specifically, the loss described in equation 4.3 is used, which averages the batch's predictions beforehand.

In this setup, after all input sequences are encoded by LSTMs, they are fed into two linear classifiers in this fashion:

$$Y_1 = f(Q + X^* + A_1), \quad Y_2 = f(Q + X^* + A_2) \tag{4.1}$$

where Q corresponds to the LSTM encoding of the question and A_1, A_2 of the two answers, X* to the concatenation of any number of LSTM encoded video features (e.g. V, T), and + denotes concatenation.

The LSTM encoding corresponds to the hidden state, and the output dimensions used here are the same as in the authors' code; 50 for the text modality and 20 for the visual.

The equations of the original MSE loss and the MSE loss used in the authors' code are, respectively, the following:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}((Y^i_{cor} - 1)^2 + (Y^i_{inc})^2) \tag{4.2}$$

$$L = (\frac{1}{N}\sum_{i=1}^{N} Y^i_{cor} - 1)^2 + (\frac{1}{N}\sum_{i=1}^{N} Y^i_{inc})^2 \tag{4.3}$$

where N is the number of samples in a batch.

Note that although we trained and evaluated our models with random order in the input choices A_1, A_2 and then grouped the predictions Y_1, Y_2 by their ground truth labels to compute the loss, validation accuracy is not affected by training with the choices in specific order, as it is implemented in the author's code.

As for the binary accuracy, it is measured in the implementation by

$$\frac{1}{M}\sum_{i=1}^{M}(Y^i_{cor} > Y^i_{inc}) \tag{4.4}$$

instead of the standard

$$\frac{1}{M}\sum_{i=1}^{M}(argmax(Y^i_1, Y^i_2) = G^i) \tag{4.5}$$

where M is the number of samples in that set and G^i is the ground truth label. The difference between the two is important, as the first is more strict in the sense that it never lets a prediction be classified as accurate (by choosing the correct one by chance) when the logits for the correct and incorrect choices are equal.

In this setup, and with using the precomputed embeddings and features provided by the authors, our reproduction results for the baselines are shown in table 4.1, reporting on mean and standard deviation over five runs.

Modalities	Reported test set A2	Reproduced val set A2
QA	57.02	64.51 (±0.58)
QAV	63.91	64.82 (±0.67)
QAT	57.87	64.54 (±0.57)
QAVT	-	64.61 (±0.72)

Table 4.1. *Comparison of reproduced with reported binary accuracy results, for LSTM baselines on BERT / D161 precomputed sequence features, using the loss in 4.3. We report results on mean and standard deviation over five runs.*

We can see a big gap between the reproduced and reported baseline for the QA modality, which can be attributed to the authors' claim[1] that the reserved test set is "more multimodal", causing uni-modal models to fail to generalize there, whereas in the validation set they overfit and exhibit over-stated abilities.

Modalities	Precomputed	From Scratch
QA	64.51 (±0.58)	74.89 (±0.52)
QAV	64.82 (±0.67)	75.33 (±0.64)

Table 4.2. *Validation set binary accuracy comparison of precomputed BERT embeddings versus computing them from scratch, for the LSTM baseline on BERT / D161 sequence features.*

Furthermore, when computing our own BERT embeddings for the questions and answers, we achieve a performance boost of 10%, resulting to 74.89% in QA, and 75.33% in QAV, as shown in table 4.2. Computing the visual features from scratch on the other hand doesn't have any effect. We take the following steps to investigate the cause of this phenomenon. First, we change the accuracy measure from

[1] https://github.com/A2Zadeh/Social-IQ/issues/1#issuecomment-521301071

equation 5.8 to equation 4.5 when using their embeddings, and observe a +4% improvement. We further change equation 5.8 from "greater" to "greater or equal" to model the extreme case of classifying all equal as accurate and measure the impact that these equal cases make. We obtain an improvement of +11%, whereas with our embeddings a difference of only +3% is noted. This is shown in the histograms of figure 4.6.

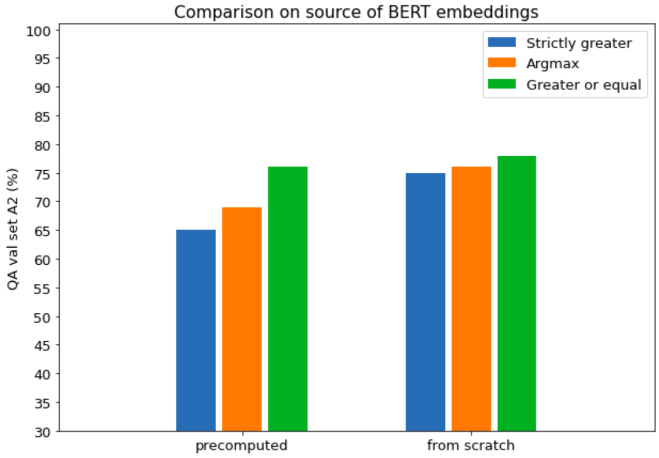


Figure 4.6. *Comparison of the validation set binary accuracy of the precomputed and computed from scratch embeddings, for different accuracy metric schemes. In the precomputed embeddings, an accuracy improvement of 11% is observed when considering equal logits as correctly classified.*

Noticing that, we count the number of samples with the same embedding for correct and incorrect choice, and for their embeddings it is 8.6% of the total samples, whereas in ours only 0.67%. In the train set these percentages are 8.6% and 0.24% respectively, and in the val set, 8.5% and 3.6%, as shown in figure 4.7. These samples are further inspected, and the 3.6% that has equal embeddings in both cases indeed corresponds to bad samples, such as having "N/A" and "Bad clip" in all answers. The rest of the equal embeddings in the precomputed have no obvious defects in their respective sample's text.

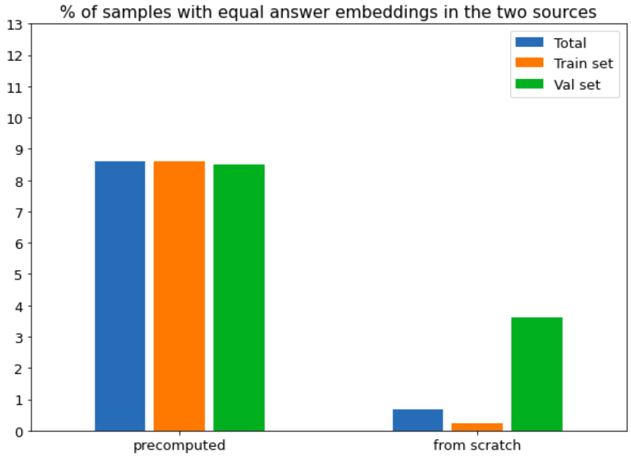


Figure 4.7. *Investigation results of the frequency of cases with correct and incorrect answers having equal embeddings, in the precomputed versus from scratch embeddings. This significant difference indicates an error in the authors' computation.*

Lastly, when calculating all combinations of binary answers for our embeddings, we end up with 2,500 more samples than in the precomputed, which can lead to better model weights and additional performance improvement.

Language Bias in Social IQ

The performance described above, obtained when computing the embeddings directly from the raw data of this multimodal dataset, is inconsistent with their statement "At a first glance, our bias analysis experiments demonstrate minimal bias in the Social-IQ dataset coming from Q+A. BERT embeddings, commonly known for their success in common-sense reasoning, show slightly higher performance than random" [2]. We conclude that Social IQ, like most multimodal and especially visual and video question answering datasets [59, 105], also suffers from the effects of language bias.

Other remarks

In addition, as for the baseline provided by the authors' proposed model TMFN, there are some peculiarities that should be reported. When running the published code exactly as is, we get a reproduction of 65.62% for TMFN. When completely removing TMFN's output from the final classifier (leaving only LSTM outputs), we obtain a performance of 66.17%. Moreover, we notice that both classifiers get both the correct and the incorrect answer representations, just in different order. When further removing that, performance drops to 62.13%, which leads us to think that TMFN's performance can be attributed to this trick. Similar peculiarities are also mentioned in [20].

4.4.2 Training with different losses

We notice that when using the standard MSE loss described in equation 4.2 instead of the one proposed in the authors' code 4.3, performance increases about 2%. Using this MSE also makes the models converge much faster, e.g. in 2 epochs instead of 10.

Another approach is training with the more intuitive cross entropy loss computed between the batch's predicted logits and the labels for the correct answer, as described below, similar to [92, 18].

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} G^{i} log(\sigma(Y_c^i)) \tag{4.6}$$

where N is the number of samples in a batch, C the number of classes, G^i the ground truth label, and σ the softmax function.

In all results below, our embeddings computed from scratch are used and accuracy is calculated according to equation 4.5 (argmax). Again in this case we tested using their embeddings which resulted in 7% lower accuracy, due to the discarded samples as explained in the previous section.

In this setup for training with cross-entropy, two alternatives are shown for the text inputs, namely with and without LSTM sequence encoding. Results for the LSTM baseline alternative with output dimension set to 512 can be seen in table 4.3, reporting on mean and standard deviation over five runs.

Modalities	LSTM
QA	77.77 (±0.31)
QAV	78.15 (±0.19)
QAT	78.42 (±0.21)
QAVT	78.72 (±0.22)

Table 4.3. *Validation set binary accuracy for the LSTM baseline on BERT / D161 sequence features, computing BERT embeddings from scratch and using Cross-Entropy loss (4.6).*

We further investigate the impact of the LSTM's output dimension in the QA baseline, and notice that it has a considerable effect, with 76%, 77% and 77.8% mean accuracy for output sizes of 50, 100, and 512 respectively. In addition, we note that bidirectionality has no impact.

In the second alternative, where LSTM modelling is omitted, to aggregate the text sequences before classification either the BERT pooler output is used (where the pooler linear layer needs to be trained as well), or equivalently a weighted average of the last hidden state embeddings (weighted by the attention mask to factor out padding tokens). Comparative results between these different embedding schemes, as well as different levels of fine-tuning are shown in table 4.4.

Embedding: Training	BERT
CLS: Frozen	60.39
CLS: Fine-tuning pooler	74.37
LH: Frozen	75.10
CLS: Fine-tuning pooler + last 3	80.35
LH: Fine-tuning last 3	80.67

Table 4.4. *Ablation study for different BERT embedding sources and fine-tuning levels.*

Similar to [92, 93], the text modality inputs can be jointly encoded by BERT along with the questions and answers, instead of separately, effectively performing text question answering. An ablation is performed comparing the input order, which can be either like:

$$\begin{aligned} E_1 &= bert([CLS] + Q + T + [SEP] + A_1) \\ E_2 &= bert([CLS] + Q + T + [SEP] + A_2) \end{aligned} \tag{4.7}$$

or

$$\begin{aligned} E_1 &= bert([CLS] + T + [SEP] + Q + A_1) \\ E_2 &= bert([CLS] + T + [SEP] + Q + A_2) \end{aligned} \tag{4.8}$$

and then passed to a classifier together with other modalities' inputs X* (e.g. V):

$$Y_1 = f(E_1, X^*), \quad Y_2 = f(E_2, X^*) \tag{4.9}$$

As shown in table 4.5, the order in equation 4.8 is more beneficial, which can be attributed to the fact that it allows for more context to be inputted in the model without being truncated (truncation in the BERT tokenization step can be performed in either of the two sentences).

Modalities	BERT
QA	74.37 (±0.06)
QAT 4.7	74.39 (±0.05)
QAT 4.8	74.59 (±0.05)

Table 4.5. *Comparison of two different BERT input schemes, in training a classifier on BERT pooler output embedding (CLS) and Cross-Entropy loss (4.6).*

To summarize and connect the analyses of the above sections to our own research and experimental process, we must make clear that the reproduction setup in section 4.4.1 with the precomputed input embeddings and training loss, is what is used in the first part of our work in order to compare our end-to-end method to the state-of-the-art techniques published on the dataset, while the more intuitive approach of computing the embeddings from scratch and training with the standard cross-entropy loss for classification shown in section 4.4.2 is used in the second part of our work, where we effectively perform text question answering in a scheme similar to the one described above.

Chapter 5

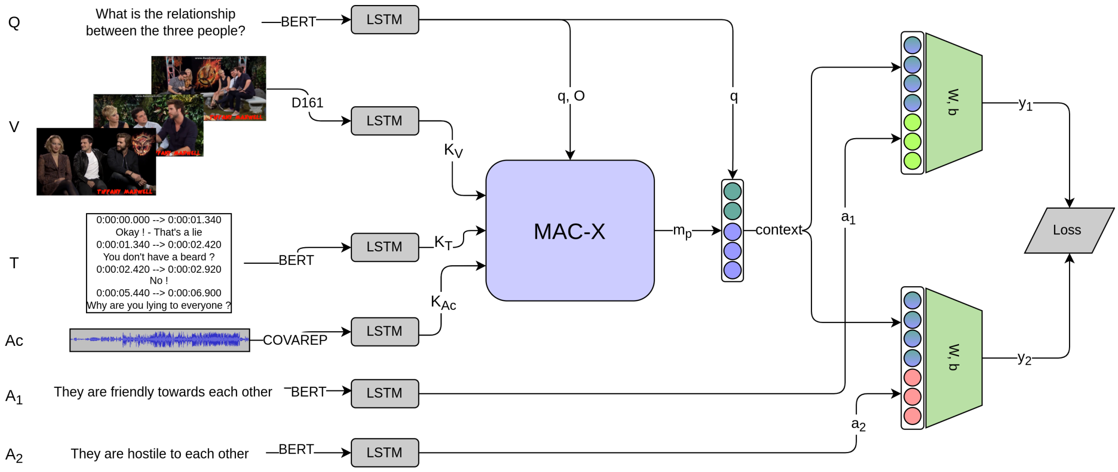
Proposed Approaches

5.1 End-to-end Approach: MAC-X

5.1.1 Overview

In this first part of our work, we propose a multimodal extension of the MAC Network for Social-IQ, called MAC-Extend (MAC-X). The motivating factors for this approach are that MAC: 1) was intended for tasks that require deliberate reasoning from facts to conclusions on account of its structured and iterative reasoning, and 2) consists of thoroughly general-purpose modules and operations. We believe that these characteristics make it very well-suited for Social-IQ, and a strong baseline for the task of Social Reasoning as well as any reasoning task.

Our model is based on the MAC Network, a recurrent architecture of length p and dimension d defined by the Memory, Attention and Composition (MAC) cell which performs an attention-based reasoning step i given a knowledge base and a query. The MAC cell is composed of three operational units, the Control Unit, the Read Unit, and the Write Unit. This pipeline reads from input features in a way that is controlled by part of the query and memory from previous readings, proceeding to incorporate that into the current memory. One of its most important features is the separation between control (c_i) and memory (m_i) that it enforces, and that the interaction between the knowledge base and the query is only mediated through probability distributions. In the succession of its total p recurrent iterations, it has the ability to represent arbitrarily complex acyclic reasoning graphs in a soft manner [1].



0:00:00.000 --> 0:00:01.340
Okay ! - That's a lie
0:00:01.340 --> 0:00:02.420
You don't have a beard ?
0:00:02.420 --> 0:00:02.920
No !
0:00:05.440 --> 0:00:06.900
Why are you lying to everyone ?

Figure 5.1. *Overview of the proposed end-to-end architecture, centered around the MAC-X Network. On the left, the question (Q), visual frames (V), dialogue transcript (T), acoustic input (Ac) as well as correct (A_1) and incorrect (A_2) answers are shown for the binary task. Their features are encoded with LSTMs, before use in MAC-X or in final classification along with last memory m_p. Two identical classifiers make the predictions y_1, y_2 which are then used to calculate the loss in equation (5.7).*

Building on these structural priors, MAC-X extracts information from multiple sources, formulates its attention over time instead of space, performs a mid-level fusion on the intermediate representations of the modalities, and ultimately facilitates multiple-choice Question Answering on multimodal data. An overview of the model's architecture for the task of Social Video QA can be seen in Figure 5.1, and the enhanced cell's architecture is shown in Figure 5.2. In the following sections, all equations and figures are described for the binary task for simplicity, and can be directly extended for the multiple choice task in which we also report results.

5.1.2 Input Units

As shown in Figure 5.1, the language modality inputs which consist of the question (Q), the dialogue transcript (T) and the correct and incorrect answers (A_1, A_2 respectively), are initially encoded with last hidden state BERT embeddings, while the visual modality (V) with Densenet161 (D161) features for each frame (at 1fps), and the acoustic modality (Ac) with COVAREP features. They are then passed through bidirectional LSTMs whose outputs constitute the knowledge bases K_V, K_T and K_{Ac} for the visual, transcript and acoustic input respectively and the contextual words O for the question. The hidden states q, a_1, and a_2 are used as the vector representation for the question and answers respectively. The output dimension of the LSTMs is d, where d is the dimension of the MAC model. Each of the knowledge bases can be described as $K_j^{L\times d} = \{k_t|_{t=1}^{L}\}$, where L is the sequence length of modality j in the time dimension t.

5.1.3 The MAC-X cell

Control Unit

The Control Unit (Figure 5.2) stays the same as in the original architecture, and can be summarized as

$$c_i = \sum_{s=1}^{S} \sigma(f_c(f_{cq}([c_{i-1}, f_q(q)]) \odot O_s)) \cdot O_s \tag{5.1}$$

where S is the number of contextual words, σ the softmax function, and f_x are single layer feedforward networks. In the equation above, attention is performed on the contextual words O based on information from the question q and the previous control c_{i-1}, in order to update the current c_i. This c_i determines what part of the question we want to extract knowledge about from the input modalities in the current reasoning step.

Multiple Read Units

For reading from the knowledge bases, a simple cloning of the Read Unit for each modality is proposed, each getting a copy of the previous control and memory (see Figure 5.2). This approach allows for the control c_i to attend to the different modalities independently at the same reasoning step, while at the same time being conditioned on a memory that is kept collectively for all of them. For example, previous information from the audio and visual modalities could be important to determine the next most useful information to integrate from the transcript. The operation of each Read Unit j is defined as

$$I_{i,t}^{j} = f_{mk}([f_m(m_{i-1}) \odot f_k(k_t^j), k_t^j]) \tag{5.2}$$

$$r_i^j = \sum_{t=1}^{L} \sigma(f_r(c_i \odot I_{i,t}^{j})) \cdot k_t^j \tag{5.3}$$

where $j = V, T, Ac$ are the different modalities. In the former of the above equations, information $I_{i,t}^{j}$ is gathered from the knowledge base of modality j at each position t in its temporal sequence. This information is considered to be only optionally related to the previous memory m_{i-1}, and so the initial k_t^j

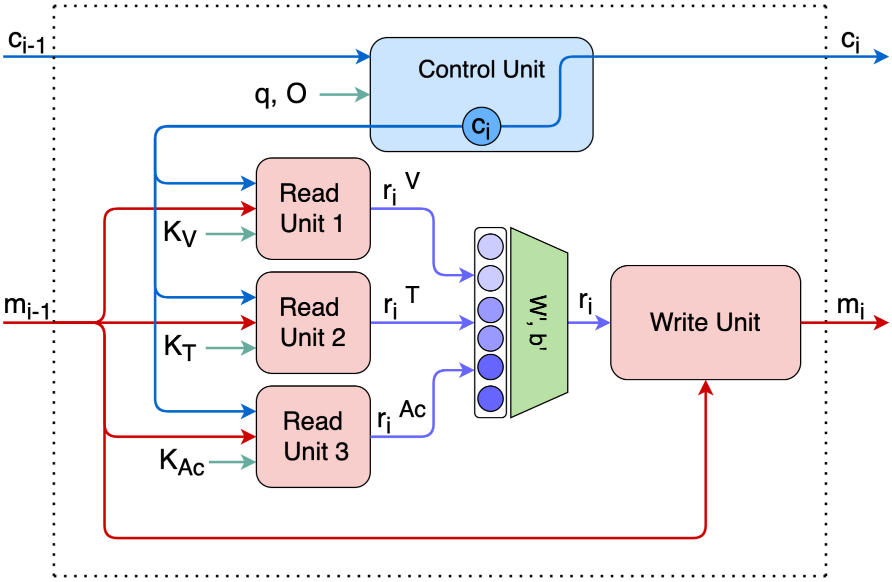


Figure 5.2. *The MAC-X recurrent cell in the i th reasoning step. The multimodal extension of the MAC cell is manifested in the cloning of the Read Unit and consequent fusion of the modalities' extracted information r_i^j before integration to memory m_i.*

is also concatenated in the input vector of equation (5.2). In equation (5.3), attention based on the current control c_i is performed on $I_{i,t}^j$, to create the current r_i^j for each Read Unit.

Multimodal Fusion

In order to perform a mid-level fusion, we fuse modalities at this stage by concatenating the intermediate extracted knowledge results r_i^j for every modality j and passing them through a feedforward layer, effectively constructing a single shared representation layer r_i for all modalities. This is shown in Figure 5.2 and in the equation

$$r_i = W'[r_i^V, r_i^T, r_i^{Ac}] + b' \tag{5.4}$$

Implementing the multimodal fusion at this innermost stage stands in contrast to simpler late fusion methods, a comparison discussed in detail in Section 5.1.6.

Write Unit

The Write Unit (Figure 5.2) integrates the collective information r_i from the Read Units to the previous memory m_{i-1} and thus obtains the current memory m_i.

$$m_i = f_{mr}([m_{i-1}, r_i]) \tag{5.5}$$

In this work we omit the optional components of the Write Unit proposed in [1] since their use did not prove to be significantly beneficial.

5.1.4 Output Unit

After p recurrent iterations of the MAC-X cell as described in the previous sections, the final memory m_p is concatenated with the question representation q to create the context on which the correct answer should be chosen (Figure 5.1). This is further concatenated with each of the answers a_1, a_2 and passed to identical two layer feedforward networks for classification, which output the predictions

$$y_1 = W[q, m_p, a_1] + b, \quad y_2 = W[q, m_p, a_2] + b \tag{5.6}$$

where y_1 and y_2 are the correct and incorrect answer predictions respectively. We then compute the composite loss

$$\mathcal{L} = (\frac{1}{N}\sum_{i=1}^{N} y_1^i - 1)^2 + (\frac{1}{N}\sum_{i=1}^{N} y_2^i)^2 \tag{5.7}$$

where N is the number of samples in a batch. We note that this is the same loss that is exhibited in the original code released for the Social-IQ baseline in [2]. The binary accuracy A2 is formulated as

$$A2 = \frac{1}{M}\sum_{i=1}^{M}(y_1^i > y_2^i) \tag{5.8}$$

where M is the total number of samples in the set for which the accuracy is calculated.

5.1.5 Experimental Setup

In all experiments the validation set is used for evaluation and comparison of the models, as the private test set is reserved by the authors for future challenges. For all input modalities, we use the precomputed embeddings published in [2]. For the LSTM baseline, after all modalities are encoded, they are concatenated and passed directly to the classifiers for final prediction. All experiments with the TMFN baseline are reproduced on the validation set, and the original code released is used. For our model (MAC-X), hyperparameters are set as $p = 12$, $d = 512$, and no self-attention or memory gate mechanisms from [1] are used. All LSTMs are bidirectional, with output dimension d for use in the MAC-X cell. For the comparison to previous state-of-the-art models in Table 5.3, we use their reported results on the validation set. In all experiments, models are trained on 32 samples per batch, with Adam optimizer and learning rate of 10^{-3}, for 10 epochs for LSTM and MAC and 50 epochs for TMFN.

5.1.6 Results and Discussion

We next show the results for the proposed architecture and reproduced baselines. All results are averaged over five runs. Input modalities are denoted as Q for the question, A for the answers, V for the visual frames, T for the dialogue transcript, and Ac for the acoustic input.

In Table 5.1 we compare our model (MAC-X) to the LSTM and TMFN baselines based on the binary accuracy (A2), in an ablation study for different combinations of the input modalities; each combination is denoted by the modalities it makes use of. It is observed that in both baselines multimodality is not necessarily beneficial to performance, and can even degrade it substantially. In contrast, MAC-X performs best when all modalities are used, marking a 0.25% absolute accuracy improvement over its single modality input counterparts, which points to the soundness of its knowledge extraction and fusion methods. At the same time it is very effective in the unimodal input settings, surpassing both the LSTM and TMFN baselines by at least five percentage points. As for the observed importance of each modality, the visual and audio modalities seem to perform best in the LSTM and TMFN baselines respectively, while MAC-X benefits fairly equally from all modalities. In addition, we show that using just the question and answer (or even just the answer) modalities in the LSTM baseline achieves performance well above random, attesting to the existence of language bias in the validation set.

Modalities	**LSTM**	**TMFN**	**MAC-X**
A	63.22 (±0.41)	-	-
QA	64.51 (±0.58)	-	-
QAV	64.82 (±0.67)	65.67 (±0.38)	**71.01** (±0.24)
QAT	64.54 (±0.57)	65.51 (±0.43)	**70.97** (±0.44)
QAAc	64.17 (±0.32)	65.89 (±0.32)	**71.00** (±0.30)
QAVTAc	63.73 (±0.71)	65.62 (±0.55)	**71.25** (±0.15)

Table 5.1. *Ablation study on input modalities and comparison to baseline models, reporting on A2 validation set accuracy.*

In Table 5.2 we present an ablation study that showcases the effectiveness of our mid-level fusion method, outperforming a late fusion baseline in both metrics. In the latter's setting, each modality goes through a completely separate MAC Network, whose outputs are fused at that late stage in the same manner as in our mid-level fusion, before entering the final classifiers. This indicates the advantage of fusing modalities at the intermediate representation stage in the models, where their collective useful information can be jointly processed further.

Models	**A2**	**A4**
MAC w. Late fusion	70.59 (±0.62)	46.46 (±0.26)
MAC-X	**71.25** (±0.15)	**47.22** (±0.60)

Table 5.2. *Ablation study on the multimodal fusion stage, reporting on the validation set with the full set of input modalities.*

In Table 5.3 we measure the performance of our proposed model against five prior state-of-the-art methods, reporting on both metrics for the validation set. We observe a $2.3 - 2.6\%$ accuracy improvement from the previous state-of-the-art in the binary task (MCQA [17]), taking variance into account. As regards the multiple choice task (A4), we obtain comparable results to the best-performing model TACO-Net [18]. Note that TACO-Net measures explicitly the consistency between each answer and modality, contributing to the robustness of the model in the multiple choice setting. Overall, through implementing and applying MAC-X we set a new leading performance for the binary task of the Social-IQ dataset.

Models	**A2**	**A4**
TMFN [2]	65.62	36.24
Removing bias [19]	67.93	-
TACO-Net [18]	68.19	**49.08**
Perceptual score [20]	68.65	-
MCQA [17]	68.80	38.30
Ours (MAC-X)	**71.25** (±0.15)	47.22 (±0.60)

Table 5.3. *Performance comparison to state-of-the-art methods on the Social-IQ validation set. We report averaged results and standard deviation over five runs.*

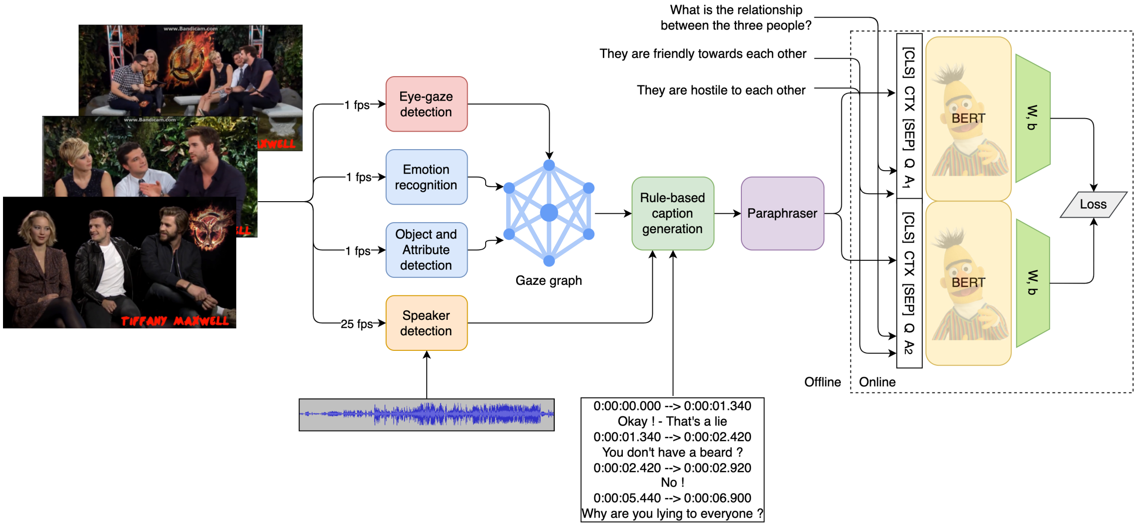
5.2 Augmentation Approach: Emogaze

5.2.1 Overview

Most of previous work regards video as a continuous source over which some attention can be computed. Although this general approach makes sense in most VQA and Video QA tasks, where questions refer to the environment, objects, or actions and events, Social Video QA revolves around people and their interactions. If we were to modularize social video, then these modules would be the people participating in interactions and the information they exchange, both verbal and non-verbal. People communicate through language and non-verbal cues such as facial expressions, looks, gestures and body language.

In addition, drawing inspiration from works such as [16], which state that natural language understanding will only be complete with the integration of multimodal and social semantics, we take the direction of training a purely NLP model with multimodal (WS3) input that contains social (WS5) clues, filtered to contain only gaze data, emotion data, object recognition grounding data, and conversation data, all translated into language, effectively performing captioning.

This is also supported by works for visual and video question answering that take advantage of the superior reasoning inference abilities of the natural language models, compared to computer vision models which tend to perform more instinctive, system 1 processing, as well as bypassing the language bias problem of multimodal models which tend to focus on the textual information since it is easier to get statistical information from (due to the sparse granularity of the linguistic domain), while ignoring information from other modalities. An additional benefit of this approach is more explainability both in success and in failure cases, as the detected attributes in the captions can reflect what the system extracts from the frames, and the intermediate results can help us decouple the inabilities of the scene understanding part from the inabilities of the answer inference part.



What is the relationship between the three people?

They are friendly towards each other

They are hostile to each other

0:00:00.000 --> 0:00:01.340
Okay ! - That's a lie
0:00:01.340 --> 0:00:02.420
You don't have a beard ?
0:00:02.420 --> 0:00:02.920
No !
0:00:05.440 --> 0:00:06.900
Why are you lying to everyone ?

Figure 5.3. *Overview of the proposed augmentation and text question answering architecture, which is composed of an offline caption generation part (left), followed by online training (right). On the left, the three detectors of eye-gaze, emotion, and objects are combined into a structure that can be viewed as the gaze graph. With the addition of the dialogue transcript and the speaker detection, a rule-based caption is produced, optionally followed by a paraphraser model. These video captions are used for multiple choice question answering in pre-trained BERT models followed by classifiers.*

In this work, we choose to enhance the verbal understanding of social interactions with the emotional content of the face carried through eye gaze. In particular, for each detected person s in a frame we find the person t their gaze is targeted at and establish the gaze $g = (s, t)$. This gaze carries an emotional and an optional verbal weight if the person is also the speaker W_g = (face emotion[, uttered phrase]). In addition, with the goal of retaining only the highest social content, gazes are filtered to keep only the speaker and the person they are looking at, resulting in a graph $G_f = \{(s,t),(t,r)\}$ for frame f. All participating nodes (people) can be described in terms of visual appearance, in attribute and object form i.e. $s :=$ attr + obj. From this formulation, we construct natural language descriptions through a set of rules, effectively performing a graph-to-text conversion. These descriptions are optionally processed by a paraphrasing network, to produce more natural and higher variation content. These human-centered video captions are used as input context in NLP BERT-like language models to perform multiple choice question answering on.

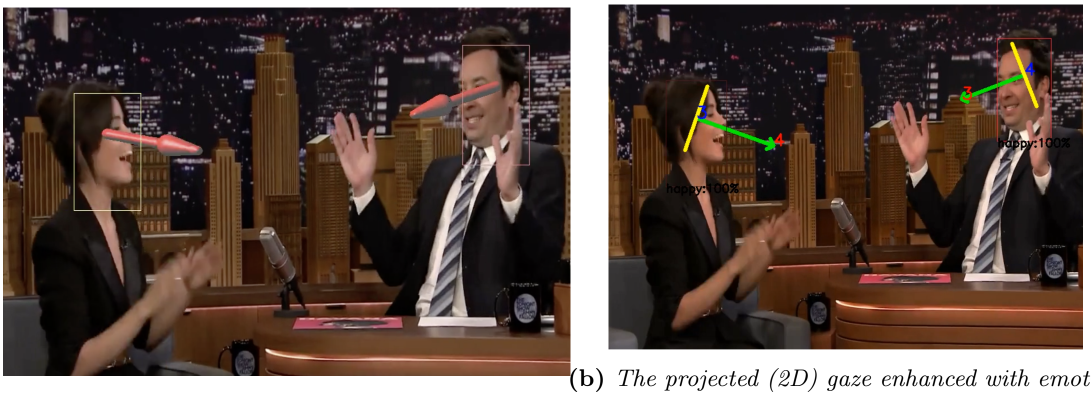
5.2.2 Feature extraction pipeline

In this work, we examine filtering multimodal interaction down to a series of key detections. These correspond to different detection modules that are used to connect multimodal concepts to natural language descriptions, performing in this way a late fusion of their predictions.

Eye gaze

First of all, an important component of social interactions is considered to be the eye-gaze between people participating in them - meaning who looks at whom when uttering a specific phrase or displaying certain characteristics such as emotions. The pretrained model gaze360[1] [106] is chosen for this purpose; it is a model with 3D gaze capabilities. In this work we leverage it for inference and subsequent 2D projection as we don't have 3D data in order to use the extra dimension. Please see the future work section for how this could be adapted.

In the gaze360 inference pipeline, first DensePose[2] [107] is used to detect accurate bounding boxes for the people in the scenes, and keep those above a confidence threshold. The detections are further cropped to contain the person's head. Then a basic tracking via Intersection over Union (IOU) of the head bounding boxes is employed, and a series of bounding box locations for each tracked id are saved. The gaze prediction model takes as input the respective consecutive cropped head images for each tracked id, to calculate their predicted gaze. The output prediction is in the form of relative coordinates for the pointed end of an arrow indicating the gaze's direction; this arrow's stem coordinates are the approximate eye coordinates. These are placed at the center of the head's bounding box width and at 65% of its height. Model outputs are in spherical coordinates and are converted to cartesian before use in the next steps.



(a) *Eye-gaze detection (3D) from gaze-360 [106].*

(b) *The projected (2D) gaze enhanced with emotion information.*

Figure 5.4. *From gaze to emogaze. The gaze detection is projected in two dimensions to compute the target for a given source gaze in the 2D video source.*

[1] https://github.com/erkil1452/gaze360

[2] https://github.com/facebookresearch/detectron2/tree/main/projects/DensePose

We utilize two out of the three output coordinates, corresponding to the axes of the image (x, y) and leave out the coordinate on the plane perpendicular to the image (z). We proceed to compute the target for a person's gaze. This is formulated as either (1) another person or (2) the audience, as we want the descriptions we generate to be as human-centered as possible. In this way, objects are not considered as targets for a gaze. The nearest person is selected via euclidean distance of the candidate targets' eyes to the extended line of the source's gaze, with the additional constraint that this target lies in the (open) half-plane defined by the gaze's direction (due to our eyes' placement in the front of our heads).

$$\hat{t} = \underset{t\in T}{argmin} \frac{|(E_t - G_s) \times (G_e - G_s)|}{||G_e - G_s||}, \quad T = \{t \mid (E_t - G_s) \cdot (G_e - G_s) > 0\} \tag{5.9}$$

The gaze is noted as directed at audience when (1) there is no person in that half plane, or (2) the predicted gaze's absolute z coordinate exceeds a threshold, meaning that it was the dominant direction and that the person is looking closest to the camera.

Emotion

The second component chosen to distil information in social videos is that of face emotion recognition. We follow up on the hypothesis that some of the intent in people's actions and words, as well as their consequences is carried by the emotional state. Often, the cause and effect of people's words and actions can be observed from their emotional state at those moments in time.

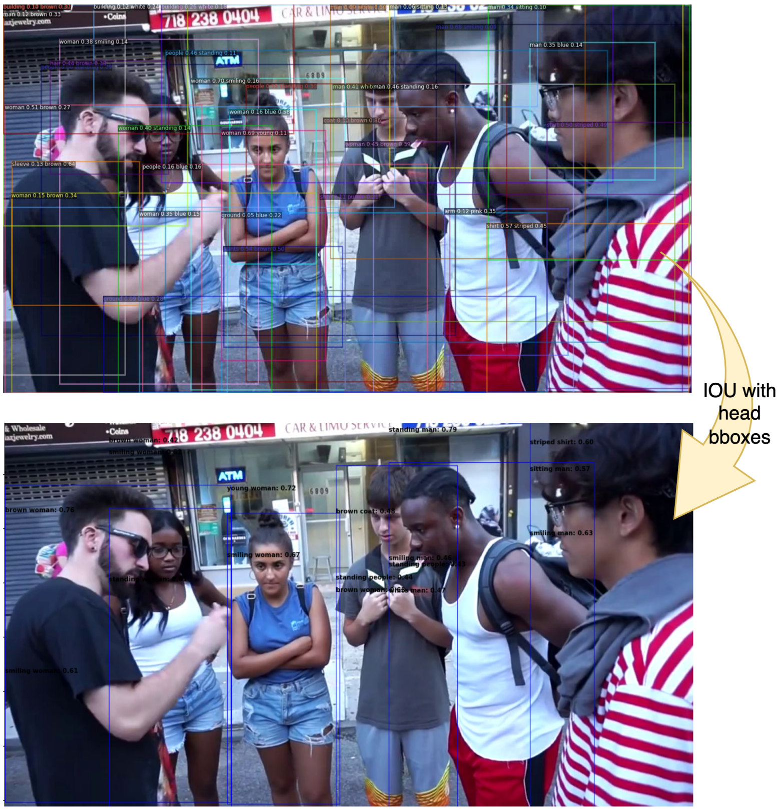


Figure 5.5. *The pre-trained F-RCNN model gives predictions for objects and attributes all over the image. We want to limit this to the social scene participants, so we filter them through an IOU threshold with the people's head bounding boxes.*

To that end, we finetuned Resnet50 [33] on an emotion recognition dataset from kaggle, FER-2013 [108]. The model outputs a prediction for one of 7 emotion classes; neutral, happy, sad, angry, scared, surprised, and disgusted. We use the face detector from FaceNet[3] [109] to detect the faces in the head

[3] https://github.com/timesler/facenet-pytorch

bounding boxes already detected from DensePose, and then feed them to the finetuned model to obtain the predicted emotion.

Object and Attribute detection

A connecting component that helps to identify who takes part in the interactions recorded in a scene, is detected attributes and objects that appear in the area of a detected person, for example what they wear or what they hold. This is in accordance with the Social IQ questions which refer to people in this way (e.g. "the man in the blue shirt"), in contrast to other datasets that use a character's name and provide these annotations in the dialogue transcript.

For this purpose we employed the Faster RCNN [36] pretrained network, finetuned in the Visual Genome dataset[4] in order to predict attributes as well as objects. Of the detections we obtain in each frame, we use those that exceed a confidence threshold to compute the Intersection over Union of their bounding boxes with the detected people from DensePose. This way, we match detected attributes to people.

Speaker detection

Another connnecting component is speaker detection, which is necessary to determine which visually detected person utters each phrase in the dialogue transcript. Using the SyncNet [12] framework we detect the speaker's bounding box in each frame. More specifically, given confidence scores for each tracked person at 25 fps (since this is the frame rate at which SyncNet operates), we choose the person that had the highest overall confidence (sum) for all 25 frames as the speaker at 1 fps. If for all 25 frames no speaker is detected that exceeds a confidence threshold, then no speaker is assigned. At the same time, given the transcript timestamp annotations provided, we group transcript utterances at the level of 1 second and match each of those to the selected speaker at that timestep. If an utterance spans multiple seconds, then it is assigned to the speaker that was detected for the majority of this time. Finally, we connect the detected speaker's bounding box to the respective person's head from DensePose through maximizing the Intersection over Union score, similar to the process in the other submodules.

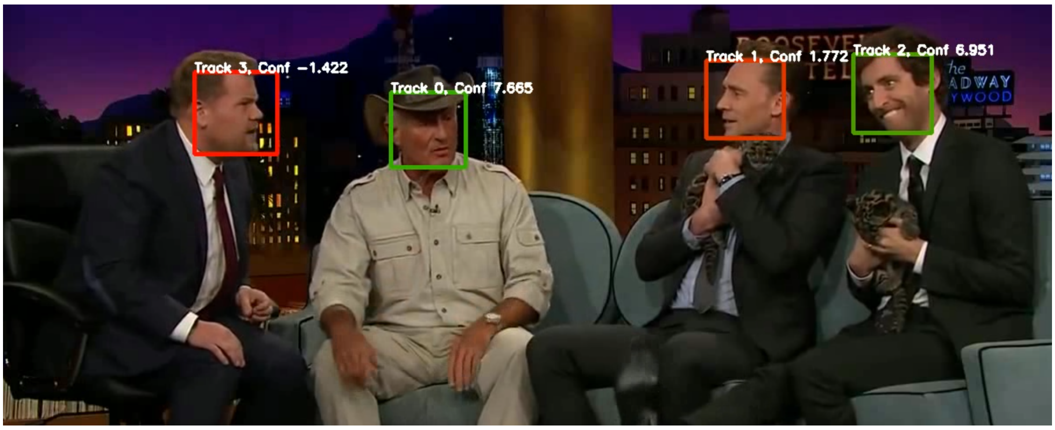


Figure 5.6. *We utilize speaker detection from SyncNet [12], to connect the dialogue utterances to the people in the scene. In this specific scene, the man in the middle is talking about the animals the two men hold, while the man on the right makes a sound of pain while a small leopard bites him.*

Apart from connecting to the transcript, we utilize speaker detection to keep only the useful detected gaze interactions; more specifically we keep only the speaker and the person he is looking at. From a social perspective, the rest of the detected people can be considered as observers or bystanders, and will tend to matter less to the overall storyline.

[4] from unc-nlp/frcnn-vg-finetuned

5.2.3 Video caption generation

Rule-based

To enhance the scene dialogues with our human-centered detections, we take the approach of constructing natural language sentences that describe who says what, and who they are looking at at that time. In this way, we ground our detections to the dialogue utterance, and construct a video caption storyline with social and affective information. To stabilize the detected gaze at the desired utterance timestamp we average it during a window around that time, as utterances can last several seconds and we want to avoid sampling. The rule-based generation used is the following:

```
"At <timestamp>, the <emo><attr> says <trans> and looks at the <target>
[, who looks <target2>]",  where
<target> = <emo2><attr2> | audience,
<target2> = back at <pronoun> | at the (<emo3><attr3> | audience),
and the [ ] mean optional, in case <target> is audience.
```

An example of three such frames after the rule-based conversion to natural language descriptions is:

```
At 0:25 the angry looking sitting young man says "I'mma direct my question to you guys
now" and looks at the audience.
At 0:29 the happy sitting man says "They know already !" and looks at the happy blonde
woman who looks back at him.
At 0:32 the happy blonde hair smiling woman says "Yeah they know already and I'm like "we
already know what's going on !"" and looks at the audience.
```

At this point, descriptions can be optionally grouped into visual scenes (change of camera shot), which are additionally returned by the SyncNet model used for speaker detection. Otherwise, they can be used as a single body of video caption text. Full descriptions usually range from 2,000-5,000 tokenized words.

Paraphrased

The idea here is that giving some variations to a rule based description will make it more natural and therefore closer to natural language pretraining priors in large transformer-based models. In this work we implement paraphrasing via backtranslation [110]. The core concept of backtranslation is that, through translating a sentence to another language and then translating back to the original (with two respective pretrained translation models e.g. transformers), the final output will be a slight variation of the input, but will not have lost its meaning. This is not the case with other models specifically trained for the task of paraphrasing, such as T5 [111], as these models tend to remove useful information and add imagined details. The models used in for backtranslation are the pre-trained transformers for translation to and from german, made available by fairseq[5], WMT18.en-de and WMT19.de-en respectively. For example, the three rule-based frame descriptions above are transformed to the following:

```
At 0:25 the angry seated young guy says, "I\'m asking you now, guys," and looks into
the audience.
At 0:29 "They already know!" says the cheerfully seated man, looking at the happy blonde
woman who is looking at him.
At 0:32 the cheerful blonde, smiling woman says: "Yes, they already know and I\'m like" we
already know what\'s going on!" and looks into the audience.
```

while T5 outputs:

```
"I'mma direct my question to you guys now" and gazes at the audience at 0:25.
At 0:29, the contentious sitting man looks at the smiling blonde woman who is
```

[5] https://github.com/facebookresearch/fairseq

```
smiling at him.
"They know already know what's going on," the happy blonde hair smiling woman
says at 0:32, and I'm like, "we already know what's going on.".
```

5.2.4 Proposed models

The basic idea is to experiment with using embeddings from large pretrained BERT-like language models, or selectively finetune some of their last layers, for the downstream task of multiple choice question answering. This is because our created augmented dataset is small (only 1015 unique video descriptions), and can be observed to have very different distribution from the corpus on which BERT is pretrained on. We experiment with two models, the original BERT [3] base, and the RoBERTa [42] large, previously finetuned for another text multiple choice question answering dataset, RACE [21]. The latter is chosen with the goal of starting with better weight initializations for the same task, and therefore containing the transfer learning scenario to just different domains instead of both different domain and task. Additionally we experiment with a more aggressive finetuning scheme for the smaller model DistilBERT [43].

Another issue when using such language models for interpreting large volumes of text is that the maximum input length is limited, e.g. at 512 or 1024 tokens depending on the model, which results in truncating the context to only the first tokens. To combat this, we experimented with summarization via extractive question answering, using BERT previously finetuned on an extractive question answering dataset, SQUAD [22]. The idea is to extract useful information from the input context, conditioned on the question, and move to answering the multiple choice based on that (smaller) extract. However, extractive question answering also has a maximum input length, which leads to idea of performing these summaries in a hierarchical manner on smaller parts of the context, and then on the concatenated result of extracts. These smaller parts can be obtained either by using visual scene segmentation (camera shot change detection from SyncNet output) or splitting into equal parts.

The way the question answering components are inputted in a BERT-like model can be described as follows, for the binary task:

$$\begin{aligned} E_1 &= bert([CLS] + CTX + [SEP] + Q + A_1) \\ E_2 &= bert([CLS] + CTX + [SEP] + Q + A_2) \end{aligned} \tag{5.10}$$

where CTX is the context that corresponds to the video description we generated, E_1 and E_2 are the final layer CLS token embeddings, and $+$ denotes concatenation. The output logits $Y_1, Y_2 = f(E_1), f(E_2)$ where f is a linear classifier are fed into standard cross-entropy loss together with the ground truth label.

5.2.5 Experimental Setup

In all experiments the validation set is used for evaluation and comparison of the models, as the private test set is reserved by the authors for future challenges.

The hyperparameters for the feature extraction pipeline described in the respective section above are described in detail as follows. For the eye-gaze detection, videos are first resampled at 1fps and the densepose backbone used for the human bounding boxes is the rcnn r50 fpn, with a threshold confidence of 0.8 for denspose person detection. The threshold for the gaze-360 bounding box IOU tracking is set to 0.5, and after processing by the gaze-360 model, the threshold for z coordinate to be dominant and therefore pointed at the camera is 0.9. For the emotion detector, we performed resnet50 finetuning on imagenet black and white images before finetuning in the emotion dataset. For the object and attribute detection, the threshold confidence for frcnn detections is set to 0.3 and the threshold confidence for the IOU between the frcnn and head bounding boxes for matching attributes to persons is 0.4, where there is an additional limit to keep a maximum of 5 matched detections per person. For the speaker detection, videos are resampled at 25 fps and the speaker is chosen with a confidence score of 5. Finally, at the rule-based generation gaze averaging is performed on a window of 3 seconds.

BERT embedding classifier is composed of two linear layers separated by tanh activation and dropout of p=0.1, according to the BertForMultipleChoice implementation from huggingface[6]. The BERT and DistilBERT models used are the base uncased provided by huggingface, and the RoBERTa RACE is the large finetuned on RACE provided by LIAMF-USP[7]. The BERT model used for extractive question answering is the base uncased finetuned on SQUAD2 from the huggingface community. BERT model layers and classifiers are trained for 20 epochs and batch size 32, with AdamW optimizer and learning rate of $5 * 10^{-5}$, using a linear scheduler with no warmup steps.

5.2.6 Results and Discussion

In Table 5.4 we present our results in an ablation study for using different levels of input augmentation, comparing the performance of two pretrained language models, BERT and RoBERTa RACE, the latter of which is pretrained in the multiple choice text QA dataset RACE [21]. The different inputs tested for ablation are the following. First the simple QA-only baseline, where in equation 5.10 the context is the empty string, as well as the QA with the dialogue transcript given in the dataset as context. Then, to test our augmentation pipeline, we present results for both the rule-based context as well as the paraphrased context. We remind that the rule-based context includes both transcript and augmentation information such as grounding and emotional gaze (emogaze), and that the paraphrased context (ctx) is the output of the paraphraser with the rule-based context as input.

In these experiments, a classifier is trained on the last layer embeddings (from the [CLS] token), and results are reported for the binary (A2) accuracy task, with mean and standard deviation over 5 runs. In addition, the context (CTX) is cropped when it exceeds the language model's maximum input tokens size. We can see that the BERT model seems to benefit more from exploiting the language bias in the questions and answers, than from using any augmentation input from the videos, which can even result in a performance drop. This is suspected to be due to the fact that it hasn't been fine-tuned for any context-based multiple-choice downstream task, and simply ignores longer sequences of input instead of attending to parts that are relevant to the question. On the contrary, RoBERTa RACE performs best when given our rule-based and paraphrased context as input. It is also interesting that the performance of the RoBERTa model drops when the simple dialogue transcript is used, but when this is grounded and enhanced with our detections it surpasses the QA baseline by 0.5%. Furthermore, the paraphrasing enhancement is observed to improve the rule-based context, which is indicative of the benefit of more natural and diverse video descriptions.

Inputs	**BERT**	**RoBERTa RACE**
QA	74.37 (±0.06)	79.05 (±0.05)
QA+transcript	74.34 (±0.05)	78.50 (±0.06)
QA+emo	**74.87** (±0.06)	79.24 (±0.05)
QA+rule-based ctx	73.12 (±0.05)	79.43 (±0.06)
QA+paraphrased ctx	74.21 (±0.07)	**79.56** (±0.06)

Table 5.4. *Ablation study for training a classifier on embeddings (CLS last layer) on different levels of input augmentation. Results are reported on validation set accuracy on average and standard deviation over 5 runs.*

However when we experiment with finetuning the encoders' layers to different extents and with different language models, it appears that the unaugmented input performs better (either only QA or QA with the dialogue transcript as context) in all cases. Since this was not the case with the frozen RoBERTa model in table 5.4, we are lead to hypothesize that, apart from the fact that the self-attention layers' previous training was on a much different distribution than our own descriptions, that the pre-training procedure (simple MLM) on which they are conditioned on is not enough to model the relations between the detections

[6] https://huggingface.co/

[7] https://huggingface.co/LIAMF-USP/roberta-large-finetuned-race

in our descriptions, which follow the repetitive format of section 5.2.3. Ideas on how effective finetuning could be performed are explored in detail in section 6.2.

Inputs	**3-BERT**	**1-RoBERTa RACE**	**5-DistilBERT**
QA	80.11	81.99	**79.83**
QA+transcript	**81.39**	**82.37**	79.28
QA+rule-based ctx	81.14	81.50	77.64
QA+paraphrased ctx	81.06	81.74	78.89

Table 5.5. *Ablation study for selectively fine-tuning layers in different language models on different levels of input augmentation. Results are reported on validation set accuracy.*

In table 5.6 we evaluate the effect of utilizing smaller extracts, or summaries, as input via different extraction schemes, again in training a classifier on BERT's last layer embeddings. This extraction is performed by extractive question answering, again using a pretrained BERT model, this time on the SQUAD [22] dataset. The different input schemes included in this ablation are, first the simple QA-only baseline and the input including our rule-based context truncated to the maximum input length, same as in table 5.4. In "QA+extract" a smaller extract is used as context along with the question and answer. This comes from a single extraction from the whole rule-based context, in which case the context, as input to the extractive-answering BERT is also truncated to the maximum input token size. In "QA+equal-parts-hierarchical extract", the input is splitted in equal parts, for each of which the extractive QA is performed, and the resulting extracts are concatenated to be used as input for another extractive QA inference. The final extract is used as context for the multiple choice question answering, again along with the question and answer. Similarly, in "QA+visual-scenes-hierarchical extract", the visual scenes from the SyncNet model [12] are used to split the input, followed by the same procedure.

So far, we have observed that with the rule-based context as additional input to the question and answer, the BERT base model actually drops in performance, which is hypothesised to be attributed to the lack of finetuning on large context question answering tasks. This is somewhat alleviated by the single extract mode, but still no significant improvement can be observed from the simple QA-only baseline. As the extract is still taken from a cropped context, there is no additional input seen by the model than in the direct question answering on the cropped context, and no additional benefit from this apart from solving the large input problem. When using the hierarchical scheme to create the final extract, we observe an improvement of 0.3% over the simple QA-only baseline, which can be attributed to the sampling of sentences from all over the context instead of simply the start, with the additional benefit of this sampling being conditioned on the question. When comparing the splitting into equal parts or visual scene parts, the equal parts seem to have better performance, although the visual scene parts still exceed the QA baseline.

Inputs	BERT
QA	74.37
QA+rule-based ctx	73.12
QA+extract	74.39
QA+equal-parts-hierarchical extract	**74.64**
QA+visual-scenes-hierarchical extract	74.43

Table 5.6. *Comparison of different summarization schemes, with the cropped rule-based context and simple QA-only baselines. The final multiple choice BERT input is the extract, while the input of the extractive question answering BERT is the rule-based context. Only the classifier is trained in the main model.*

Finally, we experiment with enhancing the question and answers with a single emotion word aggregated from the detections of the whole video, to probe the sensitivity of the dataset's questions on the overall detected emotion of each video's participants. Three different aggregation methods are tested, each choosing the emotion with the highest of the following metrics. Firstly the emotion's term frequency (tf), secondly the term frequency - inverse document frequency (tf-idf) which normalizes the term frequency of the emotion in the video by the number of videos of the training set it appears in, and thirdly a modified tf-idf (tf-idfm), where the term frequency is normalized by the frequency of the word in the training set instead of number of videos with appearance.

Inputs	BERT
QA	74.37 (±0.06)
QA+tf-emo	**74.87** (±0.06)
QA+tfidf-emo	74.81 (±0.06)
QA+tfidfm-emo	74.86 (±0.06)

Table 5.7. *Training a classifier on embeddings of the questions and answers, enhanced only with the aggregated emotion tag of the video. Results are mean and deviation over 5 runs.*

All three of these metrics for aggregation of emotion seem to perform comparably, and a 0.5% improvement over the QA-only baseline can be achieved simply by using a single emotion word to describe the video. This goes to show that Social-IQ has a big dependence on the emotion information, and that it should be further enhanced in future work.

Chapter 6

Conclusions

6.1 Discussion

In this work, we explore two very different approaches to the task of Social Video Question Answering, specifically for the Social-IQ dataset [2]. This was done to specifically explore capabilities that are often missing from machine learning systems but are much needed especially in Social Video QA, such as explicit reasoning operations and social cues detection.

In the first approach, we follow an end-to-end training scheme using pretrained CNN features and audio features, and perform modality fusion through an extension of the MAC network [1] which we have called MAC-X. More specifically we present MAC-X, a multimodal extension of the MAC Network capable of handling complex multiple choice and multiple modality reasoning tasks like Social-IQ, where we evaluate it and obtain state-of-the-art results. We conclude that structural priors as well as compositional reasoning can prove useful to Social Video Question Answering, in which - to the best of our knowledge - this direction is applied for the first time. We can further confirm from our ablation studies that MAC-X can effectively benefit from all modalities and that mid-level fusion performs considerably better than the late fusion baselines.

In the second approach, we follow the direction of performing question answering on captioning, which we obtain through augmentation of the dialogue transcripts with emotional gaze information as well as visual grounding attributes information. The emotional gaze connects the people involved in a social scene, which are filtered from bystanders via speaker detection, which also connects them to the dialogue transcript utterances. This is the first time to the best of our knowledge, that such a feature extraction pipeline specifically designed for social video, is proposed. We suggest that it provides a general framework for leveraging social information in video. Additionally, through augmenting the dialogue transcripts and effectively performing video captioning, we also provide a baseline on which to base more sophisticated social video captioning methods. Finally, we provide ablation studies for several BERT-like language models and fine-tuning levels, as well as a hierarchical summary scheme based on question conditioning via extractive question answering.

6.2 Future Work

End-to-end

First, an idea that was motivated by the analysis of intra and inter modality attention maps in chapter 3 was to use the resulting attention maps of models such as ViLBERT [8] (or any other models that calculate the alignment between modalities), as graph representations where each attention weight corresponds to the weighted connection between two nodes, which can either be homogenous or heterogenous in terms of modality, or alternatively form different subgraphs (as in works such as [112]) for each attention map. These graphs would be processed either by a GNN or a soft neurosymbolic approach using graphs such as the Neural State Machine [15], which is similar to MAC [1] but instead performs its sequential reasoning as an iterative computation of a differentiable state machine over a semantic graph.

As far as MAC-X is concerned, we plan on further experimenting with a hierarchical-MAC version where each frame is processed separately by sub-MAC networks, and the final memories from each of the frames' forward passes are processed by the main MAC network. Modality fusion can manifest either in the deeper per-frame level, which would make each sub-network a MAC-X network and the outer a vanilla MAC network, or in the outer level with a different MAC network for each modality at the frame level, and a MAC-X network to combine and fuse their respective final memories. Additionally, we plan on investigating more sophisticated techniques of mid-level fusion for the purpose of learning better intermediate multimodal representations, as well as explore a more tailored modelling of the multiple choice task [18].

Another idea for an alternative end-to-end approach is to fine-tune recent promising large scale video and language transformer models, such as CLIP-BERT [66] and HERO [91], which offer affordable end-to-end learning for video-and-language tasks through methods like sparse sampling and hierarchical transformers designed and tailored for the temporal dimension.

Augmentation

In our second approach, our utilization of the gaze prediction comes with two inherent problems. The first is the percentage of "social" videos with only one person per frame (this can happen in TV pannels) which renders gaze useless as they always look at the camera, and the second is the exact opposite case of videos with more than three people per frame, where there is an actual need for 3D space and depth, as the recipient of the gaze is often calculated incorrectly due to the 2D projection. An interesting approach since we have 3D gaze predictions [106] but not 3D video, is to use video depth estimation [113] to differentiate between gaze targets in different depths.

As far as quality improvements are concerned, there is much room for improvement, including the cases of inaccurate face emotion predictions due to training in a fairly small dataset [108], or the speaker identity when the speaker is not on the current scene / video frame. These can be approached respectively through finetuning in a larger emotion dataset such as Affectnet [114], and utilizing speaker diarization techniques in addition to speaker detection [115]. Furthermore, as for additional social cues that could be detected from the video, some ideas would be body language (like gestures and posture) [116, 117], voice prosody characteristics [118], and additionally, action recognition as further grounding reference. For the video caption generation, some ideas for enhancement are to utilize some kind of rule guided GPT [119] story generation to make descriptions more natural and randomized, as well as document retrieval from knowledge bases for additional common sense and world knowledge information.

To improve the utilization of our video captioning intermediate results by the question answering model, we plan on experimenting with different schemes such as a sliding-window BERT configuration or a voting scheme on separate per-scene answer predictions [66, 95]. Another idea based on the need to guide BERT more towards learning our specific rule based limited syntax (which is assumed to be outside the distribution of its pre-training data), is to use self-supervised pretraining such as Masked Language Modeling [120]. An interesting detail is that there are enough samples as this will be done on sentence level and not video level. Based on this concept, we would also like to explore masking similar to the relationship-attribute-object scheme used in ERNIE-ViL [64] as it is very similar to the structure of our sentences. Finally, we plan on further pretraining in other QA datasets, both text QA and through applying the feature extraction pipeline in e.g. TVQA [81], to then use in the smaller Social-IQ with limited fine-tuning.

As respects the different approaches apart from translating the inferred gaze graph into language, a very important idea to explore is using a Spatio-temporal GNN such as [121] which is specifically meant for this kind of data. For this to be implemented, a more sophisticated face tracking method needs to be employed (e.g. with face similarity) so that the same people correspond to the same nodes throughout the video, regardless of scene changes. Similarly, a soft neurosymbolic approach like [15] extended in the temporal dimension could be combined with the augmentative gaze graph approach.

Use cases

A very interesting use case of a performant social video question answering system would be to help people with ASD with an online system to which they can ask their questions and get answers, either about videos online, shows and movies, or about their real-time surrounding social situations given access through camera in an IoT device such as smart glasses.

6.3 Ethical Considerations

Some ethical considerations regarding both Social VQA, the current dataset, as well as our proposed methods and use cases are the following.

Firstly, a system such as the one described above could end up having access to people's footage without their consent, as well as making impactful wrong predictions that could negatively influence people's relationships in the real world.

Secondly, this particular dataset categorizes (and refers to) people as men and women which is a bad human bias to insert in AI agents, as it characterizes their gender identity based on their external characteristics. This is a broader issue in computer vision datasets as by definition they only have access to visual information, and removing this grounding from the social descriptions would result in poorer information regarding who is who in the video.

Third, an important aspect to consider is that especially with such data and tasks, fairness and bias analysis are necessary before use in real world applications. Such analyses can be separated in bias analysis of the data, and bias analysis of the models [122]. In the first, some things to investigate could be, for example, with what words does the word woman, man, black, white appear in most in the correct answer, and what in the question. Such analyses for the data in the case of a two-stage solution such as our augmentation approach should be applied in the intermediate results as well, to examine what biases the feature extraction models impose. The second kind of analysis refers to which of the existing biases in the data are reinforced by the proposed models, as well as what they carry from their pre-training data. For example, one could seek answers to questions such as what percentage of answers correctly answered contain the word woman, and which man, and whether this is consistent with the whole set of correct answers, which answers incorrectly answered had the word woman in the correct answer and which man (consistency with incorrect), and which questions with the word woman in the questions were incorrectly answered and which with the word man.

We end this thesis by reinforcing the importance of these ethical considerations, if and when such a system is considered for real world use-cases.

Bibliography

[1] Drew A. Hudson και Christopher D. Manning. *Compositional Attention Networks for Machine Reasoning. arXiv:1803.03067 [cs]*, 2018. arXiv: 1803.03067.

[2] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong και Louis Philippe Morency. *Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 8799–8809, Long Beach, CA, USA, 2019. IEEE.

[3] Jacob Devlin, Ming Wei Chang, Kenton Lee και Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Τεχνική Αναφορά με αριθμό arXiv:1810.04805, arXiv, 2019. arXiv:1810.04805 [cs] type: article.

[4] Alex Krizhevsky, Ilya Sutskever και Geoffrey E Hinton. *ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems*, τόμος 25. Curran Associates, Inc., 2012.

[5] Karen Simonyan και Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition.* Τεχνική Αναφορά με αριθμό arXiv:1409.1556, arXiv, 2015. arXiv:1409.1556 [cs] type: article.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser και Illia Polosukhin. *Attention Is All You Need.* Τεχνική Αναφορά με αριθμό arXiv:1706.03762, arXiv, 2017. arXiv:1706.03762 [cs] type: article.

[7] Tadas Baltrušaitis, Chaitanya Ahuja και Louis Philippe Morency. *Multimodal Machine Learning: A Survey and Taxonomy. arXiv:1705.09406 [cs]*, 2017. arXiv: 1705.09406.

[8] Jiasen Lu, Dhruv Batra, Devi Parikh και Stefan Lee. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. arXiv:1908.02265 [cs]*, 2019. arXiv: 1908.02265.

[9] Drew A. Hudson και Christopher D. Manning. *GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 6693–6702, Long Beach, CA, USA, 2019. IEEE.

[10] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi και Kentaro Inui. *Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, σελίδες 7057–7075, Online, 2020. Association for Computational Linguistics.

[11] Zhanwen Chen, Shiyao Li, Roxanne Rashedi, Xiaoman Zi, Morgan Elrod-Erickson, Bryan Hollis, Angela Maliakal, Xinyu Shen, Simeng Zhao και Maithilee Kunda. *Characterizing Datasets for Social Visual Question Answering, and the New TinySocial Dataset. arXiv:2010.11997 [cs]*, 2020. arXiv: 2010.11997.

[12] Joon Son Chung και Andrew Zisserman. *Out of Time: Automated Lip Sync in the Wild. Computer Vision – ACCV 2016 Workshops*Chu Song Chen, Jiwen Lu και Kai Kuang Ma, επιμελητές, Lecture Notes in Computer Science, σελίδες 251–263, Cham, 2017. Springer International Publishing.

[13] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia και Davide Testuggine. *The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. arXiv:2005.04790 [cs]*, 2020.

[14] Noam Chomsky. *Aspects of the Theory of Syntax.* MIT Press, Cambridge, MA, USA, 1965.

[15] Drew A. Hudson και Christopher D. Manning. *Learning by Abstraction: The Neural State Machine. arXiv:1907.03950 [cs]*, 2019. arXiv: 1907.03950.

[16] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto και Joseph Turian. *Experience Grounds Language. arXiv:2004.10151 [cs]*, 2020. arXiv: 2004.10151.

[17] Abhishek Kumar, Trisha Mittal και Dinesh Manocha. *MCQA: Multimodal Co-attention Based Network for Question Answering. arXiv:2004.12238 [cs]*, 2020. arXiv: 2004.12238.

[18] Lingyu Zhang και Richard J. Radke. *Temporal Attention and Consistency Measuring for Video Question Answering. Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, σελίδες 510–518, New York, NY, USA, 2020. Association for Computing Machinery.

[19] Itai Gat, Idan Schwartz, Alexander Schwing και Tamir Hazan. *Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies. arXiv:2010.10802 [cs]*, 2020. arXiv: 2010.10802.

[20] Itai Gat, Idan Schwartz και Alexander Schwing. *Perceptual Score: What Data Modalities Does Your Model Perceive? arXiv:2110.14375 [cs]*, 2021. arXiv: 2110.14375.

[21] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang και Eduard Hovy. *RACE: Large-scale ReAding Comprehension Dataset From Examinations. arXiv:1704.04683 [cs]*, 2017. arXiv: 1704.04683.

[22] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev και Percy Liang. *SQuAD: 100,000+ Questions for Machine Comprehension of Text.* Τεχνική Αναφορά με αριθμό arXiv:1606.05250, arXiv, 2016. arXiv:1606.05250 [cs] type: article.

[23] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick και Devi Parikh. *VQA: Visual Question Answering. 2015 IEEE International Conference on Computer Vision (ICCV)*, σελίδες 2425–2433, Santiago, Chile, 2015. IEEE.

[24] Christopher Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[25] Y. Lecun, L. Bottou, Y. Bengio και P. Haffner. *Gradient-based learning applied to document recognition. Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[26] David E. Rumelhart, Geoffrey E. Hinton και Ronald J. Williams. *Learning Internal Representations by Error Propagation.* Τεχνική Αναφορά με αριθμό, CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE, 1985.

[27] Sepp Hochreiter και Jürgen Schmidhuber. *Long Short-Term Memory. Neural Computation*, 9(8):1735–1780, 1997.

[28] Dzmitry Bahdanau, Kyunghyun Cho και Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate.* Τεχνική Αναφορά με αριθμό arXiv:1409.0473, arXiv, 2016. arXiv:1409.0473 [cs, stat] type: article.

[29] Minh Thang Luong, Hieu Pham και Christopher D. Manning. *Effective Approaches to Attention-based Neural Machine Translation.* Τεχνική Αναφορά με αριθμό arXiv:1508.04025, arXiv, 2015. arXiv:1508.04025 [cs] type: article.

[30] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel και Yoshua Bengio. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Proceedings of the 32nd International Conference on Machine Learning*, σελίδες 2048–2057. PMLR, 2015.

[31] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould και Lei Zhang. *Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. arXiv:1707.07998 [cs]*, 2018. arXiv: 1707.07998.

[32] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li και Li Fei-Fei. *ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition*, σελίδες 248–255, 2009. ISSN: 1063-6919.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren και Jian Sun. *Deep Residual Learning for Image Recognition.* Τεχνική Αναφορά με αριθμό arXiv:1512.03385, arXiv, 2015. arXiv:1512.03385 [cs] type: article.

[34] Gao Huang, Zhuang Liu, Laurensvan der Maaten και Kilian Q. Weinberger. *Densely Connected Convolutional Networks.* Τεχνική Αναφορά με αριθμό arXiv:1608.06993, arXiv, 2018. arXiv:1608.06993 [cs] type: article.

[35] Ross Girshick. *Fast R-CNN.* Τεχνική Αναφορά με αριθμό arXiv:1504.08083, arXiv, 2015. arXiv:1504.08083 [cs] type: article.

[36] Shaoqing Ren, Kaiming He, Ross Girshick και Jian Sun. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.* Τεχνική Αναφορά με αριθμό arXiv:1506.01497, arXiv, 2016. arXiv:1506.01497 [cs] type: article.

[37] Ross Girshick, Jeff Donahue, Trevor Darrell και Jitendra Malik. *Rich feature hierarchies for accurate object detection and semantic segmentation.* Τεχνική Αναφορά με αριθμό arXiv:1311.2524, arXiv, 2014. arXiv:1311.2524 [cs] type: article.

[38] Tomas Mikolov, Kai Chen, Greg Corrado και Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space.* Τεχνική Αναφορά με αριθμό arXiv:1301.3781, arXiv, 2013. arXiv:1301.3781 [cs] type: article.

[39] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee και Luke Zettlemoyer. *Deep contextualized word representations.* Τεχνική Αναφορά με αριθμό arXiv:1802.05365, arXiv, 2018. arXiv:1802.05365 [cs] type: article.

[40] Alec Radford και Karthik Narasimhan. *Improving Language Understanding by Generative Pre-Training. undefined*, 2018.

[41] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever και Dario Amodei. *Language Models are Few-Shot Learners.* Τεχνική Αναφορά με αριθμό arXiv:2005.14165, arXiv, 2020. arXiv:2005.14165 [cs] type: article.

[42] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer και Veselin Stoyanov. *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* Τεχνική Αναφορά με αριθμό arXiv:1907.11692, arXiv, 2019. arXiv:1907.11692 [cs] type: article.

[43] Victor Sanh, Lysandre Debut, Julien Chaumond και Thomas Wolf. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* Τεχνική Αναφορά με αριθμό arXiv:1910.01108, arXiv, 2020. arXiv:1910.01108 [cs] type: article.

[44] Pradeep Atrey, M. Hossain, Abdulmotaleb El Saddik και Mohan Kankanhalli. *Multimodal fusion for multimedia analysis: A survey. Multimedia Syst.*, 16:345–379, 2010.

[45] Ekaterina Shutova, Douwe Kiela και Jean Maillard. *Black Holes and White Rabbits: Metaphor Identification with Visual Features. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, σελίδες 160–170, San Diego, California, 2016. Association for Computational Linguistics.

[46] Emilie Morvant, Amaury Habrard και Stéphane Ayache. *Majority Vote of Diverse Classifiers for Late Fusion. Structural, Syntactic, and Statistical Pattern Recognition*Pasi Fränti, Gavin Brown, Marco

Loog, Francisco Escolano και Marcello Pelillo, επιμελητές, Lecture Notes in Computer Science, σελίδες 153–162, Berlin, Heidelberg, 2014. Springer.

[47] Youssef Mroueh, Etienne Marcheret και Vaibhava Goel. *Deep multimodal learning for Audio-Visual Speech Recognition. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 2130–2134, 2015. ISSN: 2379-190X.

[48] Wanli Ouyang, Xiao Chu και Xiaogang Wang. *Multi-source Deep Learning for Human Pose Estimation.* σελίδες 2329–2336, 2014.

[49] Zuxuan Wu, Yu Gang Jiang, Jun Wang, Jian Pu και Xiangyang Xue. *Exploring Inter-feature and Inter-class Relationships with Deep Neural Networks for Video Classification. Proceedings of the 22nd ACM international conference on Multimedia*, MM '14, σελίδες 167–176, New York, NY, USA, 2014. Association for Computing Machinery.

[50] A.J. Hunt και A.W. Black. *Unit selection in a concatenative speech synthesis system using a large speech database. 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, τόμος 1, σελίδες 373–376 vol. 1, 1996. ISSN: 1520-6149.

[51] T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi και K. Tokuda. *Text-to-visual speech synthesis based on parameter generation from HMM. Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, τόμος 6, σελίδες 3745–3748 vol.6, 1998. ISSN: 1520-6149.

[52] Atsuhiro Kojima, Takeshi Tamura και Kunio Fukunaga. *Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions. International Journal of Computer Vision*, 50(2):171–184, 2002.

[53] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R.G. Lanckriet, Roger Levy και Nuno Vasconcelos. *A new approach to cross-modal multimedia retrieval. Proceedings of the 18th ACM international conference on Multimedia*, MM '10, σελίδες 251–260, New York, NY, USA, 2010. Association for Computing Machinery.

[54] Oriol Vinyals, Alexander Toshev, Samy Bengio και Dumitru Erhan. *Show and Tell: A Neural Image Caption Generator.* Τεχνική Αναφορά με αριθμό arXiv:1411.4555, arXiv, 2015. arXiv:1411.4555 [cs] type: article.

[55] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney και Kate Saenko. *Translating Videos to Natural Language Using Deep Recurrent Neural Networks.* Τεχνική Αναφορά με αριθμό arXiv:1412.4729, arXiv, 2015. arXiv:1412.4729 [cs] type: article.

[56] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen και Ilya Sutskever. *Zero-Shot Text-to-Image Generation.* Τεχνική Αναφορά με αριθμό arXiv:2102.12092, arXiv, 2021. arXiv:2102.12092 [cs] type: article.

[57] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu και Mark Chen. *Hierarchical Text-Conditional Image Generation with CLIP Latents.* Τεχνική Αναφορά με αριθμό arXiv:2204.06125, arXiv, 2022. arXiv:2204.06125 [cs] type: article.

[58] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra και Devi Parikh. *Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering.* σελίδα 10, 2017.

[59] Aishwarya Agrawal, Dhruv Batra, Devi Parikh και Aniruddha Kembhavi. *Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 4971–4980, Salt Lake City, UT, 2018. IEEE.

[60] Liunian Harold Li, Mark Yatskar, Da Yin, Cho Jui Hsieh και Kai Wei Chang. *VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv:1908.03557 [cs]*, 2019. arXiv: 1908.03557.

[61] Yen Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng και Jingjing Liu. *UNITER: UNiversal Image-TExt Representation Learning. arXiv:1909.11740 [cs]*, 2020. arXiv: 1909.11740.

[62] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso και Jianfeng Gao. *Unified Vision-Language Pre-Training for Image Captioning and VQA. arXiv:1909.11059 [cs]*, 2019. arXiv: 1909.11059.

[63] Hao Tan και Mohit Bansal. *LXMERT: Learning Cross-Modality Encoder Representations from Transformers. arXiv:1908.07490 [cs]*, 2019. arXiv: 1908.07490.

[64] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu και Haifeng Wang. *ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph. arXiv:2006.16934 [cs]*, 2021. arXiv: 2006.16934.

[65] Sruthy Manmadhan και Binsu C. Kovoor. *Visual question answering: a state-of-the-art review. Artificial Intelligence Review*, 53(8):5705–5745, 2020.

[66] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal και Jingjing Liu. *Less is More: ClipBERT for Video-and-Language Learning via Sparse Sampling. arXiv:2102.06183 [cs]*, 2021. arXiv: 2102.06183.

[67] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell και Dawn Song. *Fooling Vision and Language Models Despite Localization and Attention Mechanism. arXiv:1709.08693 [cs]*, 2017. arXiv: 1709.08693 version: 1.

[68] Ilija Ilievski, Shuicheng Yan και Jiashi Feng. *A Focused Dynamic Attention Model for Visual Question Answering. arXiv:1604.01485 [cs]*, 2016. arXiv: 1604.01485.

[69] Justin Johnson, Bharath Hariharan, Laurensvan der Maaten, Li Fei-Fei, C. Lawrence Zitnick και Ross Girshick. *CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. arXiv:1612.06890 [cs]*, 2016. arXiv: 1612.06890.

[70] Caiming Xiong, Stephen Merity και Richard Socher. *Dynamic Memory Networks for Visual and Textual Question Answering.* σελίδα 10, 2016.

[71] Jin Hwa Kim, Sang Woo Lee, Dong Hyun Kwak, Min Oh Heo, Jeonghee Kim, Jung Woo Ha και Byoung Tak Zhang. *Multimodal Residual Learning for Visual QA. arXiv:1606.01455 [cs]*, 2016. arXiv: 1606.01455.

[72] Hyeonseob Nam, Jung Woo Ha και Jeonghee Kim. *Dual Attention Networks for Multimodal Reasoning and Matching. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 2156–2164, Honolulu, HI, 2017. IEEE.

[73] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao και Qi Tian. *Deep Modular Co-Attention Networks for Visual Question Answering. arXiv:1906.10770 [cs]*, 2019. arXiv: 1906.10770.

[74] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum και Jiajun Wu. *The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. arXiv:1904.12584 [cs]*, 2019. arXiv: 1904.12584.

[75] Jacob Andreas, Marcus Rohrbach, Trevor Darrell και Dan Klein. *Neural Module Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 39–48, Las Vegas, NV, USA, 2016. IEEE.

[76] Qing Li, Jianlong Fu, Dongfei Yu, Tao Mei και Jiebo Luo. *Tell-and-Answer: Towards Explainable Visual Question Answering using Attributes and Captions.* Τεχνική Αναφορά με αριθμό arXiv:1801.09041, arXiv, 2018. arXiv:1801.09041 [cs] type: article.

[77] Jialin Wu, Zeyuan Hu και Raymond J. Mooney. *Generating Question Relevant Captions to Aid Visual Question Answering.* Τεχνική Αναφορά με αριθμό arXiv:1906.00513, arXiv, 2020. arXiv:1906.00513 [cs] type: article.

[78] Ander Salaberria, Gorka Azkune, Oier Lopezde Lacalle, Aitor Soroa και Eneko Agirre. *Image Captioning for Effective Use of Language Models in Knowledge-Based Visual Question Answering.* Τεχνική Αναφορά με αριθμό arXiv:2109.08029, arXiv, 2022. arXiv:2109.08029 [cs] type: article.

[79] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim και Gunhee Kim. *TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. arXiv:1704.04497 [cs]*, 2017. arXiv: 1704.04497.

[80] Madeleine Grunde-McLaughlin, Ranjay Krishna και Maneesh Agrawala. *AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning. arXiv:2103.16002 [cs]*, 2021. arXiv: 2103.16002.

[81] Jie Lei, Licheng Yu, Mohit Bansal και Tamara L. Berg. *TVQA: Localized, Compositional Video Question Answering. arXiv:1809.01696 [cs]*, 2019. arXiv: 1809.01696.

[82] Jiyang Gao, Runzhou Ge, Kan Chen και Ram Nevatia. *Motion-Appearance Co-memory Networks for Video Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 6576–6585, Salt Lake City, UT, 2018. IEEE.

[83] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He και Chuang Gan. *Beyond RNNs: Positional Self-Attention with Co-Attention for Video Question Answering. Proceedings of the AAAI Conference on Artificial Intelligence*, 33:8658–8665, 2019.

[84] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang και Heng Huang. *Heterogeneous Memory Enhanced Multimodal Attention Model for Video Question Answering. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 1999–2007, Long Beach, CA, USA, 2019. IEEE.

[85] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim και Chang D. Yoo. *Progressive Attention Memory Network for Movie Story Question Answering. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 8329–8338, Long Beach, CA, USA, 2019. IEEE.

[86] Hyounghun Kim, Zineng Tang και Mohit Bansal. *Dense-Caption Matching and Frame-Selection Gating for Temporal Localization in VideoQA. arXiv:2005.06409 [cs]*, 2020. arXiv: 2005.06409.

[87] Thao Minh Le, Vuong Le, Svetha Venkatesh και Truyen Tran. *Hierarchical Conditional Relation Networks for Video Question Answering. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 9969–9978, Seattle, WA, USA, 2020. IEEE.

[88] Thao Minh Le, Vuong Le, Svetha Venkatesh και Truyen Tran. *Neural Reasoning, Fast and Slow, for Video Question Answering. arXiv:1907.04553 [cs]*, 2020. arXiv: 1907.04553.

[89] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy και Cordelia Schmid. *VideoBERT: A Joint Model for Video and Language Representation Learning. arXiv:1904.01766 [cs]*, 2019. arXiv: 1904.01766.

[90] Chen Sun, Fabien Baradel, Kevin Murphy και Cordelia Schmid. *Learning Video Representations using Contrastive Bidirectional Transformer. arXiv:1906.05743 [cs, stat]*, 2019. arXiv: 1906.05743.

[91] Linjie Li, Yen Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu και Jingjing Liu. *HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training. arXiv:2005.00200 [cs]*, 2020. arXiv: 2005.00200.

[92] Aisha Urooj Khan, Amir Mazaheri, Niels da Vitoria Lobo και Mubarak Shah. *MMFT-BERT: Multimodal Fusion Transformer with BERT Encodings for Visual Question Answering. arXiv:2010.14095 [cs]*, 2020. arXiv: 2010.14095.

[93] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima και Haruo Takemura. *BERT Representations for Video Question Answering. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, σελίδες 1545–1554, Snowmass Village, CO, USA, 2020. IEEE.

[94] Noa Garcia και Yuta Nakashima. *Knowledge-Based Video Question Answering with Unsupervised Scene Descriptions. Computer Vision – ECCV 2020*Andrea Vedaldi, Horst Bischof, Thomas Brox και Jan Michael Frahm, επιμελητές, τόμος 12363, σελίδες 581–598. Springer International Publishing, Cham, 2020. Series Title: Lecture Notes in Computer Science.

[95] Deniz Engin, François Schnitzler, Ngoc Q. K. Duong και Yannis Avrithis. *On the hidden treasure of dialog in video question answering. arXiv:2103.14517 [cs]*, 2021. arXiv: 2103.14517.

[96] Reginald B. Adams, Daniel N. Albohn και Kestutis Kveraga. *Social Vision: Applying a Social-Functional Approach to Face and Expression Perception. Current directions in psychological science*, 26(3):243–248, 2017.

[97] Uta Frith και Christopher D Frith. *Development and neurophysiology of mentalizing. Philosophical Transactions of the Royal Society B: Biological Sciences*, 358(1431):459–473, 2003.

[98] Kyung Min Kim, Min Oh Heo, Seong Ho Choi και Byoung Tak Zhang. *DeepStory: Video Story QA by Deep Embedded Memory Networks.* Τεχνική Αναφορά με αριθμό arXiv:1707.00836, arXiv, 2017. arXiv:1707.00836 [cs] type: article.

[99] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun και Sanja Fidler. *MovieQA: Understanding Stories in Movies through Question-Answering.* Τεχνική Αναφορά με αριθμό arXiv:1512.02902, arXiv, 2016. arXiv:1512.02902 [cs] type: article.

[100] Rowan Zellers, Yonatan Bisk, Ali Farhadi και Yejin Choi. *From Recognition to Cognition: Visual Commonsense Reasoning. arXiv:1811.10830 [cs]*, 2019. arXiv: 1811.10830.

[101] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang και Jingjing Liu. *VIOLIN: A Large-Scale Dataset for Video-and-Language Inference. arXiv:2003.11618 [cs]*, 2020. arXiv: 2003.11618.

[102] Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik και Thomas L. Griffiths. *Evaluating Theory of Mind in Question Answering.* Τεχνική Αναφορά με αριθμό arXiv:1808.09352, arXiv, 2018. arXiv:1808.09352 [cs] type: article.

[103] Xianyu Chen, Ming Jiang και Qi Zhao. *Predicting Human Scanpaths in Visual Question Answering. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 10871–10880, Nashville, TN, USA, 2021. IEEE.

[104] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria και Louis Philippe Morency. *Tensor Fusion Network for Multimodal Sentiment Analysis. arXiv:1707.07250 [cs]*, 2017. arXiv: 1707.07250.

[105] Jie Lei, Licheng Yu, Tamara L. Berg και Mohit Bansal. *TVQA+: Spatio-Temporal Grounding for Video Question Answering. arXiv:1904.11574 [cs]*, 2020. arXiv: 1904.11574.

[106] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik και Antonio Torralba. *Gaze360: Physically Unconstrained Gaze Estimation in the Wild.* σελίδες 6912–6921, 2019.

[107] Rıza Alp Güler, Natalia Neverova και Iasonas Kokkinos. *DensePose: Dense Human Pose Estimation In The Wild.* Τεχνική Αναφορά με αριθμό arXiv:1802.00434, arXiv, 2018. arXiv:1802.00434 [cs] type: article.

[108] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang και Yoshua Bengio. *Challenges in Representation Learning: A report on three machine learning contests.* Τεχνική Αναφορά με αριθμό arXiv:1307.0414, arXiv, 2013. arXiv:1307.0414 [cs, stat] version: 1 type: article.

[109] Florian Schroff, Dmitry Kalenichenko και James Philbin. *FaceNet: A unified embedding for face recognition and clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 815–823, Boston, MA, USA, 2015. IEEE.

[110] Sergey Edunov, Myle Ott, Michael Auli και David Grangier. *Understanding Back-Translation at Scale. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, σελίδες 489–500, Brussels, Belgium, 2018. Association for Computational Linguistics.

[111] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li και Peter J. Liu. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* Τεχνική Αναφορά με αριθμό arXiv:1910.10683, arXiv, 2020. arXiv:1910.10683 [cs, stat] version: 3 type: article.

[112] Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan και Xilin Chen. *Multi-Modal Graph Neural Network for Joint Reasoning on Vision and Scene Text. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 12743–12753, Seattle, WA, USA, 2020. IEEE. ZSCC: 0000007.

[113] Xuan Luo, Jia Bin Huang, Richard Szeliski, Kevin Matzen και Johannes Kopf. *Consistent Video Depth Estimation.* Τεχνική Αναφορά με αριθμό arXiv:2004.15021, arXiv, 2020. arXiv:2004.15021 [cs] type: article.

[114] Ali Mollahosseini, Behzad Hasani και Mohammad H. Mahoor. *AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. arXiv:1708.03985 [cs].

[115] Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe και Shrikanth Narayanan. *A Review of Speaker Diarization: Recent Advances with Deep Learning.* Τεχνική Αναφορά με αριθμό arXiv:2101.09624, arXiv, 2021. arXiv:2101.09624 [cs, eess] type: article.

[116] R. Santhoshkumar και M. Kalaiselvi Geetha. *Deep Learning Approach for Emotion Recognition from Human Body Movements with Feedforward Deep Convolution Neural Networks. Procedia Computer Science*, 152:158–165, 2019.

[117] Zhengyuan Yang, Amanda Kay, Yuncheng Li, Wendi Cross και Jiebo Luo. *Pose-based Body Language Recognition for Emotion and Psychiatric Symptom Interpretation.* Τεχνική Αναφορά με αριθμό arXiv:2011.00043, arXiv, 2020. arXiv:2011.00043 [cs] type: article.

[118] Leonardo Pepino, Pablo Riera και Luciana Ferrer. *Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings. arXiv:2104.03502 [cs, eess]*, 2021. arXiv: 2104.03502.

[119] Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi και Hannaneh Hajishirzi. *Reframing Instructional Prompts to GPTk's Language.* Τεχνική Αναφορά με αριθμό arXiv:2109.07830, arXiv, 2022. arXiv:2109.07830 [cs] type: article.

[120] Constantinos Karouzos, Georgios Paraskevopoulos και Alexandros Potamianos. *UDALM: Unsupervised Domain Adaptation through Language Modeling.* Τεχνική Αναφορά με αριθμό arXiv:2104.07078, arXiv, 2021. arXiv:2104.07078 [cs] type: article.

[121] Lifeng Fan, Wenguan Wang, Song Chun Zhu, Xinyu Tang και Siyuan Huang. *Understanding Human Gaze Communication by Spatio-Temporal Graph Reasoning. 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, σελίδες 5723–5732, Seoul, Korea (South), 2019. IEEE.

[122] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman και Aram Galstyan. *A Survey on Bias and Fairness in Machine Learning.* Τεχνική Αναφορά με αριθμό arXiv:1908.09635, arXiv, 2022. arXiv:1908.09635 [cs] type: article.

List of Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
GPU	Graphics Processing Unit
CV	Computer Vision
NLP	Natural Language Processing
NMT	Neural Machine Translation
LM	Language Modelling
MLM	Masked Language Modelling
QA	Question Answering
VQA	Visual Question Answering
RACE	ReAding Comprehension dataset from Examinations
EM	Expectation Maximization
PCA	Principal Components Analysis
SVM	Support Vector Machine
MLP	Multi Layer Perceptron
FFNN	Feed Forward Neural Network
MSE	Mean Squared Error
SGD	Stochastic Gradient Descent
ReLU	Rectified Linear Unit
IoU	Intersection over Union
CNN	Convolutional Neural Network
R-CNN	Region-based Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory network
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly optimized BERT approach
ViLBERT	Vision and Language BERT
GPT	Generative Pre-trained Transformer
GNN	Graph Neural Network
TMFN	Tensor Memory Fusion Network
MAC	Memory Attention Composition
MAC-X	MAC-Extend
ASD	Autism Spectrum Disorder
WS	World Scope
ToM	Theory of Mind
IoT	Internet of Things