Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

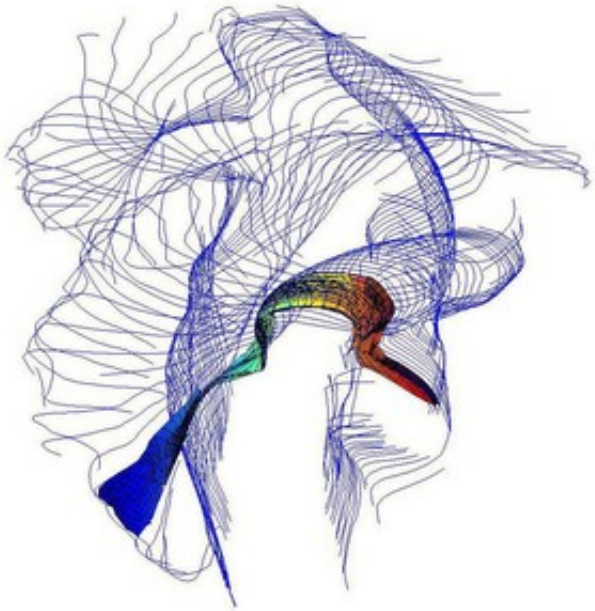
Χωροχρονική Αναγνώριση Δράσης

Μελέτη και υλοποίηση

Διπλωματικη Εργασια

του

ΒΑΓΕΝΑ ΕΥΣΤ. ΝΙΚΗΦΟΡΟΥ



Επιβλέπων: Αθανάσιος Βουλόδημος
Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Χωροχρονική Αναγνώριση Δράσης

Μελέτη και υλοποίηση

Διπλωματικη Εργασια

του

ΒΑΓΕΝΑ ΕΥΣΤ. ΝΙΚΗΦΟΡΟΥ

Επιβλέπων: Αθανάσιος Βουλόδημος
Επίκουρος Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 11η Ιουλίου 2024.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Αθανάσιος Βουλόδημος	Γεώργιος Στάμου	Ανδρέας Γεώργιος Σταφυλοπάτης
Επίκουρος Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Διπλωματικής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στη Διπλωματική μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Διπλωματική Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Νικηφόρος Βαγενάς

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.
7 Ιουλίου 2024

Περίληψη

Η αναγνώριση ανθρώπινης δράσης σε βίντεο αποτελεί έναν από τα κυριότερα προβλήματα του τομέα της όρασης υπολογιστών και έχει συγκεντρώσει το ενδιαφέρον πολλών ερευνητών λόγω της δυνατότητας εφαρμογής της σε διάφορους τομείς της ανθρώπινης δραστηριότητας, από την ιατρική έως τον κόσμο του κινηματογράφου και της μόδας. Διαχρονικά έχουν χρησιμοποιηθεί πολλά μοντέλα βαθιάς μάθησης με σκοπό την ορθή πρόβλεψη της κλάσης, στην οποία ανήκει η δράση που αναπαρίσταται σε ένα βίντεο. Η βασική αρχιτεκτονική, που εφαρμόζεται για την ανάλυση εικόνων και βίντεο είναι τα συνελικτικά νευρωνικά δίκτυα, τα οποία διαθέτουν την ικανότητα εξαγωγής χαρακτηριστικών, όπως ακμές και υφές, από τα οπτικό υλικό, μέσω της εφαρμογής ειδικών φίλτρων στα δεδομένα. Πρόσφατα μεγάλο ενδιαφέρον παρουσιάζουν οι μετασχηματιστές, μοντέλο, που δανείστηκαν οι ερευνητές της όρασης υπολογιστών από τον τομέα επεξεργασίας φυσικής γλώσσας. Οι μετασχηματιστές απαιτούν περισσότερα δεδομένα, γεγονός που αυξάνει το υπολογιστικό κόστος, αλλά μπορούν να μάθουν τους συσχετισμούς μεταξύ των διαφορετικών τμημάτων μίας εικόνας ή βίντεο. Τέλος, οι αποκρύπτοντες αυτοκωδικοποιητές είναι αρχιτεκτονικές νευρωνικών δικτύων για την αναγνώριση δράσης σε βίντεο των τελευταίων δύο χρόνων, που χρησιμοποιούν τεχνικές απόκρυψης μέρους των δεδομένων και τους οπτικούς μετασχηματιστές, για να κάνουν ταξινόμηση στα δεδομένα εισόδου.

Στόχος αυτής της διπλωματικής εργασίας είναι η αναπαραγωγή σε προγραμματιστικό περιβάλλον της εκπαίδευσης τριών από τα μοντέλα αποκρύπτοντων αυτοκωδικοποιητών, ώστε να παρουσιαστούν οι καλές τους επιδόσεις στο πρόβλημα αναγνώρισης ανθρώπινης δράσης σε βίντεο. Επιπλέον, εκτελούμε ένα πείραμα, όπου προεπεξεργαζόμαστε τα δεδομένα εισόδου με τέτοιο τρόπο, ώστε να αποκρύψουμε και στη διάσταση του χρόνου τμήματα αυτών. Με αυτόν τον τρόπο, θέλουμε να αναδείξουμε πως μπορεί στις συγκεκριμένες αρχιτεκτονικές δικτύου να μειωθεί ο χρόνος εκπαίδευσης με ελάχιστες απώλειες στην ορθότητα πρόβλεψης.

Λέξεις Κλειδιά

Αναγνώριση ανθρώπινης δράσης σε βίντεο, χωροχρονική ανάλυση δεδομένων, μετασχηματιστής, αποκρύπτων αυτοκωδικοποιητής, μάσκα απόκρυψης

Abstract

Human action recognition in video is one of the main problems in the field of computer vision and has attracted the interest of many researchers duo to its implementations in many fields of human activity, ranging from medicine to the world of cinema an fashion. Throughout the years many deep learning models have been used, in order to accurately predict the class of an action being portrayed in a video. The main architecture, that is used for image and video analysis, are the convolution neural networks, which are capable of extracting visual features, such as edges and texture, by applying special filters to the data. Recently, there has been shown a great interest towards the transformers, a model borrowed by the computer vision researchers from the field of natural language processing. The transformers need a lot of data, which increases its computational cost, but they are capable of learning how different parts of an image or a video are related to one another. Finally, masked autoencoders are architectures of neural nets for action recognition, that have appeared the last two years and use vision transformers and data masking techniques, so that they can classify the input data.

This diploma thesis aims to reproduce the training of three of these masked autoencoder models in a programming environment, so that their good performance in human action recognition tasks can be demonstrated. Moreover, we conduct an experiment, where we preprocess the input data in such a way, so that we can mask part of the data in the temporal dimension as well. In this way, we want to highlight how the training time can be reduced in certain architectures of neural nets without heavy losses in prediction accuracy.

Keywords

Human action recognition, spatiotemporal data analysis, transformer, masked autoencoder, mask

στους γονείς μου

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Βουλόδημο για την ευκαιρία που μου έδωσε να εκπονήσω την εργασία μου στο εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης καθώς και την τριμελή επιτροπή για τη δυνατότητα να παρουσιάσω τη διπλωματική μου. Επιπλέον,θα ήθελα να ευχαριστήσω την καθηγήτρια κ. Τζούβελη για τη συμβολή της στην επίβλεψη αυτής της διπλωματικής εργασίας. Επίσης, ευχαριστώ ιδιαίτερα την υποψήφια διδάκτορα Θεοφίλου Παρασκευή για την καθοδήγησή της και την εξαιρετική συνεργασία που είχαμε. Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου για την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

Αθήνα, Μάιος 2024

Νικηφόρος Βαγενάς

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Πρόλογος

Η διπλωματική εργασία διενεργήθηκε στο εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης , το οποίο βρίσκεται στα παλιά κτήρια της σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου στην πολυτεχνειούπολη της Ζωγράφου.

Κεφάλαιο 1

Εισαγωγή

Η σημερινή εποχή χαρακτηρίζεται από τη διείσδυση της τεχνητής νοημοσύνης σε ποικίλα κομμάτια της ανθρώπινης ζωής και δραστηριότητας, με σκοπό την υποστήριξη του ανθρώπου. Καθημερινά, ακόμα και χωρίς να το καταλάβουμε, κάνουμε χρήση εργαλείων τεχνητής νοημοσύνης, όπως το σύστημα προτάσεων ταινιών ή σειρών στο Netflix ή τον τρόπο που το κινητό μας από μόνο του είναι σε θέση να αναγνωρίζει επαναλαμβανόμενα πρόσωπα σε εικόνες και να οργανώνει τη συλλογή φωτογραφιών μας βάσει αυτού.

Ειδικότερα το τελευταίο παράδειγμα αποτελεί ένα εργαλείο που βασίζεται σε τεχνικές ενός συγκεκριμένου τομέα της τεχνητής μάθησης, της όρασης υπολογιστών. Αυτός ο τομέας ασχολείται με την επεξεργασία και ανάλυση εικόνων και βίντεο, από τα οποία προσπαθεί με πολλούς τρόπους να εξάγει διάφορα χαρακτηριστικά και να εντοπίσει μοτίβα, π.χ. την συνεχόμενη εμφάνιση των ίδιων προσώπων σε φωτογραφίες, που αναφέρθηκε παραπάνω.

Ένα από τα επιμέρους προβλήματα που καλείται να λύσει ο συγκεκριμένος τομέας είναι η κατηγοριοποίηση της ανθρώπινης δράσης, που περιέχει ένα βίντεο, σε μία από πολλές γνωστές κλάσεις ανθρώπινων δράσεων. Όπως μπορεί να φανταστεί κανείς, αν γίνει με επιτυχία η ταξινόμηση του περιεχομένου ενός βίντεο με τη σωστή κλάση δράσης, στην οποία ανήκει, αυτό μπορεί να εφαρμοστεί σε διάφορους τομείς. Μπορεί να χρησιμοποιηθεί σε λιγότερο σημαντικές για την ανθρώπινη ζωή εφαρμογές, όπως για παράδειγμα, για να παράγεται αυτόματα μία μικρή περίληψη του περιεχομένου ενός βίντεο στο YouTube, αλλά και σε κρίσιμες για τη ζωή μας εφαρμογές, π.χ. την παρακολούθηση ασθενών που πάσχουν από Parkinson, για να μπορέσουν οι φροντιστές τους να ειδοποιηθούν εγκαίρως σε περίπτωση που κάνουν κάποια απότομη κίνηση και χρειαστούν βοήθεια.

1.1 Αντικείμενο της διπλωματικής

Αυτή η διπλωματική έχει ως αντικείμενο την εύρεση μοντέλων τελευταίας τεχνολογίας, τα οποία με αποδοτικό τρόπο και μεγάλη ακρίβεια καταφέρνουν να προβλέψουν τις κλάσεις, στην οποία ανήκουν κάποια βίντεο. Αφού παρουσιαστούν διάφορα μοντέλα, που έχουν χρησιμοποιηθεί ανά τα χρόνια στο πρόβλημα της αναγνώρισης ανθρώπινης δράσης σε βίντεο, θα δώσουμε έμφαση στην παρουσίαση αυτών που πετυχαίνουν τον επιθυμητό στόχο της ορθής ταξινόμησης των βίντεο σε κατηγορίες χωρίς να απαιτούν υπερβολικούς υπολογιστικούς πόρους.

Ο συγκεκριμένος παράγοντας κρίνεται πολύ σημαντικός στην εργασία μας, καθώς αφε-

νός έχουμε πρόσβαση σε περιορισμένη υπολογιστική ισχύ για τη διενέργεια των πειραμάτων μας. Αφετέρου, αποτελεί μεγάλο πρόβλημα στο σύγχρονο ερευνητικό κόσμο η εύρεση μοντέλων, τα οποία μπορούν με αποδοτικό τρόπο να φέρουν εις πέρας την αποστολή της ορθής ταξινόμησης των δεδομένων σε κλάσεις, εφόσον τα διαθέσιμα δεδομένα εισόδου είναι πάρα πολλά σε αριθμό.

Τέλος, εκτός από την αναπαραγωγή και ανάδειξη της καλής επίδοσης των τελικά επιλεγμένων μοντέλων, προκειμένου να παράξουμε κάποιο πρωτότυπο ερευνητικό έργο εκτελούμε μερικά πειράματα. Έτσι είμαστε σε θέση να παρουσιάσουμε τα συμπεράσματά μας βάσει αυτών. Τα πειράματα βασίζονται και αυτά στην απλοποίηση των δεδομένων εισόδου, ώστε να μην επιβαρύνεται τόσο πολύ υπολογιστικά το μοντέλο, με τέτοιο τρόπο όμως, που να μην χειροτερεύει σημαντικά η ικανότητα πρόβλεψης των μοντέλων.

1.2 Οργάνωση του τόμου

Η εργασία αυτή είναι οργανωμένη σε έξι κεφάλαια: Στο Κεφάλαιο 2 δίνεται το θεωρητικό υπόβαθρο των βασικών μεθόδων τεχνητής νοημοσύνης και μηχανικής μάθησης και ιδιαίτερα αυτών που βρίσκουν εφαρμογή στην όραση υπολογιστών. Στο Κεφάλαιο 3 αρχικά περιγράφεται το πρόβλημα, το οποίο αποτελεί και το θέμα της παρούσας διπλωματικής. Στη συνέχεια γίνεται μία ιστορική αναδρομή όσον αφορά τα μοντέλα βαθιάς μάθησης, που έχουν χρησιμοποιηθεί για την επίλυση του προβλήματος. Στο Κεφάλαιο 4 παρουσιάζεται η μεθοδολογία και η υλοποίηση των πειραμάτων. Γίνεται αναλυτική περιγραφή των βημάτων, που ακολουθούνται, για να στηθούν τα πειράματα, σε συνδυασμό με αναφορά στις πλατφόρμες και τα προγραμματιστικά εργαλεία που χρησιμοποιούνται. Στο Κεφάλαιο 5 παρουσιάζεται με λεπτομέρεια ο τρόπος, με τον οποίο εκπονούνται τα πειράματα για κάθε μοντέλο. Επίσης αναλύονται τα αποτελέσματα που παίρνουμε από τα πειράματα αυτά. Τέλος στο Κεφάλαιο 6 δίνεται η συνεισφορά αυτής της διπλωματικής εργασίας, καθώς και μελλοντικές επεκτάσεις.

Μέρος Ι

Θεωρητικό Μέρος

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Στο κεφάλαιο αυτό παρέχουμε το απαραίτητο θεωρητικό υπόβαθρο για τον τομέα της τεχνητής νοημοσύνης. Πιο συγκεκριμένα, θα δώσουμε μία γενική εικόνα του τομέα στην πρώτη ενότητα του κεφαλαίου και στις ενότητες που ακολουθούν θα παρουσιάσουμε τον τομέα της όρασης υπολογιστών, δηλαδή τον τομέα στον οποίο ανήκει το θέμα αυτής της διπλωματικής, και την εξέλιξη των τεχνικών που χρησιμοποιούνται.

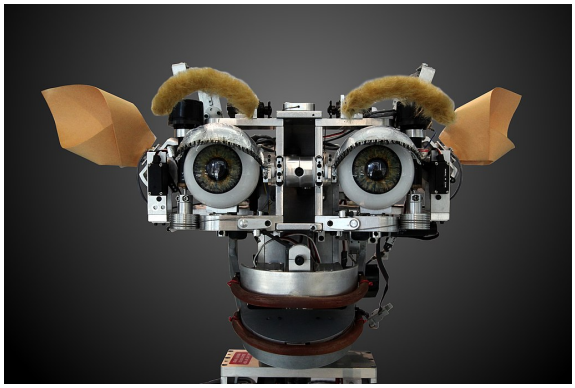
2.1 Τεχνητή Νοημοσύνη

Η τεχνητή νοημοσύνη είναι ένας τομέας έρευνας στην επιστήμη υπολογιστών, ο οποίος στοχεύει στην κατασκευή έξυπνων μηχανών. Το γενικότερο πρόβλημα της δημιουργίας μηχανών, των οποίων η νοημοσύνη προσεγγίζει την ανθρώπινη μπορεί να χωριστεί σε υποπροβλήματα. Μερικά από αυτά είναι η ικανότητα επίλυσης προβλημάτων, η αναπαράσταση γνώσης, ο σχεδιασμός και λήψη αποφάσεων, η εκπαίδευση, η επεξεργασία φυσικής γλώσσας, η κοινωνική νοημοσύνη και η γενική νοημοσύνη[5].

Όπως προαναφέρθηκε, ο στόχος του AI(Artificial Intelligence) είναι η δημιουργία μηχανών που έχουν την ικανότητα να σκεφτούν σαν άνθρωποι και συνεπώς να φέρουν εις πέρας μία πολύπλοκη ανθρώπινη δραστηριότητα. Οι μέθοδοι που χρησιμοποιούνται, για να επιτευχθεί αυτός ο στόχος εξαρτώνται από το πρόβλημα, αλλά μερικές από τις πιο κοινές είναι οι γενετικοί αλγόριθμοι[6], οι αλγόριθμοι αναζήτησης, τα νευρωνικά δίκτυα, οι τεχνικές μηχανικής και βαθιάς μάθηση. Αφού τεθούν σε εφαρμογή αυτές οι τεχνικές σε μηχανές, οι οποίες έχουν το κατάλληλο hardware, για να εκτελέσουν τις απαραίτητες κινήσεις, όπως ρομποτικά χέρια, μπορούμε να εκπαιδεύσουμε τις μηχανές, ώστε να εκτελούν κάποιες εντολές χωρίς οδηγίες από εμάς.

Η χρήση των εργαλείων τεχνητής νοημοσύνης είναι διαδεδομένη σε πολλούς τομείς της ανθρώπινης δραστηριότητας στις μέρες μας, από τον επιχειρηματικό κόσμο και τον αθλητισμό μέχρι την παιδεία και τις κυβερνήσεις τεχνολογικά αναπτυγνένων χωρών. Υπάρχουν πολλά παραδείγματα τέτοιων εργαλείων, που χρησιμοποιούνται καθημερινά σχεδόν από όλους, όπως το ChatGPT[7], το σύστημα, που μας προτείνει περιεχόμενο σε Netflix και YouTube. Εκτός από τις τεχνολογίες AI, που είναι προσβάσιμες σε όλους, υπάρχουν και εργαλεία σχεδιασμένα για χρήση μόνο από ειδικούς. Παραδείγματος χάριν, οι μετεωρολόγοι πλέον χρησιμοποιούν την τεχνητή νοημοσύνη, προκειμένου να προβλέψουν όχι μόνο τον καιρό, αλλά και το πως μπορεί η κλιματική αλλαγή να επηρεάσει την επιβίωση του είδους

μας στον πλανήτη. [8]



Εικόνα 2.1: *Kismet, ένα κεφάλι ρομπότ που κατασκευάστηκε στη δεκαετία του 1990: ένα μηχάνημα που μπορεί να καταλαβαίνει και να δείχνει συναισθήματα[1].*

2.2 Μηχανική Μάθηση

Η μηχανική μάθηση είναι ένας κλάδος της τεχνητής νοημοσύνης. Παρόλο που τόσο το AI και το ML(Machine Learning) επιδιώκουν την εύρεση λύσεων για δύσκολα προβλήματα, οι οποίες πρέπει να είναι όσο το δυνατόν πιο ακριβείς γίνεται με βάση τη γνώση, την οποία έχει λάβει η μηχανή, η μηχανική μάθηση σχετίζεται περισσότερο με την ανάλυση μεγάλου όγκου δεδομένων. Με τη βοήθεια εργαλείων της στατιστικής, τα οποία της δίνουν τη δυνατότητα να αναγνωρίσει πρότυπα στα δεδομένα και να εξάγει ένα συμπέρασμα με υψηλό βαθμό εμπιστοσύνης.

Υπάρχουν τρεις κύριες κατηγορίες μηχανικής μάθησης[9]:

- Επιβλεπόμενη μάθηση
- Μη επιβλεπόμενη μάθηση
- Ενισχυτική μάθηση

Η σημαντικότερη διαφορά ανάμεσα στις δύο πρώτες κατηγορίες σχετίζεται με την επισήμανση του συνόλου δεδομένων, στο οποίο εκπαιδεύεται και δοκιμάζεται το μοντέλο της μηχανικής μάθησης. Ειδικότερα, τα μοντέλα επιβλεπόμενης μάθησης εκπαιδεύονται γνωρίζοντας την ετικέτα της κατηγορίας, στην οποία ανήκει κάθε δεδομένο, και υπολογίζει με αυτον τον τρόπο το σφάλμα ανάμεσα στην ετικέτα που προβλέπουν εκείνα με τη σωστή. Αντιθέτως, στη μη επιβλεπόμενη μάθηση η μηχανή δεν χρειάζεται να ξέρει τις ετικέτες των δεδομένων, αφού καταφεύγει σε άλλες μεθόδους, π.χ. συσταδοποίηση, προκειμένου να βγάλει ένα συμπέρασμα για την κατηγορία στην οποία ανήκουν τα δεδομένα.

Όσον αφορά την ενισχυτική μάθηση, είναι η κατηγορία μηχανικής μάθησης, η οποία επικεντρώνεται στην εύρεση της βέλτιστης απόφασης σε ένα περιβάλλον, όπου κάθε ενέργεια επιστρέφει μία τιμή από ορισμένες συναρτήσεις επιβράβευσης. Σκοπός της μηχανής είναι η επίτευξη της μέγιστης τιμής σε αυτές τις συναρτήσεις. Διαφοροποιείται απο τις άλλες δύο κατηγορίες, γιατί η μηχανή που εκπαιδεύεται καλείται πράκτορας και αλληλεπιδρά με το

περιβάλλον της λαμβάνοντας αποφάσεις. Με αυτόν τον τρόπο δεν επιδιώκει την αναγνώριση προτύπων, αλλά την εύρεση της καλύτερης κίνησης κάθε φορά.

Προκειμένου να πετύχουν τη μέγιστη δυνατή ορθότητα πρόβλεψης των ετικετών ενός συνόλου δεδομένων στην επιβλεπόμενη και στη μη επιβλεπόμενη μάθηση και τη μεγιστοποίηση της συνάρτησης επιβράβευσης στην ενισχυτική μάθηση, τα μοντέλα μηχανικής μάθησης χρειάζονται μεγάλο όγκο δεδομένων. Για αυτό, η επιλογή του σωστού συνόλου δεδομένων και της κατάλληλης προεπεξεργασίας του είναι σημαντικό κομμάτι του ML. Αφενός, όσο μεγαλύτερος ο όγκος δεδομένων που επεξεργάζεται το μοντέλο, τόσο ακριβέστερο το αποτέλεσμα του μοντέλου, αφετέρου, μαζί με τα δεδομένα αυξάνονται ο χρόνος και η υπολογιστική δύναμη, που απαιτούνται, ώστε να γίνει η επεξεργασία τους. Αυτό είναι ένα σημαντικό πρόβλημα που αντιμετωπίζουν οι σύγχρονοι ερευνητές στον τομέα του ML και γίνεται προσπάθεια να βρεθούν πιο αποτελεσματικές μέθοδοι, για να μειωθούν οι υπολογιστικές ανάγκες της εκπαίδευσης των μοντέλων χωρίς να υπάρχουν σημαντικές απώλειες στην ακρίβεια τω αποτελεσμάτων.

2.2.1 Όραση Υπολογιστών

Ένας από τους κυριότερους τομείς της μηχανικής μάθησης, ο οποίος θα μας απασχολήσει στη συγκεκριμένη διπλωματική εργασία, είναι η όραση υπολογιστών. Πρόκειται για το σύνολο των μεθόδων, που αναλύουν και επεξεργάζονται οπτικά δεδομένα με σκοπό την εξαγωγή χρήσιμων συμπερασμάτων. Προσπαθούν αρχικά να μεταφράσουν τα pixels μίας εικόνας σε αριθμητικά δεδομένα, με τα οποία μπορεί να κάνει πράξεις ο υπολογιστής, και στη συνέχεια τα εισάγουν σε διάφορα μοντέλα, προκειμένου να μάθουν από αυτά.

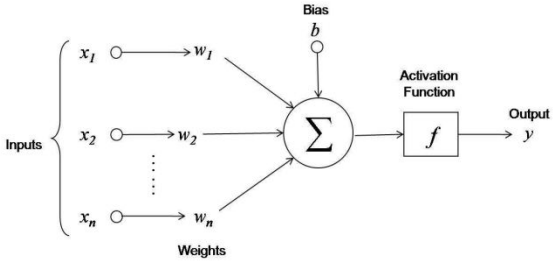
Υπάρχουν πολλές μέθοδοι για την επεξεργασία και ερμηνεία δεδομένων από τον οπτικό κόσμο, οι οποίες προσομοιάζουν τον τρόπο που ο άνθρωπος αναλύει οπτικά δεδομένα. Παρακάτω παραθέτουμε μία ιστορική αναδρομή στην εξέλιξη αυτών των μεθόδων, ξεκινώντας από τα νευρωνικά δίκτυα.

Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα είναι ένα υπολογιστικό σύστημα, το οποίο μιμείται τις σύνθετες διεργασίες, που λαμβάνουν χώρα στον ανθρώπινο εγκέφαλο. Είναι το πιο βασικό μοντέλο που χρησιμοποιείται στη μηχανική μάθηση. Αποτελούνται από συνδεδεμένους κόμβους ή νευρώνες, που αναλύουν και μαθαίνουν από τα δεδομένα, διευκολύνοντας την εύρεση λύσης σε προβλήματα όπως αναγνώριση προτύπων και λήψη αποφάσεων.

Κάθε νευρωνικό δίκτυο είναι κατασκευασμένο από επίπεδα νευρώνων. Σε αυτά περιλαμβάνονται ένα επίπεδο εισόδου, τουλάχιστον ένα κρυφό επίπεδο και ένα επίπεδο εξόδου. Κάθε νευρώνας είναι συνδεδεμένος με τους άλλους και έχει δύο χαρακτηριστικά, μία τιμή βάρους, που αντιστοιχεί σε κάθε είσοδο του νευρώνα, και μία κλίση(bias). Συνήθως, όπως φαίνεται και στην εικόνα, εφαρμόζεται μία συνάρτηση ενεργοποίησης στην έξοδο του νευρώνα, για να αποφανθεί το δίκτυο αν θα περάσει τους υπολογισμούς ενός νευρώνα στο επόμενο επίπεδο του δικτύου[10].

Ο τύπος για τον υπολογισμό της εξόδου κάθε κόμβου υπολογίζεται από τα εξής βήματα:

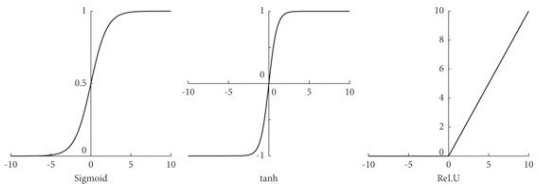


Εικόνα 2.2: *Ο νευρώνας, ο ακρογωνιαίος λίθος κάθε νευρωνικού δικτύου.*

1. Ζυγισμένο Άθροισμα: Η τιμή κάθε εισόδου πολλαπλασιάζεται με το αντίστοιχο βάρος και τα επιμέρους γινόμενα αθροίζονται. Για παράδειγμα, για n εισόδους, έχουμε αντίστοιχα n βάρη και το ζυγισμένο άθροισμα: $z = w_1 * x_1 + w_2 * x_2 + ... + w_n * x_n$

2. Προστιθέμενος συντελεστής: Η τιμή της συντελεστή συστημικού σφάλματος, που είναι χαρακτηριστική κάθε νευρώνα, προστίθεται στο ζυγισμένο άθροισμα. Αυτή η τιμή καθορίζει αν η τιμή της εξόδου του νευρώνα θα περάσει το κατώφλι της συνάρτησης ενεργοποίησης, που ακολουθεί. Το συνολικό άθροισμα γίνεται: $z = w_1 * x_1 + w_2 * x_2 + ... + w_n * x_n + b$

3. Συνάρτηση ενεργοποίησης: Το τελικό αποτέλεσμα περνάει από μία συνάρτηση ενεργοποίησης f, η οποία έχει συγκεκριμένο κατώφλι, δηλαδή απαιτεί η είσοδος να είναι σε συγκεκριμένο εύρος τιμών, ώστε να παράξει κάποια έξοδο. Η έξοδος αυτή μπορεί να είναι μη γραμμική, κάτι το οποίο χρειάζεται για την αναγνώριση σύνθετων προτύπων στα δεδομένα. Μερικές από τις πιο κοινές συναρτήσεις ενεργοποίησης είναι η σιγμοειδής[11], η συνάρτηση υπερβολικής εφαπτομένης[12] και η συνάρτηση ανόρθωσης (ReLU)[13]. Η έξοδος του νευρώνα εν τέλει διαμορφώνεται ως εξής: $output = f(z)$



Εικόνα 2.3: *Σύγκριση του εύρους τιμών της σιγμοειδούς, της υπερβολικής εφαπτομένης και της συνάρτησης ReLU. Φαίνεται ότι το κατώφλι της ReLU συγκεκριμένα είναι το 0.*

Η παραγόμενη έξοδος γίνεται η είσοδος στον επόμενου νευρώνα. Αυτή η διαδικασία μεταφοράς δεδομένων από το ένα επίπεδο στο επόμενο του δικτύου το χαρακτηρίζει ως νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης[14]. Γενικότερα, μερικές φορές τα νευρωνικά δίκτυα καλούνται και τεχνητά νευρωνικά δίκτυα (ANNs ή SNNs στα αγγλικά).

Τα νευρωνικά δίκτυα βασίζονται, όπως όλα τα μοντέλα μηχανικής μάθησης, στα δεδομένα εκπαίδευσης, για να βελτιώσουν την ορθότητα της πρόβλεψής τους, όσο συνεχίζουν να τα επεξεργάζονται. Συγκεκριμένα, υπολογίζεται κάθε φορά το σφάλμα ανάμεσα στο αποτέλεσμα του επιπέδου εξόδου του δικτύου και στο σωστό αποτέλεσμα και με βάση αυτό διορθώνονται

ανάλογα οι τιμές των βαρών και των συντελεστών συστημικού σφάλματος σε κάθε επίπεδο. Αυτή η διόρθωση καλείται διαβάθμιση κλίσης και η διαδικασία με την οποία το δίκτυο βελτιώνει μόνο του τις τιμές των παραμέτρων του, ξεκινώντας από την έξοδο και καταλήγοντας στην είσοδο ονομάζεται οπισθοδιάδοση[15].

Συνελικτικά Νευρωνικά Δίκτυα

Τα συνελικτικά νευρωνικά δίκτυα, CNNs ο αγγλικός όρος, είναι μία κατηγορία νευρωνικών δικτύων βαθιάς μάθησης, η οποία εξειδικεύεται στην ανάλυση εικόνας και βίντεο, δηλαδή στον τομέα της όρασης υπολογιστών. Πρόκειται για κανονικοποιημένα νευρωνικά δίκτυα εμπρόσθιας τροφοδότησης, που μαθαίνουν μόνα τους να εξάγουν τα οπτικά χαρακτηριστικά των δεδομένων εφαρμόζοντας φίλτρα σε αυτά[16].

Τα CNNs χρησιμοποιούν τρισδιάστατα (ή τετραδιάστατα δεδομένα, αν λάβουμε υπόψιν τη διάσταση του χρόνου στα βίντεο) για ταξινόμηση εικόνων σε κατηγορίες και αναγνώριση αντικειμένων. Αποτελούνται όπως όλα τα νευρωνικά δίκτυα από ένα επίπεδο εισόδου, κάποια κρυφά επίπεδα, στα οποία η πληροφορία μεταδίδεται από το ένα στο άλλο όπως περιγράφηκε στην παραπάνω υποενότητα, και ένα επίπεδο εξόδου.

Υπάρχουν τρία είδη κρυφών επιπέδων στα CNNs:

1. Συνελικτικό επίπεδο: Στο επίπεδο αυτό οφείλει το όνομά του το δίκτυο και είναι η ειδοποιός διαφορά των CNNs με τους άλλους τύπους νευρωνικών δικτύων. Είναι το κύριο θεμέλιο του δικτύου και χρειάζεται εκτός από τα δεδομένα εισόδου και ένα φίλτρο. Ουσιαστικά το φίλτρο εφαρμόζεται στα οπτικά δεδομένα, που είναι χωρισμένα σε εικονοστοιχεία, πολλαπλασιάζοντας τα βάρη του φίλτρου με ένα συγκεκριμένο κομμάτι της εισόδου κάθε φορά. Αυτή η διαδικασία καλείται συνέλιξη και η επανάληψή της έχει ως αποτέλεσμα τη δημιουργία ενός χάρτη χαρακτηριστικών της εικόνας.

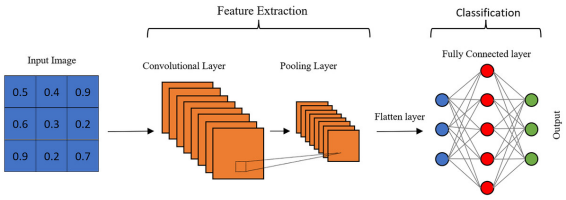
2. Επίπεδο συγκέντρωσης: Αυτό το επίπεδο εφαρμόζει ξανά ένα φίλτρο μίας συγκεκριμένης συνάρτησης στο χάρτη χαρακτηριστικών που έχει προκύψει νωρίτερα. Σκοπός αυτού του επιπέδου είναι η μείωση του μεγέθους των δεδομένων και η εξαγωγή χαρακτηριστικών. Η μείωση των επεξεργαζόμενων δεδομένων εξυπηρετεί τη μείωση του αριθμού των παραμέτρων, που πρέπει να μάθει το μοντέλο, και η εξαγωγή χαρακτηριστικών τη βελτίωση της ορθότητας της πρόβλεψης.

   Οι πιο ευρέως χρησιμοποιούμενες συναρτήσεις των φίλτρων συγκέντρωσης είναι η συγκέντρωση τοπικού μεγίστου, όπου επιλέγεται και μεταφέρεται στο επόμενο επίπεδο το στοιχείο με τη μεγαλύτερη τιμή σε μία περιοχή, και η συγκέντρωση τοπικού μέσου όρου, όπου υπολογίζεται ο μέσος όρος σε μία περιοχή και αυτός αντικαθιστά όλες τις υπόλοιπες στο επόμενο επίπεδο.

3. Πλήρως συνδεδεμένο επίπεδο: Ουσιαστικά αποτελεί το τελευταίο επίπεδο πριν την έξοδο του δικτύου, επομένως χρησιμοποιεί ό,τι χαρακτηριστικά έχουν καταφέρει να εξάγουν τα συνελικτικά επίπεδα, για να ταξινομήσει επιτυχώς την είσοδο στη σωστή κλάση. Ο συνδυασμός των αποτελεσμάτων γίνεται γραμμικά, περνώντας τα από ένα ή παραπάνω επίπεδα νευρώνων, όπου κάθε νευρώνας είναι συνδεδεμένος με κάθε νευρώνα του επόμενου επιπέδου. Με τον πολλαπλασιασμό των βαρών κάθε νευρώνα και

την προσθήκη του συντελεστή συστημικού σφάλματος επιτυγχάνεται ένας μη γραμμικός συνδυασμός των χαρακτηριστικών. Μπορούμε να επιτύχουμε και μη γραμμικό συνδυασμό των χαρακτηριστικών, εισάγοντας το αποτέλεσμα του δικτύου νευρώνων σε μια μη γραμμική συνάρτηση ενεργοποίησης.

Μερικές από τις πολλές εφαρμογές των CNNs είναι η αναγνώριση εικόνας και βίντεο, ταξινόμηση εικόνων και βίντεο, διαχωρισμός εικόνων, ανάλυση και εξαγωγή συμπερασμάτων από ιατρικές εικόνες[17], επεξεργασία φυσικής γλώσσας[18] και επεξεργασία οικονομικών δεδομένων στην πάροδο του χρόνου.



Εικόνα 2.4: *Ένα συνελικτικό νευρωνικό δίκτυο με ένα συνελικτικό επίπεδο και ένα επίπεδο συγκέντρωσης.*

Μετασχηματιστές

Οι μετασχηματιστές(transformers) είναι ένα μοντέλο βαθιάς μάθησης, το οποίο χρησιμοποιήθηκε για πρώτη φορά από ερευνητές της Google στην επεξεργασία φυσικής γλώσσας(NLP) το 2017 με το άρθρο "Attention is all you need"[19]. Το χαρακτηριστικό των transformers, που τους διαφοροποιεί από τα υπόλοιπα μοντέλα βαθιάς μάθησης είναι η χρήση ενός μηχανισμού προσοχής(attention mechanism). Αυτός ο μηχανισμός, αφού τα σειριακά δεδομένα εισόδου έχουν χωριστεί σε μονάδες επεξεργασίας, εξάγει από αυτά το περιεχόμενο τους καθώς και τη συχέτιση που έχουν μεταξύ τους. Ένα άλλο πλεονέκτημα αυτού του μηχανισμού είναι το γεγονός ότι σε αντίθεση με άλλα μοντέλα μηχανικής μάθησης, όπως τα LSTMs, έχει την ικανότητα να επεξεργάζεται όλες τις μονάδες επεξεργασίας ταυτόχρονα και όχι σειριακά.

Η κλασική εκδοχή ενός transformer αποτελείται από έναν κωδικοποιητή(encoder) και έναν αποκωδικοποιητή(decoder). Δουλειά του κωδικοποιητή είναι η εξαγωγή χαρακτηριστικών από τα δεδομένα εισόδου, μεταξύ των οποίων υπολογίζει τη συσχέτιση μέσω του attention. Έτσι παράγει την ενσωμάτωση(embedding) κάθε δεδομένου εισόδου, η οποία ουσιαστικά είναι ένα διάνυσμα τιμών, το οποίο μεταφέρει την έννοια του δεδομένου με τρόπο που είναι κατανοητός από το μοντέλο. Στη συνέχεια, τα embeddings οδηγούνται στον αποκωδικοποιητή, από τα οποία παράγεται η έξοδος του μετασχηματιστή. Η ενσωμάτωση της εισόδου στο μοντέλο υπολογίζεται μέσω των πολλαπλών επιπέδων αυτο-προσοχής(self-attention) του κωδικοποιητή και του αποκωδικοποιητή σε συνδυασμό με νευρωνικά δίκτυα εμπρόσθιας τροφοδότησης..

Συγκεκριμένα η αρχιτεκτονική ενός μετασχηματιστή είναι η εξής:

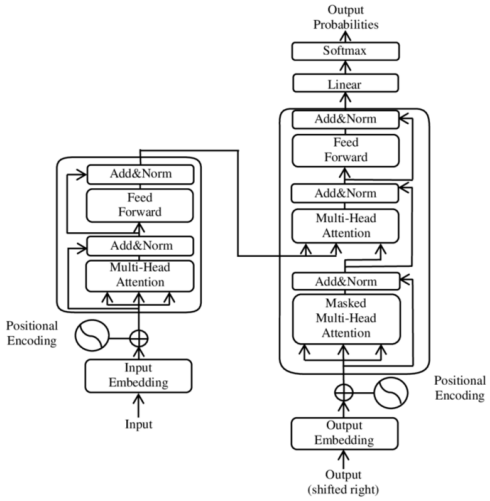
- Στάδιο δημιουργίας μονάδων επεξεργασίας από τα δεδομένα εισόδου, κατά το οποίο λαμβάνεται υπόψη και η θέση του συγκεκριμένου δεδομένου κατά την είσοδό του.

- Επίπεδο ενσωμάτωσης, όπου τα tokens και η συγκεκριμένη θέση που είχαν στην είσοδο του μοντέλου μετατρέπονται σε ένα διάνυσμα.

- Επίπεδα self-attention και νευρωνικά δίκτυα εμπρόσθιας τροφοδότησης, που αποτελούν το ¨σώμα¨ του κωδικοποιητή και του αποκωδικοποιητή.

- Επίπεδο απο-ενσωμάτωσης, όπου τα τελικά διανύσματα αναπαράστασης των δεδομένων μετατρέπονται σε κατανομή πιθανοτήτων για την κάθε μονάδα επεξεργασίας.

Πιο αναλυτικά η αχιτεκτονική ενός μετασχηματιστή μαζί με το διαχωρισμό κωδικοποιητή/αποκωδικοποιητή φαίνεται στο σχήμα:



Εικόνα 2.5: *Τα βασικά μέρη ενός μετασχηματιστή. Όπως φαίνεται, υπάρχει και ένα επίπεδο κανονικοποίησης μετά από τον υπολογισμό του attention τόσο σε encoder όσο και σε decoder. Αυτό αποδείχθηκε ότι βοηθά τη σταθερότητα των τιμών των βαρών κατά την εκπαίδευση του μοντέλου[2].*

Όπως προαναφέρθηκε, το βασικό χαρακτηριστικό των μετασχηματιστών είναι ο υπολογισμός του attention μεταξύ των διαφόρων tokens, ώστε να καταλάβει το μοντέλο πως σχετίζεται το ένα με το άλλλο. Αυτος ο υπολογισμός συμβαίνει μέσα στις μονάδες προσοχής, που αποτελούν τους κωδικοποιητές και τους αποκωδικοποιητές. Συγκεκριμένα, για κάθε μονάδα επεξεργασίας i το διάνυσμα αναπαράστασης x_i πολλαπλασιάζεται με κάθε έναν από τους πίνακες βαρών, τον πίνακα βαρών αιτημάτων W_Q, τον πίνακα βαρών κλειδιών W_K και τον πίνακα βαρών τιμών W_V, για να πάρουμε αντίστοιχα το διάνυσμα αιτήματος q_i, το διάνυσμα κλειδιού k_i και το διάνυσμα τιμής v_i. Τα βάρη της προσοχής ανάμεσα σε δύο tokens, από το i στο j, που αποτελούν μία εκτίμηση του συσχετισμού μεταξύ των δεδομένων που αναπαριστούν αυτά, υπολογίζεται ως το εσωτερικό γινόμενο:

$$q_i \cdot k_j = a_{ij}$$

. Στη συνέχεια ζυγίζουμε τα βάρη προσοχής διαιρώντας το a_{ij} με την τετραγωνική ρίζα της διάστασης των διανυσμάτων κλειδιού:

$$\sqrt{d_k}$$

, ώστε να σταθεροποιηθεί η παράγωγός τους κατά τη διάρκεια της εκπαίδευσης. Τέλος, το αποτέλεσμα εισάγετα σε μία συνάρτηση softmax, για να κανονικοποιηθούν τα βάρη.

Η συνάρτηση softmax είναι μία κανονικοποιημένη εκθετική συνάρτηση με τύπο:

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

, η οποία παίρνει ένα διάνυσμα z με διάφορες τιμές και επιστρέφει πιθαβότητες για κάθε μία από τις τιμές του, οι οποίες καθορίζονται από το πόσο μεγάλη είναι η κάθε τιμή.

Τελικά ο υπολογισμός της προσοχής από μία μονάδα επεξεργασίας προς κάθε άλλη i υπολογίζεται από το ζυγισμένο άθροισμα των όρων $v_j * a_{ij}$, για κάθε μονάδα επεξεργασίας j, όπου v_j το διάνυσμα τιμής κάθε token j. Προκειμένου να αναπαραστήσουμε τον υπολογισμό του attention για κάθε token χρησιμοποιούμε την εξίσωση:

$$Attention(Q, K, V) = softmax(\frac{Q * K^T}{\sqrt{d_k}}) * V$$

.

Ένα σύνολο των τριών πινάκων W_Q, W_K, W_V αποτελεί μία κεφαλή προσοχής(attention head). Η παραπάνω διαδικασία αφορά τον υπολογισμό της προσοχής, εμμέσως δηλαδή του βαθμού που συσχετίζονται τα δεδομένα εισόδου μεταξύ τους, όσον αφορά ένα μόνο χαρακτηριστικό ή γνώρισμά τους. Στην περίπτωση της επεξεργασίας φυσικής γλώσσας θα μπορούσε για παράδειγμα να υπολογίζεται έτσι μόνο η σχέση ρήματος-αντικειμένου μεταξύ λέξεων, οι οποίες είναι σε αυτή την περίπτωση τα δεδομένα εισόδου. Αν θέλουμε να εξάγουμε συμπεράσματα και για ένα άλλο γλωσσικό χαρακτηριστικό, όπως τις σχέσεις ρήματος-υποκειμένου, μπορούμε να χρησιμοποιήσουμε και άλλη κεφαλή προσοχής, στην οποία η διαδικασία υπολογισμού είναι η ίδια και απλώς αλλάζουν οι τρεις πίνακες W_Q, W_K, W_V. Έτσι, καταλήγουν τις περισσότερες φορές οι μετασχηματιστές, που εξάγουν διάφορες εννοιολογικές σχέσεις μεταξύ των δεδομένων εισόδου, να έχουν πολλαπλές κεφαλές προσοχής, για να έχουν μεγαλύτερη ορθότητα στην πρόβλεψη της σωστής εξόδου.

Τέλος, αξίζει να αναφερθεί ότι ο μετασχηματιστής ανήκει σε μία ειδική κατηγορία μοντέλων μηχανικής μάθησης, εφόσον αποτελεί μοντέλο αυτο-επιβλεπόμενης μάθησης. Αυτό σημαίνει πως δεν χρειάζεται επισημάνσεις για τα δεδομένα εισόδου, προκειμένου να εντοπίσει μοτίβα σε αυτά και να μπορέσει να τα κατατάξει σε γενικές κατηγορίες. Αυτή την ικανότητα οι μετασχηματιστές την αποκτούν μέσα από τον υπολογισμό της προσοχής μεταξύ των δεομένων εισόοδυ. Είναι σύνηθες φαινόμενο η προεκπαίδευση του μοντελου να γίνεται με μη επιβλεπόμενο τρόπο, αλλά στη συνέχεια να ρυθμκστεί με ακρίβεια σε ένα σύνολο δεδομένων με επιβλεπόμενο τρόπο[20].

Αποκρύπτοντες Αυτοκωδικοποιητές

Αυτό το μοντέλο, που παράχθηκε από την ερευνητική ομάδα του Facebook το 2022[21], αποτελεί μία μίξη των μετασχηματιστών και του ειδικού νευρωνικού δικτύου των αυτοκωδικοποιητών. Ο αυτοκωδικοποιητής είναι μία παλαιότερη αρχιτεκτονική νευρωνικών δικτύων, ο οποίος ουσιαστικά αποτελείται από ένα επίπεδο συμφόρησης, έναν κωδικοποιητή και έναν

αποκωδικοποιητή, η λειτουργία των οποίων όμως διαφέρει από αυτή των τμημάτων με το ίδιο όνομα σε έναν μετασχηματιστή. Συγκεκριμένα, το κομμάτι του κωδικοποιητή συμπιέζει τα δεδομένα εισόδου και τα αναπαριστά με διανύσματα. Σκοπός του είναι να ελαττώσει το μέγεθος της εισόδου, μειώνοντας τη διάσταση των δεδομένων που αναπαριστά και συμπυκνώνοντας την πληροφορία τους. Το επίπεδο συμφόρησης είναι το κομμάτι του δικτύου, όπου συγκεντρώνονται όλες οι αναπαραστάσεις της εισόδου με χαμηλότερη διάσταση σε σχέση με αυτήν, με την οποία εισήχθησαν στο δίκτυο. Σε αυτό το επίπεδο λοιπόν είναι συμπυκνωμένη όλη η πληροφορία της εισόδου, με τέτοιον τρόπο, ώστε οι αναπαραστάσεις δεδομένων κοντά σημασιολογικά και εννοιολογικά να είναι και αυτές κοντά. Στον αποκωδικοποιητή, το τελευταίο κομμάτι του νευρωνιού δικτύου, αποσυμπιέζονται οι αναπαραστάσεις γνώσεις που έχουν εξαχθεί στα προηγούμενα επίπεδα, και ανακατασκευάζεται η είσοδος από τις επιμέρους αυτές αναπαραστάσεις με όσο το δυαντόν μεγαλύτερη ακρίβεια γίνεται.

Πρόκειται για ένα μοντέλο μη επιβλεπόμενης μάθησης, εφόσον δεν χρειάζεται επισημάνσεις για τα δεδομένα εισόδου και μαθαίνει μόνο του από τα δεδομένα εισόδου, με τέτοιο τρόπο ώστε να φτιάξει ένα πιστό αντίγραφό τους στην έξοδο. Οι αυτοκωδικοποιητές έχουν εφαρμοστεί τόσο σε προβλήματα NLP[22] όσο και σε προβλήματα όρασης υπολογιστών[23]. Περισσότερη επιτυχία όμως είχε στην όραση υπολογιστών, εφόσον οι εικόνες χωρίζονται σε κομμάτια, τα pixels, οι τιμές των οποίων παρουσιάζουν μία συνέχεια μέσα σε μία εικόνα , και στο επίπεδο συμφόρησης οι αναπαραστάσεις τους τοποθετούνται ορθώς κοντά η μία με την άλλη[24]. Αντιθέτως, στην επεξεργασία φυσικής γλώσσας, οι λέξεις, τα δεδομένα εισόδου δηλαδή, είναι διακριτά αντικείμενα και κάποιες φορές δεν μπορούν με σωστό τρόπο να τοποθετηθούν κοντά οι έννοιές τους, με αποτέλεσμα να δυσκολεύεται και ο αυτοκωδικοποιητής να τοποθετήσει κοντά τις αναπαραστάσεις τους στο επίπεδο συμφόρησης, από το οποίο μετά ανασκευάζει τα δεδομένα.

Προκειμένου να ξεπεραστεί το πρόβλημα που αντιμετώπιζε το δίκτυο των αυτοκωδικοποιητών στο NLP, δοκιμάστηκε η αυτοκωδικοποίηση με χρήση μασκών στα δεδομένα εισόδου στο μοντέλο BERT με μεγάλη επιτυχία[25]. Συγκεκριμένα, το μοντέλο τοποθετεί μάσκες, αποκρύπτοντας από το επίπεδο του κωδικοποιητή, σε ορισμένα από τα δεδομένα εισόδου και στη συνέχεια το πρόβλημα μετατρέπεται από πρόβλημα ικανοποιητικής αναπαραγωγής της εισόδου στην έξοδο σε ένα πρόβλημα σωστής πρόβλεψης του δεδομένου που λείπει, της λέξης στην περίπτωση του NLP. Σε αυτή την εκδοχή του αυτοκωδικοποιητή έχουμε χρήση μετασχηματιστών στη θέση του κωδικοποιητή, ο οποίος μαθαίνει μέσα από τον υπολογισμό της προσοχής την έννοια της κρυμμένης λέξης, και τελικά υπολογίζει με κάποια πιθανότητα ποια μπορεί να είναι αυτή.

Στο επιστημονικό άρθρο της ομάδας του Facebook το 2022 έγινε η προσπάθεια να επεκταθεί αυτή η ιδέα της χρήσης μασκών στους αυτοκωδικοποιητές και σε μοντέλα στον τομέα της όρασης υπολογιστών, τα οποία αφενός θα μπορούσαν με ακρίβεια να αναπαράγουν την είσοδο μίας εικόνας στην έξοδο, αφετέρου ο υπολογισμός της αναπαράστασης όλων των pixels της εικόνας εισόδου, πόσο μάλλον ενός βίντεο, θα απαιτούσε πολύ χρόνο και πολλούς υπολογιστικούς πόρους. Με την απόκρυψη όμως ενός μέρους των δεδομένων το πρόβλημα αυτό του υπολογισμού της αναπαράστασης των δεδομένων αντιμετωπίζεται σε σημαντικό βαθμό. Μάλιστα, όπως προαναφέρθηκε, τα εικονοστοιχεία μίας εικόνας παρουσιάζουν μία συνέχεια και σχετίζονται σε μεγάλο βαθμό με τα γειτονικά τους, με αποτέλεσμα να μπορεί να εξαχθεί

ένα ασφαλές συμπέρασμα από την τιμή ενός για την τιμή των υπολοίπων. Για αυτό το λόγο, στο paper αποδεικνύεται ότι η απόκρυψη του 75% πετυχαίνει τη μείωση των δεδομένων, που θα γίνουν αντικείμενο επεξεργασίας από τον μετασχηματιστή, χωρίς να μειώνεται η ορθότητα του μοντέλου, ενώ το αντίστοιχο ποσοστό απόκρυψης όρων στην περίπτωση της επεξεργασίας φυσικής γλώσσας είναι στο 15% περίπου.

Η διαδικασία που ακολουθεί το μοντέλο ενός αποκρύπτοντος αυτοκωδικοποιητή με μία εικόνα στην είσοδο είναι η εξής:

1. Χωρίζεται η εικόνα σε τμήματα, από τα οποία στο 75% εφαρμόζεται μία μάσκα, με αποτέλεσμα να μην περάσουν στον κωδικοποιητή/μετασχηματιστή του επόμενου επιπέδου.

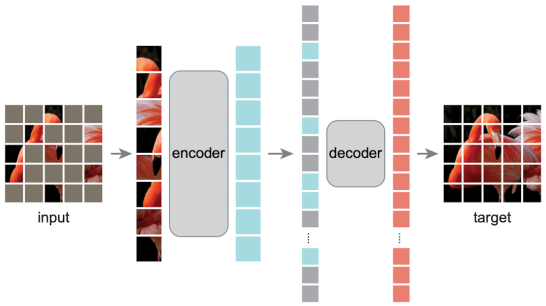
2. Στο μετασχηματιστή υπολογίζονται οι αναπαραστάσεις των τμημάτων, λαμβάνοντας υπόψη και την προσοχή από το καθένα στα υπόλοιπα.

3. Οι αναπαραστάσεις, ουσιαστικά οι μονάδες επεξεργασίας που έχει παράγει ο μετασχηματιστής, οδηγούνται στο επίπεδο συμφόρησης, όπου προστίθενται και τα τμήματα που έχουν αποκρυφθεί στην αρχή, ώς επιπλέον μονάδες επεξεργασίας.

4. Ο αποκωδικοποιητής έχοντας στη διάθεσή του πληροφορία μόνο για τα τμήματα της αρχικής εικόνας, που δεν είχαν αποκρυφθεί κατά την είσοδο, προσπαθεί να ανασκευάσει την αρχική εικόνα

Εικόνα 2.6: *Φαίνονται τα πειράματα όσον αφορά τη στρατηγική χρήσης μασκών στα τμήματα της αρχικής εικόνας, πριν αυτή εισαχθεί στον κωδικοποιητή/μετασχηματιστή. Παρατηρούμε πως η απόκρυψη του 75% των τμημάτων της εικόνας με τυχαίο τρόπο παράγει το βέλτιστο αποτέλεσμα στην έξοδο του μοντέλου. Τα παραδείγματα αυτά είναι από το σύνολο εικόνων επαλήθευσης.*

Πρέπει να σημειωθεί ότι η παραπάνω διαδικασία αποτελεί κομμάτι της προεκπαίδευσης του μοντέλου. Κατά τη διαδικασία της ρύθμισης με ακρίβεια σε συγκεκριμένο σύνολο δεδομένων εισάγονται στο μοντέλο ολόκληρες εικόνες. Έτσι ο ήδη εκπαιδευμένος κωδικοποιητής/μετασχηματιστής σε συνδυασμό με τον εκπαιδευμένο αποκωδικοποιητή προβλέπει σε ποια κατηγορία ανήκει η είσοδος. Επίσης, αξίζει να αναφερθεί πως τα τμήματα, τα οποία αποκρύπτονται από το μοντέλο, επιλέγονται τυχαία. Αυτό συμβαίνει, γιατί μετά από πειράματα των ερευνητών, παρατήρησαν πως, αν οι μάσκες εφαρμόζονται μόνο σε συγκεκριμένα κομματια της εικόνας, παραδείγματος χάριν στο περίγραμμά της, επειδή εκεί συνήθως δεν αναπαρίσταται κίνηση, η ανασκευή της εικόνας από το μοντέλο είναι κακής ποιότητας.



Εικόνα 2.7: *Το παράδειγμα ενός αποκρύπτουτος αυτοκωδικοποιητή στη διαδικασία της προεκπαίδευσης. Παρατηρείται πως το 75% των τμημάτων της εικόνας του φλαμίνγκο καλύπτουται από μάσκες και εισέρχουται ξανά στους υπολογισμούς μετά το επίπεδο του κωδικοποιητή/μεταασχηματιστή. Το γεγονός ότι ένα σημαντικό τμήμα των δεδομένων εισόδου δεν γίνεται αντικείμενο επεξεργσίας από το μεταασχηματιστή, δίνει την ευχέρεια επιλογής ενός μοντέλου με μεγάλο βάθος και πολλές παραμέτρους, συνεπώς και μεγαλύτερης ακρίβειας.*

Κεφάλαιο 3

Περιγραφή Θέματος

Στο κεφάλαιο αυτό αρχικά γίνεται μια περιγραφή της αναγνώρισης ανθρώπινης δράσης και τους λόγους για τους οποίους αυτή είναι χρήσιμη. Επιπλέον, παρουσιάζεται η εξέλιξη των εργαλείων μηχανικής μάθησης που έχουν χρησιμοποιηθεί στο πρόβλημα της αναγνώρισης ανθρώπινης δράσης σε εικόνες και βίντεο, ξεκινώντας με τα συνελικτικά νευρωνικά δίκτυα, συνεχίζοντας με τις μοντέρνες αρχιτεκτονικές των μετασχηματιστών και καταλήγοντας στους αποκρύπτοντες αυτοκωδικοποιητές, τα μοντέλα τελευταίας τεχνολογίας που αξιοποιούνται για την επίλυση τέτοιων προβλημάτων.

3.1 Αναγνώριση Ανθρώπινης Δράσης

3.1.1 Ορισμός

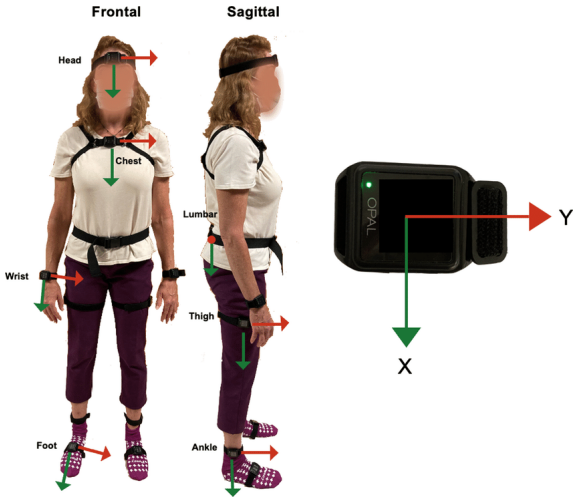
Το πρόβλημα της αναγνώρισης της ανθρώπινης δράσης, δηλαδή της κατηγοριοποίησης των πράξεων και κινήσεων που εκτελούν άνθρωποι σε εικόνες και βίντεο ανήκει στον τομέα της όρασης υπολογιστών, εφόσον έχει να κάνει με τη μετάφραση των δεδομένων από τον οπτικό κόσμο σε αναπαράσταση κατανοητή από έναν υπολογιστή. Συχνά σε τέτοια προβλήματα υπάρχουν συγκεκριμένες κατηγορίες δράσεων, που μπορεί να εκτελεί το ανθρώπινο υποκείμενο. Με αυτόν τον τρόπο, τα μοντέλα της όρασης υπολογιστών που χρησιμοποιούνται, για να επιλύσουν τέτοια προβλήματα, εκπαιδεύονται είτε με είτε χωρίς επίβλεψη σε οπτικά σύνολα δεδομένων και στη συνέχεια καλούνται να προβλέψουν σωστά τη σωστή επισήμανση για την κίνηση που υπάρχει σε κάθε δεδομένο. Αυτή η σωστή και αυτόματη από μηχάνημα πρόβλεψη της ανθρώπινης δράσης σε μία εικόνα ή ένα βίντεο βρίσκει εφαρμογές σε πολλούς τομείς όπως ο αθλητισμός, η υγεία και η ασφάλεια.

3.1.2 Συλλογή Δεδομένων

το πρώτο βήμα, για να μπορέσει να εκπαιδευθεί ένα νευρωνικό δίκτυο, αποτελεί η συγκέντρωση δεδομένων, τα οποία μπορεί να επεξεργαστεί. Στις ημέρες μας, υπάρχει τεράστιο διαθέσιμο οπτικοακουστικό υλικό στο διαδίκτυο, το οποίο δημιουργείται από τους ίδιους τους χρήστες του ίντερνετ μέσα από τις κάμερες του έξυπνου κινητού τους, αλλά και μέσα από επαγγελματικές κάμερες. Με αυτόν τον τρόπο διευκολύνεται η δουλειά των ερευνητών, αφού έχουν πρόσβαση σε πάρα πολλά δεδομένα με πάρα πολλά είδη δράσεων, τα οποία μπορούν να κατεβάσουν και να ¨ταΐσουν¨ με αυτά τα μοντέλα τους.

Παρ' όλα αυτά, υπάρχουν και πιο συγκεκριμένες κατηγορίες δράσεων, ειδικά στην περίπτωση του ιατρικού τομέα. Σε αυτόν τον τομέα ακόμα και οι παραμικρές κινήσεις των ασθενών λαμβάνονται σημαντικά υπόψιν από τους ειδικούς για τη διάγνωση της κατάστασής τους. Για αυτό, σε τέτοιες περιπτώσεις ευαίσθητες κάμερες σε συνδυασμό με αισθητήρες υψηλής ακρίβειας χρησιμοποιούνται για την παραγωγή δεδομένων με ακρίβεια.

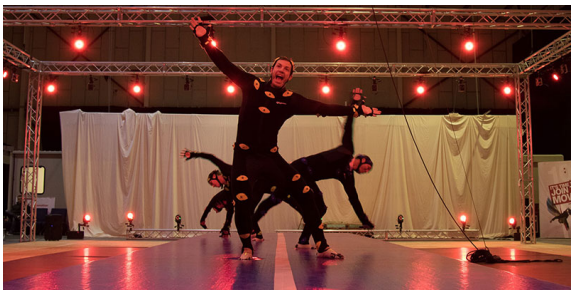
Ένα από αυτά τα συστήματα, που έχουν χρησιμοποιηθεί για τη συλλογή ειδικών δεδομένων στον τομέα της υγείας, είναι οι μικροσκοπικές μονάδες μέτρησης αδράνειας(IMUs). Πρόκειται για ελαφριά, μικρά και οικονομικά συστήματα νέας γενιάς με ενσωματωμένα τρισδιάστατα επιταχυνσιόμετρα, γυροσκόπια και μαγνητόμετρα. Έτσι μπορούν να πάρουν ακριβείς μετρήσεις και αν αξιολογήσουν την κίνηση των ανθρώπινων μυών. Τα IMUs μπορούν να παρακολουθήσουν την τροχιά των ανατομικών χαρακτηριστικών του ανθρώπου σε πραγματικό χρόνο και να εκτιμήσουν τις κινηματικές παραμέτρους του κύκλου βηματισμού του[26]. Αν και τα πρωτόκολλα αξιολόγησης δεν είναι ομογενή, αρκετές έρευνες εκτιμούν πως υπάρχει η δυνατότητα να αποτιμηθεί η ανάλυση του ανθρώπινου βηματισμού μέσω των IMUs[27, 28, 29, 30, 31]. Σε μία έρευνα που διενεργήθηκε από τους Fusca et all[32], επιλέχθηκαν 10 εθελοντές, πάνω στους οποίους τοποθετήθηκαν σημάδια για την καταγραφή κίνησης με τη χρήση του Elite (BTS) System σε συνδυασμό με αισθητήρες IMU κοντά στο κέντρο βάρους των συμμετεχόντων. Τα αποτελέσματα επιβεβαίωσαν ότι οι αισθητήρες αυτοί είναι αποτελεσματικοί για μία ακριβή αξιολόγηση των χωροχρονικών παραμέτρων του ανθρώπινου βηματισμού. Στην εικόνα που ακολουθεί φαίνεται πώς τοποθετούνται τα IMUs σε ανθρώπους.



Εικόνα 3.1: *Γυναίκα που φοράει αισθητήρες IMU.*

Υπάρχει επίσης η ανάγκη για τη χρήση ειδικών αισθητήρων στον κόσμο του κινηματογράφου. Στο σύγχρονο σινεμά, όπου είναι πολύ διαδεδομένη η χρήση της τεχνολογίας CGI, κρίνεται απαραίτητη η ακριβής καταγραφή των κινήσεων των ηθοποιών κατά το γύρισμα, ώστε στη συνέχεια της παραγωγής της ταινίας να παραχθεί με τη βοήθεια του CGI το επιθυμητό αποτέλεσμα. Συγκεκριμένα, οι ηθοποιοί καλούνται αν φορέσουν μία ειδική στολή αποτελούμενη από αισθητήρες, που ονομάζονται MOCAPS και είναι τοποθετημένοι στα καίρια σημεία του σώματός τους, ώστε να καταγράφεται η κίνησή τους. Η θέση των άκρων και του κορμού του ηθοποιού υπολογίζεται από την ποσότητα εκπεμπόμενου φωτός που αντα-

νακλάται από τους αισθητήρες πίσω στις κάμερες. Σε συνδυασμό με αυτούς τους αισθητήρες χρησιμοποιούνται και άλλα όργανα μέτρησης της κίνησης των ηθοποιών, όπως γυροσκόπια, επιταχυνσιόμετρα και μαγνητόμετρα [33, 34].



Εικόνα 3.2: *Στολές καταγραφής κινήσεων με αισθητήρες.*

3.1.3 Προεπεξεργασία Δεδομένων

Το στάδιο που έπεται της συλλογής δεδομένων είναι η προεπεξεργασία τους. Η χρήση τεχνικών προεπεξεργασίας των δεδομένων είναι πολύ σημαμντική, καθώς δεν είναι όλα τα δεδομένα στην κατάλληλη μορφή, για να εισαχθούν στα μοντέλα μηχανικής μάθησης. Είναι αρκετά συχνό φαινόμενο τα δεδομένων να περιέχουν θόρυβο, οποίος αφενός δυσκολεύει την επεξεργασία τους και αφετέρου μπορεί να εκπαιδεύσει με λάθος τρόπο το μοντέλο. Με αυτόν τον τρόπο προβλέπει λανθασμένα αποτελέσματα.

Στην παραπάνω ενότητα παρουσιάσαμε ορισμένους αισθητήρες, που παράγουν δεδομένα χρήσιμα για την εκπαίδευση μοντέλων. Το θέμα αυτής της διπλωματικής αφορά τον τομέα της όρασης υπολογιστών, θα εστιάσουμε σε τεχνικές προεπεξεργασίας οπτικών δεδομένων και συγκεκριμένα εικόνων. Αυτές οι τεχνικές μπορούν εύκολα να εφαρμοστούν και σε βίντεο, αν θεωρήσουμε τα βίντεο μία αλληλουχία εικόνων.

Ένας από τους κυριότερους τρόπους προεπεξεργασίας εικόνων αποτελεί το φιλτράρισμα τους. Υπάρχει ένας μεγάλος αριθμός χωρικών φίλτρων που εφαρμόζονται στα οπτικά δεδομένα και χωρίζονται σε δύο κατηγορίες, τα γραμμικά και μη γραμμικά φίλτρα[35]. Αρχικά, τα γραμμικά φίλτρα εφαρμόζονταν, για να αφαιρούν τον θόρυβο από τις εικόνες, αλλά δεν είχαν τη δυνατότητα να διατηρήσουν και τις υφές τους. Για παράδειγμα, το μέσο φιλτράρισμα[36] έχει χρησιμοποιηθεί για τη μείωση του γκαουσιανού θορύβου, αλλά μπορεί να υπερλειάνει τις εικόνες με χαρακτηριστικά υψηλού θορύβου[37]. Το φίλτρο Wiener[38, 39] προσπαθεί να ξεπεράσει αυτήν την αδυναμία, και για αυτό όμως υπάρχει η πιθανότητα να θολώσει ¨αιχμηρά¨ χαρακτηριστικά των εικόνων. Εφαρμόζοντας μη γραμμικά φίλτρα, όπως το διάμεσο φίλτρο[40] και το ζυγισμένο μέσο φίλτρο[41], ο θόρυβος μπορεί να αποκρυφθεί χωρίς αναγνώριση της ακριβούς πηγής του θορύβου. Έχουν την ικανότητα αυτά τα φίλτρα να διαχωρίζουν μεταξύ του κανονικού σήματος και του θορύβου με τη χρήση στατιστικών εργαλείων.

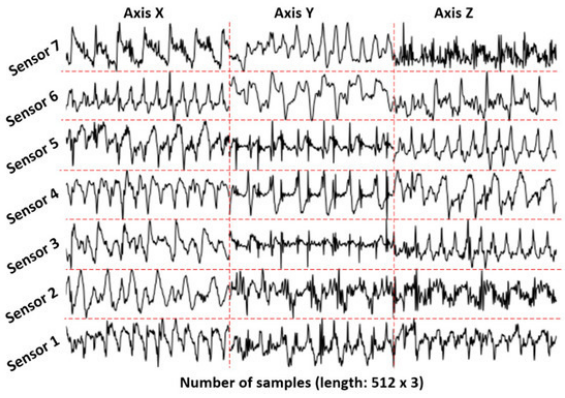
Μία αρκετά διαδεδομένη τεχνική για τη μείωση του θορύβου στις εικόνες αποτελεί και ο καταμερισμός εικόνας. Συγκεκριμένα, η εικόνα χωρίζεται σε κομμάτια βάσει ορισμένων κριτηρίων, όπως είναι η ένταση των εικονοστοιχείων, το χρώμα και η υφή. Με αυτόν τον τρόπο δεν απομονώνεται μόνο ο θόρυβος, αλλά και σε αρκετές περιπτώσεις και κάποιο αντικείμενο

ενδιαφέροντος, που διαδραματίζει σημαντικό ρόλο στην εκπαίδευση του μοντέλου[42]. Μερικοί από τους τρόπους να καταμερίσουμε μία εικόνα είναι η εφαρμογή μίας τιμής κατωφλιού για την ένταση των τιμών των pixels[43], ο εντοπισμός ακμών και η συσταδοποίηση[44].

Υπάρχουν και μέθοδοι προεπεξεργασίας των δεδομένων που δεν έχουν ως κύριο στόχο τους την αφαίρεση του θορύβου, αλλά την επαύξηση της εικόνας. Η επαύξηση των δεδομένων εν γένει συμβάλλει στο να αυξηθεί η ποικιλία των διαθέσιμων δεδομένων καταπολεμώντας έτσι την έλλειψή τους σε πολλούς τομείς. Μία από τις πιο κλασικές τεχνικές αποτελεί η αλλαγή μεγέθους ή διαστάσεων της εικόνας, η οποία αποσκοπεί στην ύπαρξη μίας ομοιογένειας όσον αφορά τα δεδομένα εισόδου, να μην υπάρχουν δηλαδή εικόνες διαφορετικών αναλύσεων. Μειώνοντας τις διαστάσεις, μπορούμε εμμέσως να μειώσουμε και τον όγκο πληροφορίας, που έχει να διαχειριστεί το μοντέλο. Έτσι, ελαττώνεται και ο απαιτούμενος χρόνος εκπαίδευσης αλλά και οι υπολογιστικοί πόροι.

Άλλες μέθοδοι, που έχουν να κάνουν με την επαύξηση της εικόνας με αλλαγή χρώματος για παράδειγμα, είναι η εξομοίωση ιστογράμματος[45] και η ανάλυση κύριων συνιστωσών[46]. Πρόκειται για πιο εξελιγμένες μεθόδους, οι οποίες ουσιαστικά κωδικοποιούν την εικόνα και την αναπαριστούν με διαφορετικό τρόπο. Με αυτόν τον τρόπο, έχουν τη δυνατότητα να προβάλλουν χαρακτηριστικά, που δεν φαίνονται με ευκολία πριν την εφαρμογή των τεχνικών αυτών. Χρήση των συγκεκριμένων εργαλείων γίνεται για την εξαγωγή χαρακτηριστικών όπως οι ακμές, τα σχήματα και τα σημεία αλλαγής χρωματισμού σε μία εικόνα.

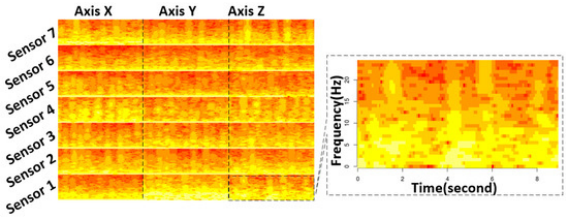
Τέλος, για χάρη πληρότητας της ενότητας, αναφέρουμε μερικές τεχνικές προεπεξεργασίας, που μετατρέπουν τα δεδομένα από ειδικούς αισθητήρες, όπως αυτοί που είδαμε παραπάνω, σε συνοπτικά διαγράμματα. Αυτά τα εργαλεία δεν χρησιμοποιούνται τόσο για την προετοιμασία των δεδομένων για το μοντέλο, αλλά για να παραχθεί μία οπτική οργάνωση και σύνοψη των δεδομένων, ώστε να είναι πιο εύκολα κατανοητά από τους ερευνητές.Η μέθοδος ακατέργαστων διαγραμμάτων μετατρέπει τα δεδομένα από τους αισθητήρες σε χρονοσειρά εικόνων. Οι τρεις χωρικοί άξονες ομαδοποιούνται ανά στήλη και τα δεδομένα που έχουν ληφθεί από έναν αισθητήρα συγκεντρώνονται στην ίδια γραμμή. Έτσι προκύπτει ένα δισδιάστατο ασπρόμαυρο διάγραμμα με τρεις στήλες, μία για κάθε άξονα, και με τόσες γραμμές, όσοι ήταν και οι αισθητήρες[47].



Εικόνα 3.3: *Ακατέργαστο διάγραμμα για δεδομένα επιταχυνσιόμετρου.*

Το φασματογράφημα είναι ένας άλλος τρόπος μεταφοράς δεδομένων από αισθητήρες σε γράφημα και παράγεται ειδικά για δεδομένα ήχου, που συγκεντρώνονται πολλές φορές κατά

την καταγραφή βίντεο. Αναπαριστά τις συχνότητες του σήματος στη διάσταση του χρόνου. Περιέχει το τετράγωνο του βραχυπρόθεσμου μετασχηματισμού Fourier, ο οποίος καθορίζει τη συχνότητα και τη φάση του ημιτόνου, που περιγράφει το σήμα με την πάροδο του χρόνου. Η διαδικασία υπολογισμού και δημιουργίας ενός φασματογραφήματος είναι η διαίρεση ενός σήματος μεγάλης χρονικής διάρκειας σε μικρότερα παράθυρα και ο υπολογισμός του μετασχηματισμού Fourier σε κάθε επιμέρους παράθυρο. Στο τελικό διάγραμμα αναπαρίστανται οι τιμές για κάθε τέτοιο μικρό παράθυρο[48].



Εικόνα 3.4: *Φασματογράφημα με δεδομένα επιτάχυνσης από 7 αισθητήρες.*

3.1.4 Επιλογή Μοντέλου

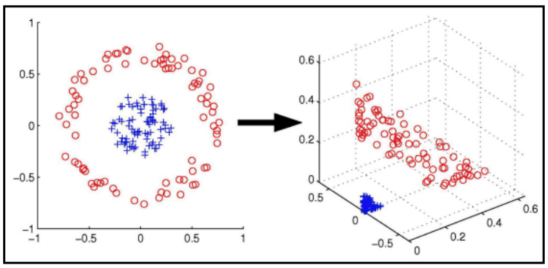
Τη διαδικασία της προεπεξεργασίας των δεδομένων ακολουθεί η επιλογή του μοντέλου, το οποίο θα αναλύσει τα δεδομένα εισόδου, θα μάθει από αυτά και θα παράξει τη σωστή έξοδο. Υπάρχουν δύο κατηγορίες μοντέλων, τα κλασικά μοντέλα μηχανικής μάθησης (τυχαία δάση, κρυφά μοντέλα Markov και μηχανές διανυσμάτων υποστήριξης) και οι αλγόριθμοι βαθιάς μάθησης (RNNs, LSTMs, CNNs, και transformers). Αν και σε ένα άρθρο του J. Gao et all.[49] συγκρίνεται η απόδοση των τεχνικών βαθιάς μάθησης με αυτή των κλασικών μεθόδων και αποδεικνύεται ότι οι πρώτες ξεπερνάνε τις δεύτερες όσον αφορά την ορθότητα των αποτελεσμάτων, την ανθεκτικότητα σε παραλλαγές των δεδομένων εισόδου καθώς και στην ικανότητα να μαθαίνουν πολύπλοκα χαρακτηριστικά, κρίνουμε σημαντικό να παρουσιαστούν τεχνικές και των δύο κατηγοριών.

Οι αλγόριθμοι δέντρων αποφάσεων χρησιμοποιούνται για την εύρεση μη γραμμικών σχέσεων μεταξύ διαφόρων κατηγοριών. Είναι συχνή η χρήση τους σε προβλήματα αναγνώρισης ανθρώπινης δράσης, όπου η είσοδος είναι δεδομένα αισθητήρων, όπως επιταχυνσιόμετρα ή γυροσκόπια. Το πλεονέκτημα αυτού του μοντέλου είναι πως έχει τη δυνατότητα να ερμηνεύσει και να διαχειριστεί τόσο διακριτά όσο και συνεχή δεδομένα. Για αυτό είναι ιδανικό για επεξεργασία μετρήσεων αισθητήρων, των οποίων οι τιμές είναι συνεχείς. Τα τυχαία δάση είναι μία εξέλιξη των δέντρων απόφασης, εφόσον αποτελούνται από ένα σύνολο δέντρων αποφάσεων και αυτό τους δίνει τη δυνατότητα να επεξεργάζονται θορυβώδη δεδομένα ή δεδομένα πολλών διαστάσεων.

Τα κρυφά μαρκοβιανά μοντέλα αποτελούνται από έναν αριθμό καταστάσεων, οι οποίες είναι συνδεδεμένες η καθεμία με την επόμενή της, συνεπώς προκύπτει ένα διασυνδεδεμένο δίκτυο καταστάσεων. Κάθε κατάσταση εκπέμπει ένα σύμβολο με κάποια πιθανότητα και το μεταφέρει στην επόμενη κατάσταση πάλι με ορισμένη πιθανότητα. Ξεκινώντας από την κατάσταση 0, μία ακολουθία καταστάσεων προσπελάζεται με βάση τις πιθανότητες μετάβασης από κατάσταση σε κατάσταση, μέχρι να φτάσει σε μία τελική κατάσταση, από όπου δεν υπάρχει πιθανότητα μετάβασης σε επόμενη. Κάθε μία από τις καταστάσεις εκπέμπει και

ένα σήμα με συγκεκριμένη πιθανότητα που ακολουθεί συγκεκριμένη κατανομή, με αποτέλεσμα να έχει δημιουργηθεί έως το τέλος μία αλληλουχία συμβόλων. Τα HMMs έχουν κριθεί από ερευνητές κατάλληλα όχι μόνο για την κατηγοριοποίηση ακουστικών δεδομένων, καθώς αποτελούν ένα χρήσιμο εργαλείο της αναγνώρισης φωνής, αλλά και για τη μοντελοποίηση βιολογικών τρισδιάστατων δομών, όπως πρωτεΐνες[50]. Πρόσφατα έχουν υπάρξει και πειράματα με τα μοντέλα αυτά και στον τομέα της αναγνώρισης ανθρώπινης δράσης[51].

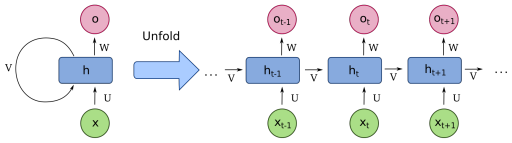
Οι μηχανές διανυσμάτων υποστήριξης είναι μοντέλα επιβλεπόμενης μάθησης, που χρησιμοποιούνται για την ανάλυση δεδομένων σε προβλήματα ταξινόμησης αλλά και παλινδρόμησης. Η εφαρμογή των SVMs δεν περιορίζεται στην γραμμική ταξινόμηση, εφόσον επιτρέπουν τη χρήση συναρτήσεων πυρήνα, οι οποίοι αλλάζουν τη διάσταση των δεδομένων. Έτσι το μοντέλο έχει τη δυνατότητα να βρει ένα χώρο πολλαπλών διαστάσεων, όπου η είσοδος πλέον θα είναι γραμμικά διαχωρίσιμη. Για αυτό το λόγο, τα SVMs έχουν καλή απόδοση σε προβλήματα μη γραμμικής ταξινόμησης και προσφέρουν την ευελιξία του να μπορούν να εφαρμοστούν σε ένα μεγάλο εύρος προβλημάτων. Όσον αφορά τη συνεισφορά τους στα προβλήματα αναγνώρισης ανθρώπινης δράσης, τα οπτικά δεδομένα σε πρώτη φάση πρέπει να μετατραπούν σε αριθμητικές τιμές και διανύσματα, τα οποία μπορεί αν επεξεργαστεί το μοντέλο. Σε αυτό το πλαίσιο έχουν χρησιμοποιηθεί σε συνεργασία με μοντέλα ειδικά για την όραση υπολογιστών, όπως τα συνελικτικά νευρωνικά δίκτυα, για να αναγνωρίσουν αποδοτικά ανθρώπινες κινήσεις σε εικόνες και βίντεο[52].



Εικόνα 3.5: *Παράδειγμα μίας μηχανής διανυσμάτων υποστήριξης, που χρησιμοποιεί μία συνάρτηση πυρήνα, προκειμένου να αλλάξει τη διάσταση των δεδομένων και να μπορέσει να τα ταξινομήσει.*

Η πρώτη τεχνική βαθιάς μάθησης που θα παρουσιάσουμε είναι τα επαναλαμβανόμενα νευρωνικά δίκτυα. Πρόκειται για ένα τεχνητό νευρωνικό δίκτυο διπλής κατεύθυνσης με πολλά επίπεδα. Η ιδιαίτερη αυτή αρχιτεκτονική του επιτρέπει η έξοδος των νευρώνων να επηρεάζει την είσοδο των προηγούμενων, των επόμενων ακόμα και των ίδιων των κόμβων. Συνεπώς, δεν ανήκει στην κατηγορία των δικτύων εμπρόσθιας κατεύθυνσης, εφόσον δεν λειτουργεί μόνο προς μία κατεύθυνση. Η κλασική εκδοχή των RNNs συνδέει τις εξόδους όλων των νευρώνων με τις εισόδους όλων των νευρώνων, όπως φαίνεται στην κάτωθι εικόνα.

Μολονότι το RNN είναι ένα πολύ χρήσιμο μοντέλο, παρουσιάζει ένα βασικό πρόβλημα με δύο όψεις. Λόγω της ικανότητάς του να μεταφέρει την είσοδο από τον έναν κόμβο στον επόμενο ή στον προηγούμενο, τα επαναλαμβανόμενα νευρωνικά δίκτυα εμφανίζουν είτε το πρόβλημα της εκμηδένισης είτε το πρόβλημα της εκτόξευσης της τιμής της κλίσης. Πιο συγκεκριμένα, κατά τη διάρκεια της οπισθοδιάδοσης για την διόρθωση των παραμέτρων



Εικόνα 3.6: *Συμπιεσμένο(αριστερά) και ξεδιπλωμένο(δεξιά) βασικό επαναλαμβανόμενο νευρωνικό δίκτυο.*

του δικτύου, η τιμή της κλίσης στους νευρώνες των αρχικών επιπέδων(των επιπέδων που βρίσκονται κοντά στην είσοδο του δικτύου) προκύπτει από τον πολλαπλασιασμό των τιμών της παραγώγου της συνάρτησης σφάλματος ως προς τα βάρη και ως προς τον συντελεστή συστημικού σφάλματος όλων των επόμενων κόμβων. Επομένως, από τη μία μεριά, αν είναι η τιμή των παραγώγων αρκετά μεγαλύτερη του ένα, ο πολλαπλασιασμός τους οδηγεί στην εκτίναξη της τιμής της κλίσης στους αρχικούς νευρώνες στο άπειρο. Από την άλλη μεριά, αν είναι σημαντικά μικρότερη του ένα, τότε μέσω του πολλαπλασιασμού όλων αυτών των παραγώγων, η τιμή της κλίσης στους νευρώνες των πρώτων επιπέδων μειώνεται με εκθετικό ρυθμό. Σε καμία από τις δύο περιπτώσεις, η τιμή της κλίσης δεν επιτρέπει στον αλγόριθμο διαβάθμισης κλίσης, που τη χρησιμοποιεί, να βρει με επιτυχία τη βέλτιστη τιμή παραμέτρων.

Μία λύση σε αυτό το πρόβλημα προσφέρουν τα νευρωνικά δίκτυα μακράς βραχύχρονης μνήμης. Είναι μία εξέλιξη των RNNs, αφού είναι σχεδιασμένα να διατηρούν τις σημαντικές χρονικές συνδέσεις μεταξύ νευρώνων μακρινών επιπέδων, αλλά και ταυτόχρονα να διαγράφουν πληροφορία, όσον αφορά την έξοδο προηγούμενων κόμβων στο δίκτυο που δεν κρίνονται τόσο χρήσιμοι για την εξαγωγή αποτελέσματος. Με αυτόν τον τρόπο αποφεύγεται ο κίνδυνος να μηδενιστεί ή να απειριστεί η τιμή της κλίσης στους νευρώνες. Λόγω της ικανότητάς τους να διατηρούν τη σύνδεση μεταξύ κόμβων, που απέχουν πολλά επίπεδα μεταξύ τους, τα LSTMs έχουν τη δυνατότητα να μαθαίνουν μακροπρόθεσμες εξαρτήσεις μεταξύ δεδομένων. Η χρήση τους ενδείκνυται για την αναγνώριση ανθρώπινης δράσης σε οπτικοακουστικό υλικό, αλλά με βάση μελέτες[53] έχουν αποδειχθεί χρήσιμα στην ανάλυση της χρονικής διάστασης των βίντεο σε συνδυασμό με τα CNNs.

3.2 Αναγνώριση Δράσης με Συνελικτικά Νευρωνικά Δίκτυα

Αφιερώνουμε ξεχωριστές ενότητες σε επίλυση προβλημάτων αναγνώρισης ανθρώπινης δράσης με τη χρήση CNNs, Transformers και Masked Autoencoders, εφόσον αποτελούν τις αρχιτεκτονικές τελευταίας τεχνολογίας στον τομέα της όρασης υπολογιστών. Συγκεκριμένα, τα συνελικτικά νευρωνικά δίκτυα, εφαρμόζοντας τη συνέλιξη στα οπτικά δεδομένα όπως περιγράφηκε στο θεωρητικό μέρος, αποτελούσαν μέχρι και την εμφάνιση των μετασχηματιστών, το ιδανικό μοντέλο για την επεξεργασία και εξαγωγή χαρακτηριστικών από εικόνα και βίντεο. Σε αυτήν την ενότητα θα παρουσιάσουμε αρχικά μία από τις κλασικές εφαρμογές των CNNs για την αναγνώριση δράσης και έπειτα ένα πιο εξελιγμένο συνελικτικό μοντέλο, το οποίο όμως χρειάστηκε να δανειστεί κάποιες ιδέες από τους μετασχηματιστές, για να ανταγωνιστεί την επίδοσή τους.

Το πρώτο παράδειγμα βασίζεται σε άρθρο[54], που δημοσιεύτηκε στο συνέδριο IAPR

του 2015. Το δίκτυο, που προτείνει, επιτυγχάνει αναγνώριση ανθρώπινης δράσης σε βίντεο βασισμένο στο σκελετό των ανθρώπων σε αυτά. Πιο συγκεκριμένα, αν θεωρήσουμε το ανθρώπινο σώμα ένα σύνολο κοκάλων και συνδέσμων, τότε μπορούμε να αναπαραστήσουμε την ανθρώπινη κίνηση με τις συντεταγμένες τους στον τρισδιάστατο χώρο[55]. Πλέον υπάρχουν οικονομικοί αισθητήρες βάθους, οι οποίοι σε συνδυασμό με αλγορίθμους εκτίμησης στάσης σκελετού σε πραγματικό χρόνο[56], παρέχουν με μεγάλη αξιοπιστία την τιμή των συντεταγμένων των αρθρώσεων.

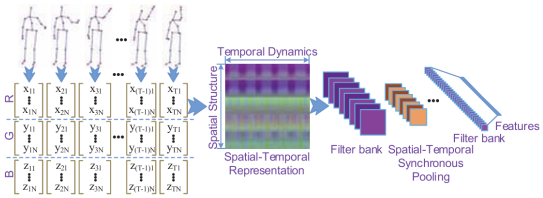
Η μέθοδος που ακολουθούν οι ερευνητές στο συγκεκριμένο άρθρο χωρίζεται σε δύο μέρη:

1. **Τη μεταφορά της πληροφορίας από τους αισθητήρες για τον ανθρώπινο σκελετό κατά την πάροδο του χρόνου σε εικόνα:**

   Σκοπός είναι η συμπύκνωση των δεδομένων από τους αισθητήρες για τις αρθρώσεις σε κάθε καρέ του βίντεο σε μία εικόνα, ώστε αυτή με τη σειρά της να εισαχθεί στο CNN. Οι συγγραφείς προτείνουν την εισαγωγή δεδομένων από πέντε βασικά μέλη του ανθρώπινου σκελετού, τα τέσσερα άκρα και τον κορμό, και την προβολή αυτών σε τρία ορθοκανονικά επίπεδα. Αυτές οι τιμές συνδέονται, δημιουργώντας τις γραμμές ενός πίνακα. Οι στήλες του πίνακα προκύπτουν από την εισαγωγή των δεδομένων αυτών σε κάθε καρέ. Τέλος, για να παραχθεί από τον πίνακα αυτό μία εικόνα, η πραγματικές τιμές των δεδομένων μετατρέπονται σε τιμές pixels και η τελική διάσταση της εικόνας τίθεται σε 60x60.

2. **Την επιλογή κατάλληλης αρχιτεκτονικής του συνελικτικού νευρωνικού δικτύου για την εξαγωγή των χαρακτηριστικών από την εικόνα:**

   Οι ερευνητές εκμεταλλεύονται την ικανότητα φιλτραρίσματος των συνελικτικών δικτύων και έτσι χρησιμοποιούν τέσσερα επίπεδα φίλτρων, συνέλιξης και συγκέντρωσης τοπικού μεγίστου. Με αυτόν τον τρόπο εξάγονται τα κατάλληλα χωροχρονικά χαρακτηριστικά, ικανά να διαφοροποιήσουν τις πράξεις στα βίντεο μεταξύ τους. Για να γίνει η κατηγοριοποίηση της δράσης σε κάθε βίντεο, τα εξαγόμενα από το CNN χαρακτηριστικά περνάνε από ένα δίκτυο εμπρόσθιας τροφοδότησης δύο πλήρως συνδεδεμένων επιπέδων.



Εικόνα 3.7: *Μία περιληπτική απεικόνιση της προτεινόμενης μεθόδου. Οι τρεις διαστάσεις ((R, G, B), στις οποίες αναλύονται τα δεδομένα ενώνονται μεταξύ τους ανά καρέ του βίντεο και στη συνέχεια οι τιμές κάθε καρέ τοποθετούνται χρονολογικά σε έναν πίνακα. Ο πίνακας ποσοτικοποιείται και κανονικοποιείται, με αποτέλεσμα τη δημιουργία μίας εικόνας, που εισάγεται στο ιεραρχικά δομημένο CNN. Έτσι τελικά προκύπτει η κατηγοριοποίηση του αναλυόμενου βίντεο.*

Αναμφίβολα, την τελευταία πενταετία η κυριαρχία του μετασχηματιστή έχει εξαπλωθεί από την επεξεργασία φυσικής γλώσσας και στην όραση υπολογιστών, καθιστώντας τα συνε-

λικτικά νευρωνικά δίκτυα απαρχαιωμένη τεχνική για την αναγνώριση ανθρώπινης δράσης. Ορισμένοι ερευνητές από το Facebook και το πανεπιστήμιο Berkeley, θέλοντας να αναδείξουν τη δύναμη της συνέλιξης στην ανάλυση οπτικών δεδομένων, κατάφεραν σε ένα άρθρο τους[57] να δημιουργήσουν ένα συνελικτικό νευρωνικό δίκτυο νέας γενιάς, το ConvNeXt που ξεπέρασε την επίδοση των μετασχηματιστών τελευταίας τεχνολογίας σε προβλήματα αναγνώρισης κίνησης, Αυτό το κατάφεραν χρησιμοποιώντας μερικές από τις τακτικές, που χρησιμοποιούνται από τους μετασχηματιστές.

Κατ' αρχάς, δοκίμασαν στρατηγικές εκπαίδευσης συνηθισμένες στους Transformers και αυτές πράγματι βελτίωσαν την επίδοση του αρχικού συνελικτικού μοντέλου ResNet[58] στο σύνολο δεδομένων από εικόνες ImageNet. Αύξησαν τις εποχές εκπαίδευσης από 90 σε 300, χρησιμοποίησαν τεχνικές επαύξησης των δεδομένων (Mixup[59], Cutmix[60], RandAugment[61] και Random Erasing[62]) σε συνδυασμό με τεχνικές κανονικοποίησης (Stochastic Depth και Label Smoothing[63]). Επίσης προτείνεται η χρήση μίας στοχαστικής εκδοχής του βελτιστοποιητή Adam, ο οποίος είναι ένας προσαρμοστικός στο μέγεθος των δεδομένων και στο σφάλμα της εξόδου αλγόριθμος διόρθωσης των παραμέτρων των νευρώνων, το AdamW[64].

Επίσης υιοθετούνται κάποιες μακροστρατηγικές από τους μετασχηματιστές. Αυτές είναι η αύξηση των μονάδων επεξεργασίας σε κάθε επίπεδο σε συνδυασμό με την εφαρμογή φίλτρων σε μη επικαλυπτόμενα παράθυρα της εικόνας. Ένα άλλο δυνατό χαρακτηριστικό των μετασχηματιστών είναι η δυνατότητά τους να βλέπουν ταυτόχρονα όλη την εικόνα ή μάλλον τα τμήματα, στα οποία αυτή χωρίζεται. Η αλλαγή που μπορεί να γίνει στα CNNs, προκειμένου να αποκτηθεί στο βαθμό που είναι εφικτό αυτή η δυνατότητα, είναι η αποφυγή χρήσης παραθύρων μικρών διαστάσεων στο συνελικτικό επίπεδο του μοντέλου. Γνώρισμα των Transformers αποτελεί και η εξαγωγή του βαθμού συσχέτισης ενός τμήματος των δεδομένων με κάθε άλλο, δηλαδή δεν γίνεται ταυτόχρονα ο υπολογισμός για εξαγωγή χαρακτηριστικών σε όλα τα τμήματα. Οι ερευνητές του άρθρου προσπάθησαν να μιμηθούν αυτό το γνώρισμα εφαρμόζοντας συνέλιξη κατά βάθος στα κανάλια, στα οποία χωρίζεται η εικόνα κατά την επεξεργασία της. Αυτό είναι ένα είδος συνέλιξης, στο οποίο δεν εφαρμόζεται ένα φίλτρο σε όλη την είσοδο, αλλά ξεχωριστό σε κάθε κανάλι της εισόδου, παίρνοντας έτσι στην έξοδο διαφορετικά φιλτραρισμένα τμήματα της εισόδου, όπως και στους μετασχηματιστές. Η κατά βάθος συνέλιξη έχει χρησιμοποιηθεί ήδη σε άλλα μοντέλα, όπως το MobileNet[65] και το Xception[66].

Τέλος, δοκιμάστηκαν και κάποιες αλλαγές σε λεπτομέρειες του μοντέλου. Συγκεκριμένα, η αντικατάσταση των συναρτήσεων ενεργοποίησης ReLU με συναρτήσεις GELU[67] σε συνδυασμό με τη μείωση της χρήσης των συναρτήσεων ενεργοποίησης, καθώς και η μείωση των επιπέδων κανονικοποίησης, σε μία προσπάθεια μίμησης των μετασχηματιστών. Άλλη μία αλλαγή στην αρχιτεκτονική του κλασικού ResNet, που έγινε στο ίδιο πλαίσιο του να μοιάσει στην αντίστοιχη των Transformers, είναι η πρόσθεση και άλλων επιπέδων μείωσης μεγέθους των δεδομένων, από τα οποία υπήρχε στο αρχικό δίκτυο μόνο ένα. Αυτή η μείωση δεδομένων στα συνελικτικά νευρωνικά δίκτυα επιτυγχάνεται με την εφαρμογή φίλτρων στα συνελικτικά επίπεδα με βήμα ίσο με δύο.

Με την υιοθέτηση όλων των παραπάνω τεχνικών, επιτυγχάνεται ανταγωνιστική επίδοση των συνελικτικών νευρωνικών δικτύων, σε σχέση με αυτή των μετασχηματιστών, όχι μόνο στην

αναγνώριση εικόνων, αλλά και στον εντοπισμό αντικειμένων και στο διαχωρισμό αντικειμένων σε εικόνες.

3.3 Αναγνώριση Δράσης με Μετασχηματιστές

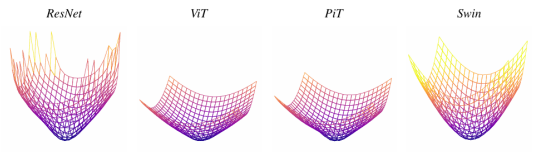
3.3.1 Οπτικός Μετασχηματιστής

Η επανάσταση στον τομέα της όρασης υπολογιστών ήρθε το 2020 με τη δημοσίευση από την ερευνητική ομάδα της Google ενός άρθρου[68] που χρησιμοποιούσε την αρχιτεκτονική των μετασχηματιστών για την αναγνώριση εικόνων και κατάφερε να πετύχει πιο ανταγωνιστική επίδοση από αυτή των, έως και εκείνη τη στιγμή επικρατούντων, CNNs. Βέβαια η αλήθεια είναι ότι η μεταφορά του μοντέλου των Transformers από το NLP στην όραση υπολογιστών δεν έγινε από τη μία μέρα στην άλλη και υπήρξε ένα μεταβατικό στάδιο, όπου χρησιμοποιούνταν πολυτροπικοί μετασχηματιστές για την επίλυση προβλημάτων που συνδύαζαν φυσική γλώσσα και εικόνες, όπως το VilBERT[69]. Σε αυτές τις εφαρμογές είχαν εντυπωσιακή επίδοση οι υβριδικοί μετασχηματιστές, εφόσον μπορούσαν με σχετικά μεγάλη ακρίβεια να απαντήσουν σε ερωτήσεις με τα σωστά τμήματα της εικόνας και αντίστροφα να αναγνωρίζουν αντικείμενα σε συγκεκριμένα τμήματα μίας εικόνας. Ωστόσο δεν υπήρχε ένα μοντέλο μετασχηματιστή αμιγώς για την επεξεργασία εικόνας.

Στο άρθρο του 2020 όμως, οι συγγραφείς προτείνουν το χωρισμό της εικόνας σε τμήματα μεγέθους 16x16 εικονοστοιχείων, τα οποία εισάγονται στη συνέχεια σε ένα μετασχηματιστή. Πριν την εισαγωγή τους στο δίκτυο, μετατρέπονται από σύνολο pixels σε διανύσματα, ώστε να μπορεί να τα επεξεργαστεί το μοντέλο, μέσω ενός γραμμικού μετασχηματισμού. Επίσης, στα διανύσματα που προκύπτουν προστίθενται οι ενσωματώσεις θέσης, με σκοπό να λάβει υπόψιν κατά τους υπολογισμούς του ο μετασχηματιστής σε ποιο σημείο της εικόνας ανήκει το κάθε τμήμα. Ακολουθεί η εισαγωγή των διανυσμάτων που αναπαριστούν τα διαφορετικά τμήματα της αρχικής εικόνας στο μετασχηματιστή, ο οποίος τα επεξεργάζεται όπως περιγράψαμε στο θεωρητικό μέρος. Το μοντέλο πετυχαίνει τα υψηλότερα ποσοστά ορθότητας στο πρόβλημα αναγνώρισης εικόνων στα σύνολα δεδομένων ImageNet, ImageNet-Real και CIFAR-100, ξεπερνώντας τα προηγούμενα συνελικτικά μοντέλα τελευταίας τεχνολογίας. Τέλος, οι ερευνητές έκανα πειράματα με τρεις εκδοχές του μοντέλου, που διέφεραν ως προς το πλήθος των επιπέδων του δικτύου και άρα το πλήθος των παραμέτρων, το ViT-Base, ViT-Large και ViT-Huge.

Σε αυτό το σημείο αξίζει να αναφερθούμε στις διαφορές μεταξύ μετασχηματιστών και συνελικτικών νευρωνικών δικτύων και τα σημεία όπου υπερτερεί το κάθε μοντέλο, όπως αναλύεται στο άρθρο[70], για να γίνει κατανοητή η μετάβαση από το ένα μοντέλο στο άλλο αλλά και η αξία συνδυασμού τους, όπως είδαμε στην προηγούμενη ενότητα το ConvNeXt. Πρώτα απ' όλα λόγω της δυνατότητας ύπαρξης πολλών κεφαλών προσοχής στους μετασχηματιστές, τα μοντέλα αυτά περιλαμβάνουν από την αρχή όλη την εικόνα, εφόσον επεξεργάζονται ταυτόχρονα όλα τα κομμάτια της, σε αντίθεση με τα συνελικτικά μοντέλα, που επεξεργάζονται παράθυρα της εισόδου στην αρχή και μόνο στο τέλος εξάγουν χαρακτηριστικά για όλη την εικόνα. Για αυτό το λόγο, οι μετασχηματιστές είναι πιο ανθεκτικοί σε αλλοιωμένα δεδομένα. Αντιθέτως, τα συνελικτικά δίκτυα δεν μπορούν να απομονώσουν το τροποποιημένο τμήμα της

εικόνας και να το μην το λάβουν υπόψιν στην εξαγωγή του αποτελέσματος, εφόσον περνάει η τροποποίηση από όλα τα επίπεδα του δικτύου. Ένα ακόμη προτέρημα των Transformers είναι ομαλότερη διαδρομή που οδηγεί στο ολικό ελάχιστο της συνάρτησης σφάλματος, το οποίο υπολογίστηκε από τους συγγραφείς του άρθρου μέσω της σύγκρισης των ιδιοτιμών των Εσσιανών πινάκων της συνάρτησης σφάλματος σε transformers και ResNet.



Εικόνα 3.8: *Η ομαλότερη εικόνα της γραφικής παράστασης της συνάρτησης σφάλματος στα μοντέλα μετασχηματιστών σε σχέση με το συνελικτικό μοντέλο ResNet.*

Από την άλλη μεριά, σίγουρα οι μετασχηματιστές απαιτούν μεγαλύτερο όγκο δεδομένων, εφόσον σε μικρά σύνολα δεδομένων υπερπροσαρμόζεται σε αυτά[71]. Επιπλέον, απαιτούν περισσότερη ενέργεια για να εκπαιδεύσουν τις πολλές παραμέτρους του μοντέλου, περισσότερο χρόνο και περισσότερους υπολογιστικούς πόρους. Αυτό είναι λογικό, γιατί υπολογίζουν τους συσχετισμούς μεταξύ ενός τμήματος της εικόνας με όλα τα άλλα τμήματα και χρειάζεται να αποθηκεύονται τα αποτελέσματα της προσοχής μεταξύ όλων των patches. Για αυτούς τους λόγους η εκπαίδευση μετασχηματιστών δεν είναι πολύ φιλική προς το περιβάλλον, σε μία εποχή, που είναι πολύ σημαντική η χρήση όχι ενεργειακά κοστοβόρων τεχνολογιών.

Ένα τελευταίο πολύ ενδιαφέρον συμπέρασμα του άρθρου είναι πως CNNs και Transformers μπορούν να λειτουργήσουν συμπληρωματικά. Αναφέρεται στο άρθρο πως οι μετασχηματιστές ειδικεύονται στα δεδομένα, τα οποία αναλύουν, γιατί εκπαιδεύονται σε μία συγκεκριμένη αλληλουχία τμημάτων εικόνας, ενώ τα συνελικτικά δίκτυα εντοπίζουν με την εφαρμογή των κατάλληλων φίλτρων στα δεδομένα κάθε φορά τα ίδια χαρακτηριστικά, όπως ακμές και υφές της εικόνας. Αντιστρόφως, τα συνελικτικά δίκτυα ειδικεύονται στα κανάλια R, G, B, στα οποία χωρίζονται οι εικόνες της εισόδου, αφού τα φίλτρα εφαρμόζονται σε κάθε κανάλι ξεχωριστά, σε αντίθεση με τους μετασχηματιστές που διαχειρίζονται κάθε κανάλι, κάθε τμήμα στο οποίο χωρίζεται η είσοδος δηλαδή, ισότιμα, υπολογίζοντας με τον ίδιο τρόπο την προσοχή σε όλα. Τέλος, βασισμένοι σε ανάλυση Fourier των χαρακτηριστικών, που εξάγουν τα δύο μοντέλα, οι συγγραφείς καταλήγουν στο συμπέρασμα πως οι μετασχηματιστές λειτουργούν ως φίλτρα διέλευσης χαμηλών συχνοτήτων, έχοντας τη δυνατότητα να εξάγουν σχήματα από τις εικόνες, ενώ τα CNNs λειτουργούν ως φίλτρα διέλευσης υψηλών συχνοτήτων, εντοπίζοντας ακμές και υφές στις εικόνες.

Το άρθρο ολοκληρώνεται με την πρόταση ενός υβριδικού δικτύου, που αποτελεί αλληλουχία συνελικτικών επιπέδων και επιπέδων με πολλαπλές κεφαλές προσοχής, το AlterNet. Αν και η ιδέα ότι μπορούν να λειτουργήσουν συμπληρωματικά η μία προς την άλλη οι δύο αρχιτεκτονικές είναι ενδιαφέρουσα και αφήνει πολύ χώρο για έρευνα, η αλήθεια είναι πως οι μετασχηματιστές παραμένουν το κυρίαρχο μοντέλο στο πεδίο της όρασης υπολογιστών[72]. Για αυτό στις επόμενες υποενότητες θα ασχοληθούμε με βελτιωμένες εκδοχές του ViT και με τη χρήση μετασχηματιστών στην αναγνώριση δράσης σε βίντεο, μιας και αυτό είναι το πρόβλημα που πραγματεύεται και αυτή η εργασία.

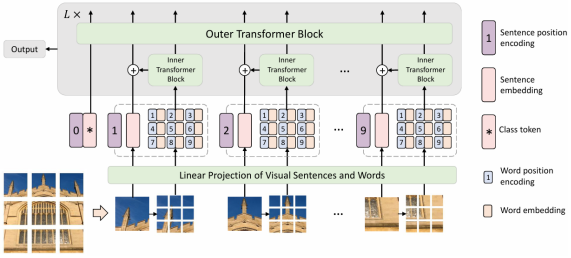
3.3.2 Εξελιγμένοι Οπτικοί Μετασχηματιστές

Η πρώτη παραλλαγή των Vision Transformers ήρθε από την ερευνητική ομάδα του Facebook με το άρθρο τους για το DEiT[73]. Το ενδιαφέρον με το συγκεκριμένο μοντέλο είναι πως καταφέρνει να πετύχει ανταγωνιστικές επιδόσεις στο πρόβλημα αναγνώρισης εικόνας με σχεδόν 300 φορές λιγότερα δεδομένα σε σχέση με το ViT. Συγκεκριμένα, το μοντέλο του άρθρου εκμεταλλεύεται τεχνικές που βρίσκουν εφαρμογή στα CNNs, προκειμένου να μη χρειαστεί να εκπαιδευθεί σε περισσότερα δεδομένα από αυτά που υπάρχουν στο σύνολο δεδομένων ImageNet. Αυτές είναι διάφορες τεχνικές επαύξησης των εικόνων, όπως οι Mixup, Cutmix, Auto-Augment και Rand-Augment σε συνδυασμό με πειραματισμό με διαφορετικούς βελτιστοποιητές διαβάθμισης κλίσης (SGD και AdamW). Μία ακόμη στρατηγική που χρησιμοποιείται για τη δημιουργία φαινομενικά νέων δεδομένων από τα ήδη υπάρχοντα είναι η αύξηση των διαστάσεων των εικόνων κατά τη διάρκεια της ρύθμισης με ακρίβεια του μοντέλου στα δεδομένα, αλλά με μία ειδική κοινωνικοποίηση των τιμών των εικονοστοιχείων των παραγόμενων εικόνων. Προκειμένου το μέτρο του διανύσματος της εικόνας πριν και μετά την αύξηση των διαστάσεων να μην αλλάξει σημαντικά, καθιστώντας το μοντέλο ανίκανο να αναγνωρίσει την ομοιότητα των δύο, εφαρμόζεται η τεχνική της δικυβικής παρεμβολής[74] με την απαιτούμενη κοινωνικοποίηση.

Η δεύτερη πρωτοτυπία του DEiT βασίζεται και αυτή στη λογική της εκπαίδευσης χωρίς μεγάλες απαιτήσεις σε δεδομένα, εφόσον χρησιμοποιεί μία τεχνική μεταφοράς γνώσης από ένα προεκπαιδευμένο συνελικτικό δίκτυο. Οι ερευνητές προσέθεσαν μία νέα μονάδα επεξεργασίας στην είσοδο του μετασχηματιστή, τη μονάδα απόσταξης γνώσης, που περιέχει κωδικοποιημένη την πρόβλεψη του δασκάλου συνελικτικού μοντέλου. Με αυτόν τον τρόπο, το DEiT λειτουργεί με ένα σχήμα δασκάλου-μαθητή και προσπαθεί όχι μόνο να ελαχιστοποιήσει την τιμή της συνάρτησης σφάλματος διασταυρωμένης εντροπίας των αποτελεσμάτων του[75], αλλά και να ελαχιστοποιεί την απόσταση μεταξύ της πρόβλεψής του και αυτής του δασκάλου.

Το μοντέλο TnT είναι μία άλλη παραλλαγή του οπτικού μετασχηματιστή, ο οποίος κατάφερε να ξεπεράσει την επίδοση του DEiT με παρόμοιο υπολογιστικό κόστος[76]. Βασίζεται στην ιδέα χρήσης ενός μετασχηματιστή για τον υπολογισμό των συσχετισμών μεταξύ των εικονοστοιχείων, από τα οποία αποτελούνται τα επιμέρους τμήματα που έχει χωριστεί η αρχική εικόνα, πριν την επεξεργασία των τμημάτων αυτών από το δεύτερο μετασχηματιστή. Με άλλα λόγια οι συγγραφείς του άρθρου αυτού καταφέρνουν να εκπαιδεύσουν το μοντέλο τους να εξάγει χαρακτηριστικά εντός κάθε τμήματος, στα οποία διαιρείται η εικόνα, χωριστά, αλλά και χαρακτηριστικά της συνολικής εικόνας, υπολογίζοντας την προσοχή μεταξύ των δεδομένων εισόδου σε επίπεδο pixels και σε επίπεδο κομματιών, που απαρτίζουν την εικόνα. Το ακόλουθο σχήμα αναπαριστά τον τρόπο που εισάγεται ο ένας μετασχηματιστής εντός του άλλου:

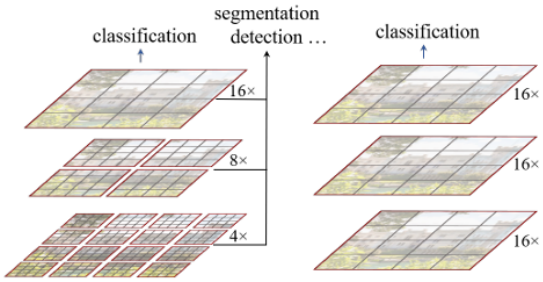
Κάποιοι άλλοι ερευνητές το 2021 εντόπισαν το πρόβλημα της δυνατότητας κλιμάκωσης των οπτικών μετασχηματιστών σε εικόνες υψηλής ανάλυσης, καθώς, αν ακολουθηθεί η τακτική τμηματοποίησης της εικόνας εισόδου σε κομμάτια διαστάσεων 16x16 pixels, τότε ο αριθμός αυτών των τμημάτων σε περιπτώσεις εικόνων HD αυξάνεται δραματικά και απαιτούνται πολλοί πόροι για την επεξεργασία τους. Επίσης τονίζεται πως σε προβλήματα όπως



Εικόνα 3.9: *Παρουσιάζεται η αρχιτεκτονική του TnT. Όπως φαίνεται η αρχική εικόνα χωρίζεται σε τμήματα, που αντιστοιχούν σε προτάσεις ενός κειμένου στο πεδίο του NLP, και κάθε μία από αυτές σε επιμέρους κομμάτια, τις λέξεις κάθε πρότασης. Αφού κάθε τμήμα προβληθεί γραμμικά ως ένα διάνυσμα, τα διανύσματα των τμημάτων κάθε patch εισάγονται σε έναν εσωτερικό μετασχηματιστή. Στη συνέχεια, η έξοδος του εσωτερικού transformer για τα κομμάτια ενός τμήματος προστίθεται στην ενσωμάτωση θέσης και στο διάνυσμα του τμήματος αυτού, με σκοπό να εισαχθούν στο δεύτερο μετασχηματιστή. Αυτός μετά από επεξεργασία των δεδομένων εισόδου, προβλέπει την κατηγορία, στην οποία ανήκει η αρχική εικόνα.*

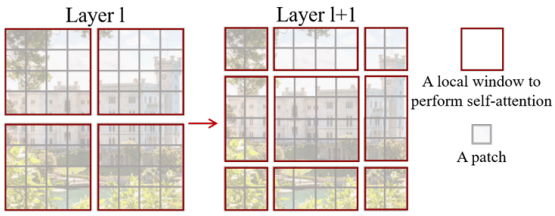
αυτό του διαχωρισμού μίας εικόνας σε αντικείμενα διαφορετικών κατηγοριών απαιτείται η ανάλυση κάθε εικονοστοιχείου, για να επιλυθεί με μεγάλη ακρίβεια, οπότε οι μετασχηματιστές δεν μπορούν να αποφύγουν την ενδελεχή επεξεργασία εικόνων υψηλής ευκρίνειας απλώς χωρίζοντάς τες σε τμήματα που καλύπτουν μεγαλύτερη επιφάνεια.

Σε αυτό το πρόβλημα απαντούν οι ερευνητές με το μοντέλο Swin Transformer[77], το οποίο βασίζεται στην ιδέα της επεξεργασίας μίας εικόνας ανά παράθυρο από τα συνελικτικά δίκτυα. Το δίκτυο που προτείνουν παρουσιάζει μία ιεραρχική δομή, καθώς είναι οργανωμένο σε στάδια με βάση το μέγεθος των τμημάτων, στα οποία χωρίζει την εικόνα. Αρχικά, διαιρείται σε μικρών διαστάσεων τμήματα η εικόνα, αλλά για να μην αυξηθεί το υπολογιστικό κόστος τετραγωνικά σε σχέση με το πλήθος των τμημάτων, που προκύπτουν, υπολογίζεται η προσοχή μόνο εντός συγκεκριμένων παραθύρων, που περιλαμβάνουν γειτονικά patches. Αφού με αυτόν τον τρόπο έχουν υπολογιστεί οι συσχετισμοί σε επίπεδο pixels και έχουν εξαχθεί τα χαρακτηριστικά υψηλής ανάλυσης από την εικόνα, στα επόμενα στάδια συγχωνεύονται τμήματα μεταξύ τους, δημιουργώντας τμήματα μεγαλύτερων διαστάσεων. Και σε αυτά τα στάδια υπολογίζεται η προσοχή μόνο μεταξύ γειτονικών τμημάτων εντός ενός παραθύρου. Έτσι, όσο βαθαίνει το δίκτυο, μειώνεται και η ανάλυση της εικόνας, η οποία γίνεται αντικείμενο επεξεργασίας και εξάγονται πιο γενικά χαρακτηριστικά αυτής.



Εικόνα 3.10: *Συγκρίνεται η αρχιτεκτονική στα διάφορα στάδια των Swin Transformers στα αριστερά και των ViTs στα δεξιά. Είναι εμφανές πως οι Swin Transformers αντιγράφουν την τακτική των CNNs να επεξεργάζονται αρχικά μικρότερα τμήματα των δεδομένων και όσο βαθαίνει το δίκτυο να συγχωνεύουν τα τμήματα αυτά.*

Ενδιαφέρον στο συγκεκριμένο άρθρο, το οποίο δίνει και το όνομά του στο μοντέλο (Μετασχηματιστής Μετατοπιζόμενων Παραθύρων) αποτελεί το γεγονός πως σε κάθε στάδιο τα παράθυρα, μόνο εντός των οποίων υπολογίζονται οι συσχετισμοί των τμημάτων, μετατοπίζονται. Ειδικότερα, στα διάφορα επίπεδα, που συναποτελούν κάθε στάδιο του μοντέλου, αλλάζει η επιφάνεια της εικόνας και επομένως τα τμήματά της, που περιέχονται σε ένα παράθυρο, ώστε να μπορέσουν να υπολογιστούν οι συσχετισμού μεταξύ τμημάτων, που δεν είχαν υπολογιστεί νωρίτερα χωρίς ταυτόχρονα να ξεφεύγει αυτός ο υπολογισμός από το επίπεδο της γειτονιάς ενός τμήματος της εικόνας.



Εικόνα 3.11: *Ένα παράδειγμα της αλλαγής της μορφής και της μετατόπισης των παραθύρων, εντός των οποίων υπολογίζεται η προσοχή μεταξύ των εμπεριεχόμενων patches.*

Δεδομένου ότι οι μετασχηματιστές είναι μοντέλα βαθιάς μάθησης με μεγάλες απαιτήσεις σε υπολογιστική δύναμη, υπάρχουν πολλές εργασίες που επικεντρώνονται στην αντιμετώπιση αυτού του μειονεκτήματος[78, 79, 80]. Θα παρουσιάζουμε μία συγκεκριμένη που έχει πολύ ενδιαφέρον λόγω του απλού αλγορίθμου που χρησιμοποιεί, για να μειώσει τον αριθμό μονάδων επεξεργασίας του μετασχηματιστή. Πρόκειται για το ToMe[81], μία τεχνική που συγχωνεύει μέχρι ένα συγκεκριμένο βαθμό, που ορίζεται από το χρήστη του μοντέλου, τις μονάδες επεξεργασίας στις οποίες αναλύεται μία εικόνα. Η τεχνική αυτή λαμβάνει χώρο μετά το επίπεδο υπολογισμού της προσοχής στο μετασχηματιστή, μέσω της οποίας έχει δημιουργηθεί μία μετρική ομοιότητας ανάμεσα στις μονάδες επεξεργασίας, και τις συγχωνεύει με βάση έναν αλγόριθμο ήπιου ταιριάσματος διμερούς γραφήματος.

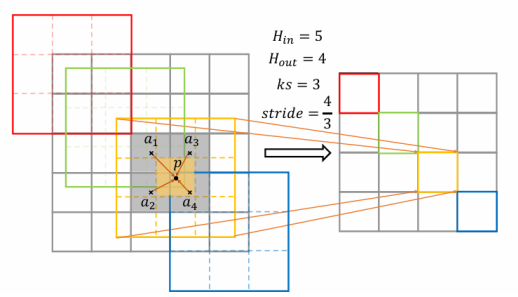
Τα βήματα που ακολουθούνται είναι:

1. Χωρίζονται τυχαία οι μονάδες επεξεργασίας σε δύο σύνολα A και B ίδιου μεγέθους.
2. Φέρνουμε μία ακμή από κάθε στοιχείο του A στο πιο όμοιό του στοιχείο από το B.
3. Κρατάμε τις r ακμές ανάμεσα στα πιο όμοια στοιχεία.
4. Συγχωνεύουμε τις μονάδες επεξεργασίας, που είναι συνδεδεμένες μεταξύ τους, διατηρώντας το μέσο όρο των χαρακτηριστικών τους.
5. Ενώνουμε ξανά όλες τις εναπομείνασες μονάδες επεξεργασίας σε ένα σύνολο.

Με αυτόν τον τρόπο, μειώνεται όσο κρίνεται απαραίτητο η ανάλυση της εικόνας εισόδου στο δίκτυο εμπρόσθιας τροφοδότησης, ώστε να επισπευσθεί αυτό το στάδιο επεξεργασίας του μετασχηματιστή. Συνολικά, εφαρμόζοντας αυτήν την τεχνική σε blocks μετασχηματιστών, μειώνεται ο χρόνος εκπαίδευσης του μοντέλου χωρίς σημαντικές απώλειες στην ορθότητα των αποτελεσμάτων του.

Ένα ακόμη μοντέλο μετασχηματιστή, που προσπαθεί να μειώσει το πλήθος μονάδων επεξεργασίας, που πρέπει να εισαχθούν σε κάθε transformer block, αποτελεί το FDViT[82]. Το δίκτυο αυτό δανείζεται την ιδέα της μείωσης μεγέθους των δεδομένων εισόδου από την ιεραρχική αρχιτεκτονική των συνελικτικών νευρωνικών δικτύων, τα οποία μειώνουν μέσω της συνέλιξης τη διάσταση των δεδομένων. Έτσι και ο μετασχηματιστής αυτού του άρθρου προσπαθεί να μειώσει το μέγεθος των δεδομένων που μεταφέρονται από transformer block σε transformer block μέσω εφαρμογής φίλτρων, που συγχωνεύουν με συγκεκριμένο τρόπο τις μονάδες επεξεργασίας μεταξύ τους. Για να γίνει με βέλτιστο τρόπο αυτό, έχει εισαχθεί ένας αποκρύπτων αυτοκωδικοποιητής που εκπαιδεύει το επίπεδο μείωσης μεγέθους των δεδομένων να διατηρεί τις μονάδες επεξεργασίας με την περισσότερη πληροφορία. Αυτό το επιτυγχάνει, δοκιμάζοντας την απόκρυψη συνδυασμών διαφόρων tokens και στη συνέχεια προσπαθώντας να ανακατασκευάσει την αρχική εικόνα. Τα tokens που είναι τα πιο κρίσιμα για την ανακατασκευή της εικόνας μαθαίνει να τα διατηρεί, ενώ τα υπόλοιπα να τα συγχωνεύει με τα πρώτα.

Μία ακόμη πρωτοτυπία του FDViT είναι η εφαρμογή φίλτρων στα επίπεδα μείωσης μεγέθους με βήμα όχι απαραίτητα ακέραιο αριθμό. Με αυτόν τον τρόπο, μπορούν να παραχθούν νέοι χάρτες χαρακτηριστικών από τους παλιούς με τη διαστατικότητα, την οποία κρίνει ο αποκρύπτων αυτοκωδικοποιητής ότι είναι απαραίτητη, για να περικλείεται όλη η πληροφορία του αρχικού. Παρατίθεται μία εικόνα ως παράδειγμα:



Εικόνα 3.12: *Στο αριστερό κομμάτι της εικόνας βρίσκεται ο χάρτης χαρακτηριστικών αρχικών διαστάσεων και δεξιά ο χάρτης χαρακτηριστικών που θέλουμε να παραχθεί. Τα χρωματιστά κουτάκια στον τελικό χάρτη προκύπτουν από την εφαρμογή των φίλτρων αντίστοιχου χρώματος στον αρχικό. Παρατηρείται ότι το βήμα ανάμεσα σε αυτά τα φίλτρα δεν είναι ακέραιο, αλλά ισούται με 4/3. Αυτός ο λόγος υπολογίζεται, αν υπολογιστεί ο λόγος της απόστασης του κέντρου του κίτρινου τετραγώνου p από τα κέντρα των γειτονικών τετραγώνων a_1 και a_3. Υπάρχουν πολλοί τρόποι που αυτά τα φίλτρα μπορούν να συγκεντρώσουν την πληροφορία εντός τους, όπως συγκέντρωση τοπικού μέσου, συγκέντρωση τοπικού μεγίστου, διγραμμική παρεμβολή[3] κ.α.*

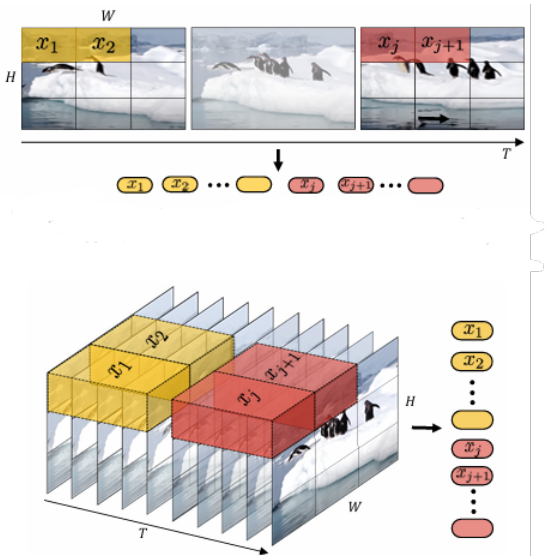
3.3.3 Αναγνώριση Δράσης σε Βίντεο με Μετασχηματιστές

Αν και αυτή η παρουσίαση της εξέλιξης των μοντέλων των οπτικών μετασχηματιστών παρουσιάζει ενδιαφέρον, το θέμα της διπλωματικής είναι η αναγνώριση ανθρώπινης δράσης σε βίντεο, οπότε θα περάσουμε στο πως έχουν καταφέρει οι μετασχηματιστές να αντεπεξέλθουν σε αυτό το πρόβλημα.

Ένα από τα πρώτα μοντέλα που βασίστηκε στη λειτουργία των Vision Transformers για

κατηγοριοποίηση δράσης σε βίντεο είναι το ViViT[83]. Στο άρθρο παρουσιάζονται τέσσερις διαφορετικές εκδοχές του μοντέλου, το οποίο στην περίπτωση των βίντεο δεν έχει να υπολογίσει την προσοχή μόνο στο χώρο, σε κάθε καρέ του βίντεο, αλλά και την προσοχή χρονικά, δηλαδή την προσοχή μεταξύ των καρέ. Οι εκδοχές αυτές διαφέρουν ως προς τον αριθμό μετασχηματιστών που χρησιμοποιούν και επομένως ως προς την πολυπλοκότητα και την αποδοτικότητά τους.

Προτού περάσουμε στην περιγραφή των εκδοχών του μοντέλου και των διαφορών τους, αναφέρουμε δύο τρόπους τμηματοποίησης των βίντεο και ενσωμάτωσης των τμημάτων αυτών. Από τη μία μεριά είναι ο διαχωρισμός του βίντεο σε καρέ και η δημιουργία τμημάτων σε κάθε καρέ με τον ίδιο τρόπο που δημιουργούνται στο ViT. Προφανώς με αυτόν τον τρόπο προκύπτει μεγάλος αριθμός μονάδων επεξεργασίας για το μετασχηματιστή, ο οποίος επιβαρύνει το χρόνο υπολογισμού και τη μνήμη, στην οποία πρέπει να αποθηκευτούν όλα τα δεδομένα. Για αυτό προτείνουν και την σωληνοειδή τμηματοποίηση των βίντεο, κατά την οποία δεν δημιουργούνται απλώς τμήματα σε κάθε καρέ, αλλά σωλήνες αποτελούμενοι από τμήματα στην ίδια θέση σε συνεχόμενα καρέ. Αναλόγως πόσο αναλυτικοί θέλουν οι ερευνητές να είναι οι υπολογισμοί τους , επιλέγουν το πλήθος καρέ, τα οποία θα περιλαμβάνουν οι μη αλληλοκαλυπτόμενοι σωλήνες. Και στην περίπτωση αυτή προκύπτουν από την τμηματοποίηση του βίντεο πληροφορίες για τα δεδομένα εισόδου τόσο σε χρονικό όσο και σε χωρικό επίπεδο.



Εικόνα 3.13: *Στο πάνω μέρος φαίνεται η κλασική τμηματοποίηση ανά καρέ και στο κάτω η σωληνοειδής τμηματοποίηση και ενσωμάτωση των δεδομένων από το βίντεο στο μουτέλο.*

1. Η πρώτη εκδοχή του μοντέλου ουσιαστικά αποτελεί μία αντιμετώπιση του προβλήματος με χρήση ωμής βίας. Ειδικότερα, χρησιμοποιείται ένας μετασχηματιστής, που δέχεται ως είσοδο τις ενσωματώσεις όλων των χωροχρονικών τμημάτων, στα οποία έχει διαιρεθεί το βίντεο. Με αυτόν τον τρόπο υπολογίζεται η προσοχή μεταξύ όλων των τμημάτων. Αυτός δεν είναι ο βέλτιστος τρόπος επεξεργασίας της εισόδου ιδιαίτερα, όσον αφορά το υπολογιστικό κόστος που απαιτείται.

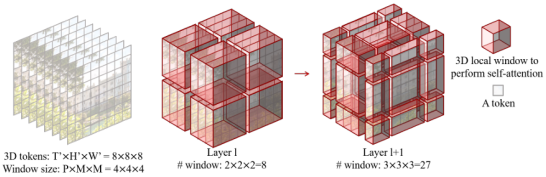
2. Στη δεύτερη εκδοχή του μοντέλου χρησιμοποιούνται δύο μετασχηματιστές, εφόσον υ-

πολογίζεται η προσοχή σε επίπεδο χωρικό αρχικά και σε χρονικό στη συνέχεια. Αφού έχει τελειώσει η επεξεργασία στα χωρικά τμήματα, τα αποτελέσματά της ενώνονται με τις ενσωματώσεις θέσεις των χρονικών τμημάτων και γίνονται αντικείμενο επεξεργασίας από το δεύτερο μετασχηματιστή. Η εκδοχή αυτή δανείζεται την ιδέα της ύστερης συγχώνευσης από άλλα papers[84]. Αυτή η εκδοχή είναι λιγότερη απαιτητική υπολογιστικά.

3. Η τρίτη εκδοχή έχει τον ίδιο αριθμό παραμέτρων με την πρώτη, αλλά επεξεργάζεται πολύ πιο αποδοτικά τα δεδομένα, αφού με έναν μετασχηματιστή υπολογίζει την παραγοντοποιημένη αυτο-προσοχή[85], δηλαδή την προσοχή σε δύο βήματα, πρώτα το χωρικό και μετά το χρονικό. Παρουσιάζει με αυτήν την τεχνική την ίδια υπολογιστική πολυπλοκότητα με τη δεύτερη εκδοχή, έχοντας όμως καλύτερη επίδοση.

4. Η τέταρτη εκδοχή παραγοντοποιεί επίσης την προσοχή, αλλά στο επίπεδο του εσωτερικού γινομένου μέσα στην κεφαλή υπολογισμού της προσοχής. Με αυτόν τον τρόπο ο ένας μετασχηματιστής που χρησιμοποιείται υπολογίζει δύο διαφορετικά βάρη για κάθε μονάδα επεξεργασίας στη διάσταση του χώρου και στη διάσταση του χρόνου. Έτσι, επιτυγχάνεται καλύτερη επίδοση στα αποτελέσματα του μοντέλου και μεγαλύτερος έλεγχος σε κάθε κεφαλή υπολογισμού προσοχής.

Όπως το προηγούμενο μοντέλο εμπνεύστηκε από το ViT, εμφανίστηκε ένα μοντέλο για το πρόβλημα της αναγνώρισης δράσης σε βίντεο εμπνευσμένο από το Swin Transformer, το Video Swin Transformer[86]. Στο άρθρο αυτό το καρέ κάθε βίντεο τμηματοποιείται όπως και στο Swin Transformer, δηλαδή αρχικά δημιουργούνται patches μικρών διαστάσεων, τα οποία στα βαθύτερα στάδια συγχωνεύονται μεταξύ τους, ακολουθώντας τη λογική της συνέλιξης των CNNs. Η κύρια διαφορά σε σχέση με το αντίστοιχο μοντέλο που επεξεργάζεται εικόνες είναι πως στα εσωτερικά επίπεδα κάθε σταδίου, τα παράθυρα, εντός των οποίων γίνονται οι υπολογισμοί μεταξύ των περικλειόμενων τμημάτων, είναι τρισδιάστατα και όχι δισδιάστατα. Με αυτόν τον τρόπο το μοντέλο καταφέρνει να υπολογίσει όχι μόνο το συσχετισμό του τμήματος ενός καρέ με τα γειτονικά του στη διάσταση του χώρου, αλλά και να υπολογιστούν οι συσχετισμοί του με τα τμήματα της χωρικής του γειτονιάς.



Εικόνα 3.14: *Η εφαρμογή των μετατοπιζόμενων τρισδιάστατων παραθύρων σε δύο διαφορετικά επίπεδα του ίδιου σταδίου του Video Swin Transformer.*

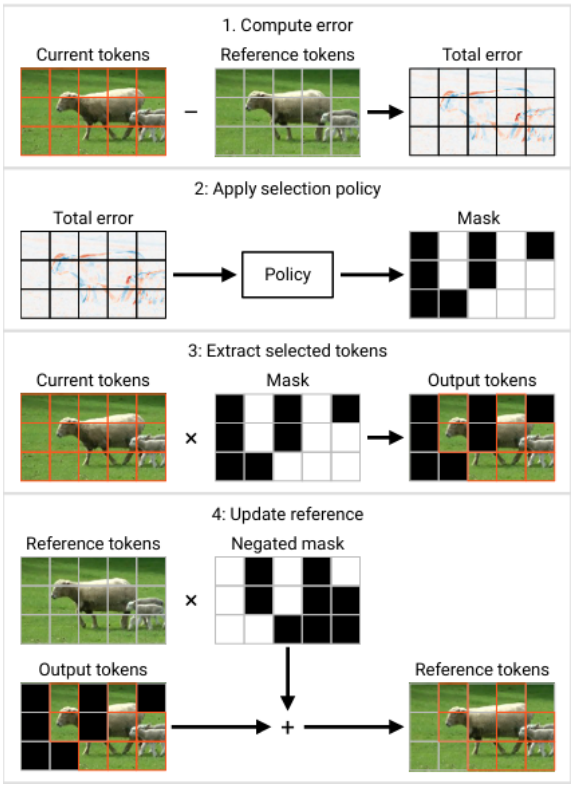
Η μεγαλύτερη πρόκληση που έχουν να αντιμετωπίσουν στο πεδίο της αναγνώρισης δράσης σε βίντεο οι μεταχηματιστές είναι ο τεράστιος όγκος δεδομένων, που πρέπει να κωδικοποιήσουν και να επεξεργαστούν, και οι τεράστιες απαιτήσεις σε υπολογιστικό κόστος και χρόνο που επιφέρει. Στο άρθρο[87] οι ερευνητές εισηγούνται ένα νέο μοντέλο οπτικού μετασχηματιστή, που βασίζεται στην τεχνική STTS για τη μείωση των δεδομένων, που τελικά θα γίνουν

αντικείμενο επεξεργασίας. Σε πρώτη φάση, όλα τα καρέ του βίντεο υπό επεξεργασία, περνάνε από ένα ρηχό συνελικτικό νευρωνικό δίκτυο, από το οποίο εξάγονται κάποια χαρακτηριστικά των καρέ, όσο αφορά την κίνηση που εμφανίζεται σε αυτά και το πόση πληροφορία περιέχουν σχετικά με αυτήν. Έπειτα, μέσω της εφαρμογής μίας συνάρτησης βαθμολόγησης στα εξαγόμενα χαρακτηριστικά, κατατάσσονται τα καρέ του βίντεο από αυτό, που κρίνεται πως περιέχει την περισσότερη πληροφορία, στο καρέ, που περιέχει τη λιγότερη. Με βάση την υπερπαράμετρο k, τα k πιο πλούσια σε πληροφορία καρέ περνάνε στο επόμενο βήμα του αλγορίθμου, ενώ τα υπόλοιπα αποκρύπτονται από το μοντέλο.

Στη συνέχεια, στο επίπεδο κάθε καρέ ξεχωριστά γίνεται ξανά μία επιλογή των περιοχών του καρέ, τα οποία κρίνονται τα πιο σημαντικά για την αναγνώριση δράσης στο βίντεο. Ακολουθείται μία διαδικασία ταξινόμησης των τμημάτων, στα οποία χωρίζεται το κάθε καρέ, από αυτά που έχουν επιλεχθεί, με βάση τις μικροδιαφορές που παρουσιάζουν σε σχέση με τα γειτονικά τους, οι οποίες συνεπάγονται περισσότερη πληροφορία. Ο αλγόριθμος STTS κρατάει τις k πιο κατατοπιστικές περιοχές των καρέ, οι οποίες μπορεί να αποτελούνται από παραπάνω από ένα patches και αποκρύπτει τα υπόλοιπα. Τελικά, στο μετασχηματιστή, που καλείται να εντοπίσει συσχετισμούς χρονικούς και χωρικούς μεταξύ των δεδομένων του βίντεο, εισάγεται ένα μόνο ποσοστό των αρχικών δεδομένων, το οποίο επισπεύδει την εκπαίδευση του μοντέλου και δεν χρειάζεται πολλούς πόρους. Το αξιοσημείωτο με το μοντέλο STTS είναι πως καταφέρνει με τον παραπάνω έξυπνο τρόπο να μειώσει κατά 10% το υπολογιστικό κόστος χάνοντας μόνο 1% σε ορθότητα προβλέψεων.

Το μοντέλο των Eventful Transformers αποτελεί και αυτό μία προσπάθεια των ερευνητών στο πλαίσιο της μείωσης του υπολογιστικού κόστους. Στο άρθρο[88] περιγράφεται η ιδέα της αγνόησης της περιττής χρονικής πληροφορίας, που παρατηρείται στα καρέ, δηλαδή τα τμήματα της εικόνας που δεν αλλάζουν κατά τη διάρκεια του βίντεο. Αφού το κάθε καρέ έχει τμηματοποιηθεί και έχουν εισαχθεί όλα τα τμήματα του πρώτου καρέ σε ένα transformer block, για τα επόμενα καρέ δεν υπολογίζονται πάλι οι ενσωματώσεις όλων των τμημάτων τους, αλλά μόνο αυτών, στα οποία έχει παρατηρηθεί αλλαγή σε σχέση με το προηγούμενο καρέ. Ουσιαστικά αποκρύπτονται από τον κωδικοποιητή του μετασχηματιστή, που αναλαμβάνει την επεξεργασία των επόμενων καρέ, όσα τμήματα των καρέ δεν έχουν αλλάξει και συνεπώς δεν χρειάζεται εκ νέου να υπολογιστεί η προσοχή μεταξύ αυτών και των υπολοίπων. Πιο παραστατικά απεικονίζεται παρακάτω η μέθοδος του άρθρου:

Ανάλογα με την πολιτική, αυστηρή ή όχι, που θέλουμε να ακολουθήσουμε κατά την επιλογή των τμημάτων του τρέχοντος καρέ με νέα πληροφορία, μπορούμε να ελέγξουμε κατά πόσο θέλουμε να επιβαρύνουμε υπολογιστικά το μοντέλο μας με επεξεργασία μεγάλου ποσοστού των τμημάτων κάθε καρέ. Έτσι επιτυγχάνεται η επιθυμητή ισορροπία ανάμεσα σε μείωση πολυπλοκότητας και ελάττωση ορθότητας πρόβλεψης αναγνώρισης της δράσης. Ενδεικτικά, οι συγγραφείς του άρθρου διατηρώντας σε κάθε καρέ περίπου το 25% των τμημάτων, στα οποία παρατηρήθηκε κάποια αλλαγή σε σχέση με το αμέσως προηγούμενο, καταφέρνουν να μειώσουν 2.4 φορές το απαιτούμενο υπολογιστικό κόστος για την επεξεργασία ενός βίντεο με απώλεια μόνο 1.62% στην ορθότητα των αποτελεσμάτων του μοντέλου.



Εικόνα 3.15: *Στο 1 υπολογίζεται το σφάλμα μεταξύ του καρέ αναφοράς, το οποίο είναι το πρώτο του βίντεο στην αρχή, και του τρέχοντος καρέ υπό επεξεργασία. Στο 2 έχοντας εντοπίσει τα τμήματα του καρέ, στα οποία έχει παρατηρηθεί κίνηση, με βάση την πολιτική που ορίζει πόσο μεγάλη πρέπει να είναι η διαφορά μεταξύ των διαδοχικών τμημάτων ίδιας θέσης, ώστε να θεωρηθεί αμελητέα, παράγεται ένας χάρτης απόκρυψης με χρήση μασκών των τμημάτων που δεν αλλάζουν σημαντικά. Στο 3 εφαρμόζεται αυτός ο χάρτης μασκών στο τρέχον καρέ και λαμβάνονται τα τμήματα του καρέ, τα οποία περιέχουν πληροφορία για την κίνηση σε αυτό. Τέλος στο 4 εφαρμόζεται ο συμπληρωματικός χάρτης μάσκας με το καρέ αναφοράς, για να παραχθούν τα τμήματα χωρίς νέα πληροφορία, και συγχωνεύεται το αποτέλεσμα με το αποτέλεσμα του 3 και προκύπτει η είσοδος στο μετασχηματιστή.*

3.4 Αναγνώριση Δράσης σε Βίντεο με Αποκρύπτοντες Αυτοκωδικοποιητές

Περνάμε από το πεδίο της επιβλεπόμενης μάθησης στο πεδίο της αυτο-επιβλεπόμενης μάθησης για την εκπαίδευση μοντέλων, που θα είναι ικανά να αναγνωρίσουν ανθρώπινη δράση σε βίντεο. Στις υποενότητες που ακολουθούν περιγράφονται τρία μοντέλα αποκρύπτοντων αυτοκωδικοποιητών τελευταίας τεχνολογίας, οι οποίοι βασίζονται στην αυτο-επιβλεπόμενη μάθηση και αποτελούν τα μοντέλα, με τα οποία έχουν γίνει τα πειράματα της διπλωματικής.

3.4.1 VideoMAE

Το VideoMAE[89] αποτελεί τη μεταφορά της ιδέας του αποκρύπτοντος αυτοκωδικοποιητή από εικόνες σε βίντεο. Οι συγγραφείς του άρθρου, αν και αναγνωρίζουν την επανάσταση που έφερε στην όραση υπολογιστών η χρήση των μετασχηματιστών, εντοπίζουν το πρόβλημα,

που σχετίζεται με την ανάγκη των μοντέλων αυτών για τεράστιο όγκο δεδομένων[90], για να εκπαιδευθούν υπό επίβλεψη. Αυτό σε συνδυασμό με το γεγονός ότι δεν υπάρχουν τόσα διαθέσιμα δεδομένα βίντεο όσα εικόνων, οδήγησαν τους ερευνητές στον πειραματισμό με αποκρύπτοντες αυτοκωδικοποιητές, που είναι μοντέλα αυτο-επιβλεπόμενης μάθησης, στο πεδίο αναγνώρισης δράσης σε βίντεο.

Ο τρόπος που προετοιμάζονται τα βίντεο για επεξεργασία από το μοντέλο είναι ο εξής:

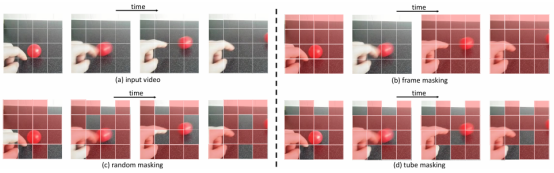
- Το βίντεο χωρίζεται σε καρέ, καθένα από τα οποία στη συνέχεια χωρίζεται σε τμήματα. Φυσικά κάθε τμήμα αποτελείται από τρία κανάλια, εφόσον το χρώμα των εικονοστοιχείων που τα αποτελούν αναλύεται σε κόκκινο, πράσινο και μπλε χρώμα.

- Μεταφέρονται τα τμήματα από τις δύο στις τρεις διαστάσεις, ενώνοντας τα τμήματα ίδιας θέσης δύο συνεχόμενων καρέ. Υπολογίζονται οι ενσωματώσεις των τρισδιάστατων tokens και, πριν εισαχθούν στον κωδικοποιητή του μοντέλου, ενώνονται με τις ενσωματώσεις θέσης.

Τα δύο κυριότερα χαρακτηριστικά των βίντεο, που τα διαφοροποιούν ως προς την επεξεργασία τους από τις εικόνες, είναι η περιττή πληροφορία που υπάρχει σε μεγάλο κομμάτι των καρέ του βίντεο, εφόσον πολλά pixels δεν αλλάζουν τιμή με την πάροδο του χρόνου, και ο συσχετισμός μεταξύ τμημάτων από χρονικά γειτονικά καρέ. Το πρώτο χαρακτηριστικό ενέχει τον πρόβλημα επιβάρυνσης του μοντέλου μέσω της επεξεργασίας παραπάνω δεδομένων, που δεν έχει κάποιο όφελος ως προς την απόδοση του μοντέλου. Το δεύτερο μπορεί να οδηγήσει το μοντέλο στην εκμάθηση χρονικών συσχετισμών χαμηλού επιπέδου μεταξύ pixels στην προσπάθειά του να γεμίσει τα κρυμμένα από αυτό κενά στο στάδιο της εκπαίδευσης και όχι την εξαγωγή συνολικών χαρακτηριστικών του βίντεο, που θα συμβάλλουν στην ορθή πρόβλεψη του αποτελέσματος.

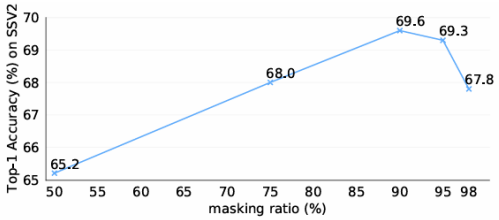
Η πρόταση του άρθρου, για να ξεπεραστούν οι δύο αυτοί κίνδυνοι, είναι αφενός η απόκρυψη ενός πολύ υψηλού ποσοστού των τρισιδιάστατων τμημάτων του βίντεο, της τάξης του 90%. με αυτόν τον τρόπο αποφεύγονται οι περιττοί υπολογισμοί που αφορούν τμήματα των καρέ χωρίς σημαντική πληροφορία για την αναγνώριση δράσης στο βίντεο. Αφετέρου, δεν γίνεται απόκρυψη των τμημάτων των καρέ με τυχαίο τρόπο, αλλά αποκρύπτονται ολόκληροι σωλήνες αποτελούμενοι από τμήματα διαδοχικών καρέ στην ίδια θέση. Έτσι, δεν αναλώνεται ο μετασχηματιστής στην εύρεση συσχετισμών μεταξύ τμημάτων ενός καρέ με τα γειτονικά του στη διάσταση του χώρου και στη διάσταση του χρόνου, ώστε να μπορέσει να συμπληρώσει τα κενά των γειτόνων του αυτών, που έχουν αποκρυφθεί από το μοντέλο. Αντιθέτως, επικεντρώνεται στην εύρεση χαρακτηριστικών υψηλού επιπέδου, που αφορούν όλο το βίντεο και βοηθούν περισσότερο στην κατηγοριοποίηση της δράσης που απεικονίζεται σε αυτό. Μετά από αυτήν την τεχνική απόκρυψης των δεδομένων, ενσωματώνονται σε διανύσματα και εισάγονται στο μοντέλο ViT, τη ραχοκοκαλιά του αποκρύπτοντος αυτοκωδικοποιητή σε βίντεο.

Υιοθετώντας τις παραπάνω στρατηγικές, ο αποκρύπτων αυτοκωδικοποιητής για βίντεο καταφέρνει με εκπαίδευση σε λίγα δεδομένα (περίπου τρεις με τέσσερις χιλιάδες βίντεο) να πετύχει ανταγωνιστικές τιμές ορθότητας σε σύνολα δεδομένων όπως το Kinetics-400 (87.4% ορθότητα), Something-Something V2 (91.3% ορθότητα) και UCF101 (62.6% ορθότητα). Α-

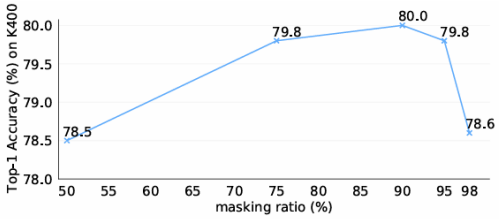


Εικόνα 3.16: *Στο πάνω αριστερά κομμάτι αναπαρίσταται το βίντεο που εισάγεται στο μοντέλο, ενώ στο πάνω δεξιά φαίνεται ένας τρόπος απόκρυψης τμημάτων του βίντεο, όπου καλύπτουται με μάσκες ολόκληρα καρέ. Αυτός ο τρόπος δεν είναι βέλτιστος, αφού δεν βοηθά την εύρεση συσχετισμών μεταξύ των δεδομένων στο πεδίο του χρόνου. Στο κάτω αριστερά κομμάτι είναι η απόκρυψη τμημάτων κάθε καρέ με τυχαίο τρόπο, η οποία ως τακτική αντιμετωπίζει το πρόβλημα της εστίασης σε συσχετισμούς χαμηλού επιπέδου. Τέλος, στο κάτω δεξιά μέρος αναπαρίσταται η σωληνοειδής απόκρυψη τμημάτων των καρέ, που πετυχαίνει τα βέλτιστα αποτελέσματα.*

ποδεικνύεται λοιπόν πως έχει σημασία η ποιότητα και όχι η ποσότητα όσον αφορά τα δεδομένα εκπαίδευσης στην αυτο-επιβλεπόμενη μάθηση. Παρατίθεται και ένα διάγραμμα της επίδοσης του μοντέλου συναρτήσει του ποσοστού των τμημάτων που αποκρύπτονται, όπου φαίνεται ότι με απόκρυψη του 90% των τμημάτων κάθε καρέ προκύπτει η υψηλότερη επίδοση.



(a) Performance on SSV2



(b) Performance on Kinetics-400

Εικόνα 3.17: *Τόσο στα πειράματα με το Kinetics-400 όσο και στα πειράματα με το Something-Something V2 παρατηρείται πως για ποσοστό απόκρυψης 90% σημειώνεται η υψηλότερη ορθότητα πρώτης επιλογής.*

3.4.2 VideoMAE V2

Αν και το VideoMAE είναι ένα πολύ αποδοτικό μοντέλο δεδομένου του μικρού αριθμού βίντεο, στα οποία προεκπαιδεύεται, σε ένα άρθρο του 2023[91] περιγράφονται οι δύο μεγαλύτεροι περιορισμοί, που παρουσιάζει το μοντέλο αυτό. Από τη μία πλευρά, δεν μπορεί να κλιμακωθεί το μοντέλο του οπτικού μετασχηματιστή, που αποτελεί τη ραχοκοκαλιά του αυτοκωδικοποιητή, δηλαδή δεν μπορεί να χρησιμοποιηθεί μετασχηματιστής με περισσότερα επίπεδα και περισσότερες παραμέτρους στη θέση του κωδικοποιητή του VideoMAE, καθώς το υπολογιστικό κόστος θα ήταν τεράστιο. Από την άλλη μεριά, παρόλο που οι αποκρύπτοντες

αποκωδικοποιητές, όπως και το VideoMAE, ως μοντέλα αυτο-επιβλεπόμενης μάθησης έχουν την ικανότητα να βγάζουν αποδοτικά αποτελέσματα με λίγα δεδομένα στη διάθεσή τους, αναπόφευκτα προεκπαίδευση σε μικρότερο σύνολο δεδομένων συνεπάγεται υπερπροσαρμογή σε αυτό.

Για αυτό, οι συγγραφείς του άρθρου, ορισμένοι από τους οποίους συμμετείχαν και στο άρθρο για το VideoMAE, προτείνουν το VideoMAE V2, μία εξελιγμένη εκδοχή του προηγούμενου μοντέλου αποκρύπτοντος αυτοκωδικοποιητή για βίντεο. Η πρωτοτυπία της νέας αυτής εκδοχής που αντιμετωπίζει το πρόβλημα κλιμάκωσης με τη χρήση μετασχηματιστή περισσότερων παραμέτρων, ως ραχοκοκκαλιά του αυτοκωδικοποιητή, είναι η χρήση διπλής μάσκας στο δίκτυο. Η ιδέα της απόκρυψης ενός ποσοστού των τμημάτων από τα δεδομένα εισόδου στον κωδικοποιητή επεκτείνεται και στον αποκωδικοποιητή, στον οποίο πλέον εφαρμόζεται επίσης μάσκα. Εκμεταλλευόμενοι μία εργασία[92] που αποδεικνύει ότι είναι δυνατή η αποδοτική ανακατασκευή μίας εικόνας με λιγότερα διαθέσιμα τμήματα, οι ερευνητές προτείνουν την απόκρυψη μερικών τμημάτων που έχουν γίνει αντικείμενο επεξεργασίας από τον κωδικοποιητή, πριν την είσοδό τους στον αποκωδικοποιητή. Να σημειωθεί πως σε αυτή τη μάσκα δεν ακολουθείται η ίδια τακτική με τη σωληνοειδή απόκρυψη τμημάτων, εφόσον σε αυτήν την περίπτωση της αποκωδικοποίησης των δεδομένων χρειάζεται η εκμάθηση συσχετισμών χαμηλού επιπέδου, μεταξύ τμημάτων στην ίδια χωροχρονική γειτονιά, για να συμπληρωθούν όλα τα κρυμμένα τμήματα. Για αυτό υιοθετείται η τακτική της απόκρυψης μίας ποικιλίας, όσον αφορά τη θέση τους, τμημάτων κάθε καρέ, που ονομάζεται απόκρυψη κινούμενης κυψελίδας.

Ο περιορισμός του μικρού όγκου διαθέσιμων δεδομένων εκπαίδευσης, που οδηγεί στην υπερπροσαρμογή του μοντέλου, ξεπερνιέται από τη δημιουργία ενός νέου υπερσυνόλου δεδομένων βίντεο. Ειδικότερα, συγκεντρώνοντας ένα τεράστιο όγκο βίντεο μικρής διάρκειας χωρίς επισήμανση από τα ήδη υπάρχοντα σύνολα δεδομένων Kinetics, Something-Something, AVA και Web-Vid, αλλά και βίντεο από το Instagram δημιούργησαν ένα τεράστιο υβριδικό σύνολο δεδομένων, αποτελούμενο από 1.35 εκατομμύρια βίντεο.

Αυτό το καινούργιο υπερσύνολο δεδομένων αποτελεί σημαντικό κομμάτι της προοδευτικής προεκπαίδευσης του μοντέλου, που προτείνεται στο άρθρο. Τα στάδια της προεκπαίδευσης είναι:

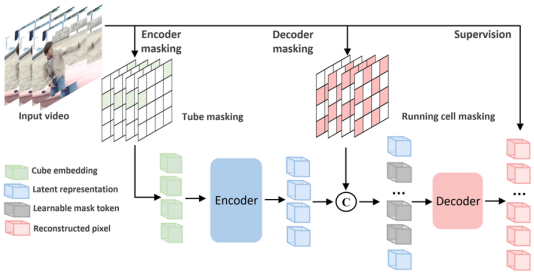
1. Προεκπαίδευση χωρίς επιτήρηση του μοντέλου με το υβριδικό σύνολο δεδομένων που κατασκευάστηκε.

2. Μετά την κατασκευή ενός συνόλου δεδομένων από τη συνένωση ήδη υπάρχοντων συνόλων δεδομένων βίντεο με επισημάνσεις προεκπαιδεύεται εκ νέου το μοντέλο με αυτό. Έτσι μαθαίνει ο αυτοκωδικοποιητής σημασιολογικά χαρακτηριστικά σχετικά με την αναγνώριση δράσης από πολλαπλές πηγές.

3. Γίνεται ρύθμιση με ακρίβεια σε συγκεκριμένα σύνολα δεδομένων, με σκοπό να εξαχθεί γνώση εστιασμένη σε συγκεκριμένα προβλήματα.

Το αποτέλεσμα όλων αυτών των πρωτοτυπιών είναι η μείωση του υπολογικού κόστους με την απόκρυψη δεδομένων όχι μόνο στον κωδικοποιητή αλλά και στον αποκωδικοποιητή, ώστε να μπορέσει να χρησιμοποιηθεί ένας οπτικός μετασχηματιστής ενός δισεκατομμυρίου

παραμέτρων ως ραχοκοκαλιά. Συγκεκριμένα, χρησιμοποιείται ένα Vit-g στη θέση του κωδικοποιητή του μοντέλου και ένα πιο ρηχό μοντέλο στη θέση του αποκωδικοποιητή. Συνολικά επιτυγχάνονται αποτελέσματα τελευταίας τεχνολογίας στην αναγνώριση δράσης σε βίντεο, αποδεικνύοντας πως σημαντικό ρόλο δεν παίζει μόνο η ποιότητα των δεδομένων προεκπαίδευσης αλλά και η ποικιλία τους.



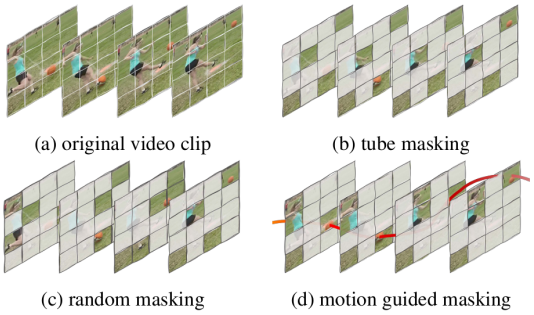
Εικόνα 3.18: *Η αρχιτεκτονική του VideoMAE V2 με κύριο χαρακτηριστικό τη χρήση διπλής μάσκας Η μία εφαρμόζεται στα δεδομένα εισόδου πριν την εισαγωγή τους στον κωδικοποιητή και η δεύτερη εφαρμόζεται στα κρυμμένα χαρακτηριστικά που έχει εξάγει μετά από επεξεργασία ο αποκωδικοποιητής πριν την είσοδό τους στον αποκωδικοποιητή. Με αυτόν τον τρόπο γίνεται πιο αποδοτικό το μοντέλο όσον αφορά τη μνήμη και το υπολογιστικό κόστος. Το σφάλμα ανάμεσα στην αρχική και την τελική είσοδο υπολογίζεται από τη σύγκριση των τμημάτων της εισόδου, που αποκρύφθηκαν από τον κωδικοποιητή.*

3.4.3 MGMAE

Ένας άλλος τρόπος να βελτιωθεί η απόδοση του VideoMAE, εκτός από την κλιμάκωση του μοντέλου οπτικού μετασχηματιστή που χρησιμοποιείται και την κλιμάκωση των δεδομένων εκπαίδευσης, αποτελεί η απόκρυψη των δεδομένων εισόδου από τον κωδικοποιητή με πιο αποδοτικό τρόπο. Αυτή η ιδέα αναλύθηκε στο άρθρο[93] του 2023 με την παρουσίαση του μοντέλου του MGMAE, ένα μοντέλο αποκρύπτοντος αυτοκωδικοποιητή που εφαρμόζει ειδική μάσκα στα δεδομένα. Ειδικότερα, οι συγγραφείς παρατήρησαν ότι, αν και ο προτεινόμενος τρόπος απόκρυψης των τμημάτων των καρέ από τα βίντεο με τη χρήση σωλήνων κατά μήκος διαδοχικών καρέ περιορίζει το πρόβλημα της διαρροής πληροφορίας, όσον αφορά τη δράση που εκτυλίσσεται στο βίντεο, δεν είναι ο βέλτιστος. Αυτός ο τρόπος απόκρυψης εστιάζει στην περίπτωση, κατά την οποία υπάρχουν μικρές αλλαγές στην κίνηση των αντικειμένων ή των πρωταγωνιστών από καρέ σε καρέ, ενώ στην πραγματικότητα υπάρχουν και πολλές περιπτώσεις βίντεο με απότομες αλλαγές, όσον αφορά την κίνηση που διαδραματίζεται σε αυτά. Για αυτό, η προτεινόμενη μέθοδος εκμεταλλεύεται τον εντοπισμό της οπτικής ροής στα καρέ του βίντεο, ώστε να εξασφαλίσει ότι κάθε κινούμενο αντικείμενο ή πρόσωπο στο βίντεο θα παραμείνει ορατό καθ' όλη τη διάρκεια αυτού, ανεξάρτητα από το πόσο γρήγορα αλλάζει η γωνία καταγραφής ή η θέση τους.

Γενικά, η διαφορά των βίντεο από τις εικόνες είναι πως τα πρώτα περιλαμβάνουν εκτός από τη διάσταση του χώρου και αυτή του χρόνου. Συνεπώς, τα μοντέλα που καλούνται να αναγνωρίσουν το είδος δράσης που εκτελείται σε αυτά, πρέπει να λάβουν υπόψιν και την εκ των προτέρων υπάρχουσα χρονική πληροφορία σχετικά με την κίνηση από καρέ σε καρέ του βίντεο. Αυτή την επιπλέον διαθέσιμη πληροφορία αγνοεί η τακτική της τυχαίας απόκρυψης

τμημάτων από τα βίντεο, εφόσον με κάποια πιθανότητα μπορεί να μην συμπεριλάβει τα τμήματα, στα οποία διαδραματίζεται η δράση και έτσι να χάσει την πληροφορία σχετικά με την εκτελούμενη κίνηση. Από την άλλη, η σωληνοειδής απόκρυψη τμημάτων στα βίντεο εισόδου βασίζεται στην υπόθεση ότι σε ένα μεγάλο μέρος των καρέ κάθε βίντεο παρατηρείται μικρή ή καθόλου κίνηση και καλύπτει με μάσκα τμήματα, που βρίσκονται στην ίδια θέση σε όλα τα καρέ του βίντεο, ώστε να μειώσει έτσι την πιθανότητα να αποκρύψει από τον κωδικοποιητή κομμάτια του βίντεο καθοριστικά για την αναγνώριση δράσης. Αυτή η τακτική όμως δεν καλύπτει την περίπτωση των βίντεο με συνεχείς και γρήγορες αλλαγές στην κίνηση, στην οποία ακόμα και η μάσκα σωληνοειδών τμημάτων του βίντεο αυξάνει την πιθανότητα να μείνουν εκτός επεξεργασίας αντικείμενα, ζώα ή άνθρωποι σημαντικοί για την αναγνώριση δράσης στο βίντεο. Αντιθέτως, η δημιουργία μάσκας απόκρυψης των δεδομένων βάσει της εξέλιξης της κίνησης σε ένα βίντεο με τη βοήθεια της οπτικής ροής, ώστε να είναι ορατά αυτά τα σημαντικά αντικείμενα στα τμήματα όλων των καρέ, αποφεύγει τον παραπάνω κίνδυνο αγνόησής τους.



Εικόνα 3.19: *Στο a φαίνεται το αρχικό βίντεο εισόδου χωρισμένο σε τμήματα, κάθε ένα από τα οποία θα καλυφθεί ή όχι από τις μεθόδους απόκρυψης που ακολουθούν. Στο b αναπαρίσταται η σωληνοειδής μάσκα απόκρυψης, που χρησιμοποιείται από το VideoMAE. Στο c φαίνεται η μάσκα απόκρυψης με τυχαίο τρόπο κατά μήκος των καρέ ενός βίντεο. Στο d παρουσιάζεται η κινητικά κατευθυνόμενη μάσκα απόκρυψης, η οποία φαίνεται πως ακολουθεί την εξέλιξη της κίνησης στα διαδοχικά καρέ.*

Η οπτική ροή ουσιαστικά κωδικοποιεί την κίνηση κάθε εικονοστοιχείου από ένα καρέ του βίντεο στο επόμενο, παρακολουθώντας έτσι τη μετακίνηση αντικειμένων συνολικά σε αυτό. Υπάρχουν διαθέσιμοι στο ίντερνετ εκτιμητές οπτικής ροής, όπως το RAFT[94], που δίνουν τη δυνατότητα σύλληψης της πληροφορίας σχετικής με την κίνηση σε βίντεο, αλλά και offline εκτιμητές, που χρησιμοποιούν τον αλγόριθμο TVL1[95] για την εξαγωγή πυκνών ροών ανάμεσα σε όλα τα διαδοχικά καρέ ενός βίντεο. Η κύρια ιδέα, που παρουσιάζεται στο άρθρο, είναι η μεταφορά της γνώσης για την οπτική ροή από τους αλγορίθμους στην μάσκα απόκρυψης, που εφαρμόζεται σε κάθε καρέ του βίντεο, με σκοπό την παραμονή στοιχείων καθοριστικών για την αναγνώριση δράσης στα τμήματα του βίντεο που θα εισαχθούν στο μοντέλο. Αυτό επιτυγχάνεται με τα εξής βήματα:

1. Ορισμός του καρέ-βάσης, για το οποίο αποδεικνύεται μέσα από πειράματα ότι η βέλτιστη επιλογή είναι το μεσαίο καρέ του βίντεο, έστω I_b, όπου b ο δείκτης του καρέ στο βίντεο.

2. Αρχικοποιείται τυχαία μία μάσκα απόκρυψης M_b για το καρέ-βάση.

3. Εξαγωγή των πυκνών οπτικών ροών F για όλο το βίντεο με μία διαδικασία διπλής κατεύθυνσης, που ξεκινά από το I_b.

4. Αλλοίωση της αρχικής μάσκας M_b βάσει των πυκνών ροών F, που έχουν υπολογιστεί, και κατασκευή των μασκών απόκρυψης για όλα τα υπόλοιπα καρέ υπό την καθοδήγηση των ροών F με χρήση της μεθόδου οπίσθιας αλλοίωσης. Χρησιμοποιείται οπίσθια αλλοίωση τόσο για τα προηγούμενα όσο και για τα επόμενα καρέ, καθώς η εμπρόσθια αλλοίωση παρουσιάζει κενά στη μετάδοση της πληροφορίας για την οπτική ροή στο επόμενο καρέ.

Με αυτήν την καινοτόμα τεχνική το MGMAE πετυχαίνει καλύτερες επιδόσεις από το VideoMAE στη ρύθμιση με ακρίβεια πάνω σε σύνολα δεδομένων αναγνώρισης δράσης, όπως το Kinetics-400 και το Something-Something. Μάλιστα αξίζει να σημειωθεί πως ειδικά στο Something-Something, ένα κινησιοκεντρικό σύνολο δεδομένων, παρατηρείται σημαντικότερη βελτίωση της ορθότητας στις προβλέψεις. Αυτό σημαίνει πως η κινητικά κατευθυνόμενη απόκρυψη τμημάτων από τα βίντεο του μοντέλου προσαρμόζεται στις απότομες αλλαγές των κινήσεων και είναι ικανή συλλάβει παρά τις μεταβολές αυτές πληροφορία για τη χρονική εξέλιξη της δράσης σε βίντεο.

3.5 Εφαρμογές

Η αναγνώριση δράσης σε εικόνες και κυρίως σε βίντεο χρησιμοποιείται σε διάφορους τομείς της ανθρώπινης δραστηριότητας, διευκολύνοντας και επισπεύδοντας την εξαγωγή συμπερασμάτων.

Πρώτα απ' όλα ο τομέας της υγείας χρειάζεται εργαλεία, που μπορούν να ερμηνεύσουν τις κινήσεις ασθενών, για να μπορούν οι ειδικοί να καταλάβουν την κατάστασή τους. Μοντέλα μηχανικής μάθησης[96] έχουν χρησιμοποιηθεί στο πλαίσιο της φροντίδας των ηλικιωμένων, ώστε να έχουν τη δυνατότητα οι οικογένειες τους ή το ιατρικό προσωπικό που τους παρακολουθεί να έχουν εικόνα της κατάστασής τους όλο το εικοσιτετράωρο και σε περίπτωση ανάγκης να κληθεί εγκαίρως βοήθεια. Είναι γνωστό ότι οι ασθενείς με Alzheimer δυσκολεύονται να εκτελέσουν καθημερινές πράξεις σημαντικές για την επιβίωσή τους, οπότε συστήματα τεχνητής νοημοσύνης βοηθούν τους φροντιστές τους να τους παρακολουθούν τακτικά εξ αποστάσεως και να κερδίσουν ως ένα βαθμό την αυτονομία τους οι ασθενείς[97]. Τελευταίος αλλά όχι λιγότερο σημαντικός τομέας της υγείας, στον οποίο βρίσκουν εφαρμογή μέθοδοι αναγνώρισης δράσης, είναι αυτός της ψυχικής υγείας. Γεγονός αποτελεί ότι πολλοί άνθρωποι στη σύγχρονη εποχή αντιμετωπίζουν προβλήματα που τους βαραίνουν ψυχολογικά σιωπηλά και υπάρχει ο κίνδυνος να αναπτύξουν κάποια ψυχική ασθένεια. Για να μειωθεί αυτός ο κίνδυνος και να διευκολυνθεί η δουλειά των ιατρών, έχουν δημιουργηθεί συστήματα , που βασίζονται στην ανίχνευση μορφασμών του προσώπου, διάγνωσης ψυχολογικών διαταραχών[98].

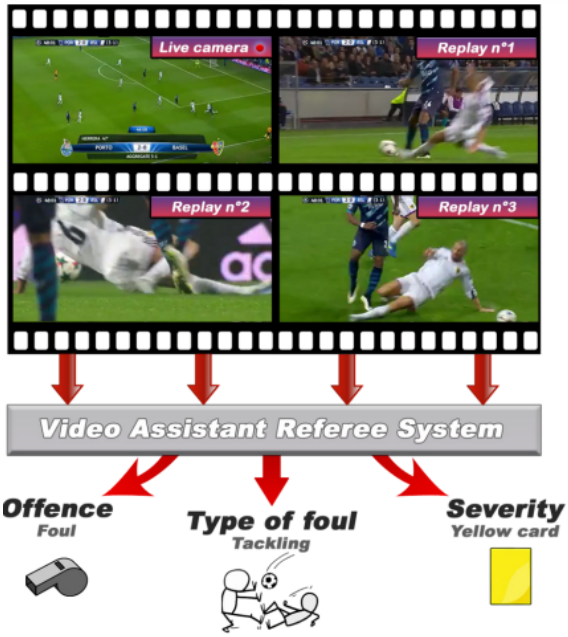
Ένας άλλος τομέας, που αναφέρθηκε παραπάνω, και χρησιμοποιεί τεχνολογίες αναγνώρισης δράσης είναι αυτός του κινηματογράφου. Τόσο κατά τη διάρκεια των γυρισμάτων

όσο και κατά τη διάρκεια της επεξεργασίας μίας ταινίας ή σειράς πριν τη δημοσίευσή της στο κοινό μπορούν να χρησιμοποιηθούν μέθοδοι αναγνώρισης κίνησης, για να διευκολυνθεί η παραγωγή του έργου. Από τη μία, η χρήση έξυπνων καμερών[99], που μπορούν με ευκολία να αναγνωρίζουν την κίνηση σε μία σκηνή και να την ακολουθούν, ακόμα και σε σκηνές με πολύ γρήγορη δράση. Από την άλλη, μπορεί η αναγνώριση των κινήσεων των χειλιών[100] των ηθοποιών στο στάδιο της επεξεργασίας του έργου να συμβάλλει στο συγχρονισμό του οπτικού υλικού που κατέγραψαν οι κάμερες με τις φωνητικές καταγραφές των διαλόγων της σκηνής. Στη σύγχρονη εποχή γίνεται προσπάθεια οι ταινίες και οι σειρές που κυκλοφορούν να είναι προσβάσιμες από όλους τους ανθρώπους ανεξαρτήτως της κατάστασής τους. Για αυτό, η αναγνώριση ανθρώπινης δράσης με συνδυασμό ήχου και εικόνας σε σκηνές κινηματογραφικών έργων χρησιμεύει στην αυτόματη δημιουργία υποτίτλων[101], που περιγράφουν τη σκηνή για συνανθρώπους μας με δυσκολίες στην ακοή.

Με αφορμή τη διεξαγωγή του Ευρωπαϊκού Πρωταθλήματος Ποδοσφαίρου και των Ολυμπιακών Αγώνων αξίζει να αναφερθεί και η συμβολή των αυτόματων συστημάτων αναγνώρισης δράσης στην ανάλυση δεδομένων στον αθλητισμό. Πρώτον, στα ομαδικά αθλήματα υπάρχει η δυνατότητα παρατήρησης του τρόπου παιχνιδιού και εντοπισμού τακτικών[102, 103] τόσο της ομάδας σε προπονήσεις ή επίσημους αγώνες όσο και της αντίπαλης ομάδας, για να προετοιμάσει κατάλληλα ο προπονητής τη στρατηγική του. Οι προπονητές μπορούν επίσης να χρησιμοποιήσουν τεχνολογίες αναγνώρισης κινήσεων, για να εντοπίζουν πρόωρα ανωμαλίες στις κινήσεις αθλητών, με σκοπό την αποφυγή τραυματισμών[104]. Έτσι θα μπορούν να προσαρμόσουν την προετοιμασία των αθλητών τους με πιο ασφαλή για την υγεία τους τρόπο. Ειδικά στο ποδόσφαιρο είναι γνωστό ότι τα τελευταία χρόνια έχει εισαχθεί η τεχνολογία VAR, για να βοηθήσει τους διαιτητές να παίρνουν ορθές αποφάσεις. Αν και ο ανθρώπινος παράγοντας αποτελεί σημαντικό κομμάτι, λόγω της εμπειρίας και της κατανόησης του διακυβεύματος κάθε φάσης, παρουσιάζει ενδιαφέρον η εφαρμογή ενός μοντέλου[4] με την ικανότητα να κρίνει μόνο του αμερόληπτα την παραβίαση κανονισμών σε έναν αγώνα.

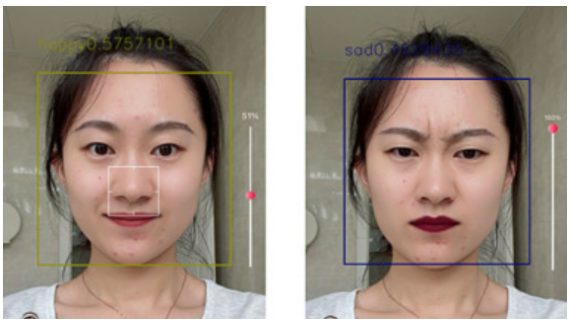
Ο τομέας της παρακολούθησης για λόγους ασφαλείας σε δημόσιους χώρους, όπου υπάρχει υψηλή πιθανότητα παράνομης δραστηριότητας, όπως τα μέσα μαζικής μεταφοράς, ακόμα και τρομοκρατικών ενεργειών, όπως θεατρικές αίθουσες, είναι ιδιαίτερα σημαντικός για την αποτροπή εγκλημάτων. Η αναγνώριση δράσης μπορεί να βρει εφαρμογή σε συστήματα παρακολούθησης χώρων με κάμερες[105], ώστε να βοηθήσει τους υπεύθυνους παρακολούθησης να εντοπίζουν με μεγαλύτερη ακρίβεια και με ταχύτητα περίεργες κινήσεις, που συνδέονται με παράνομες συμπεριφορές, ώστε να προλάβουν το έγκλημα πριν αυτό συμβεί. Πρέπει να σημειωθεί ωστόσο ότι η εφαρμογή της τεχνητής νοημοσύνης σε τέτοια συστήματα οφείλει να γίνει με μεγάλη προσοχή και υπευθυνότητα, γιατί αφενός προκύπτουν ζητήματα ιδιωτικότητας των πολιτών. Αφετέρου, έχει μεγάλη σημασία ο εντοπισμός του μεγαλύτερου δυνατού ποσοστού των πραγματικά εγκληματικών δραστηριοτήτων, οπότε το μοντέλο αναγνώρισης δράσης πρέπει να σημειώνει καλές επιδόσεις στη μετρική της ανάκλησης. Η ανάκληση μετρά το λόγο των σωστά επισημασμένων ως παράνομων ενεργειών προς το σύνολο των πραγματικά παράνομων ενεργειών που εντοπίζει το σύστημα δίνοντας έτσι έμφαση στην ελαχιστοποίηση του αριθμού των λανθασμένα επισημασμένων ως μη παράνομων ενεργειών.

Ο χώρος της μόδας προσφέρεται για τη χρήση τεχνικών της όρασης υπολογιστών για τον πειραματισμό όσον αφορά το συνδυασμό ενδυμάτων, χρωμάτων και υφών[106]. Ειδικότε-



Εικόνα 3.20: *Παράδειγμα του προτεινόμενου συστήματος εικονικού βοηθού διαιτητή στο άρθρο [4], όπου εντοπίζει την κατηγορία foul που συμβαίνει στον αγώνα και αποφασίζει την ποινή του παίκτη που το προκάλεσε.*

ρα η αναγνώριση ανθρώπινης δράσης μπορεί να χρησιμοποιηθεί από τις εταιρείες μόδας όχι μόνο στο επίπεδο σχεδίασης ρούχων, όπου μπορούν να βγάλουν συμπεράσματα για το πόσο διευκολύνουν διάφορες ανθρώπινες κινήσεις και χειρονομίες τα ρούχα, αλλά και στο επίπεδο εξυπηρέτησης πελατών. Δεδομένου ότι είναι πολύ διαδεδομένη η αγορά ρούχων, υποδημάτων και προϊόντων ομορφιάς μέσω του Ίντερνετ στις μέρες μας, τεχνικές αναγνώρισης δράσης επιτρέπουν στους καταναλωτές, που επισκέπτονται τα online καταστήματα εταιρειών, να έχουν πρόσβαση σε εικονικά δοκιμαστήρια[107]. Με αυτόν τον τρόπο μπορούν να σιγουρευτούν για την αγορά τους και να ελαχιστοποιήσουν το ρίσκο μη συμβατότητας που συνεπάγεται η παραγγελία προϊόντων χωρίς δοκιμή εξ αποστάσεως.



Εικόνα 3.21: *Μία τεχνική αναγνώρισης δράσης εστιασμένη στον εντοπισμό μορφασμών του προσώπου βρίσκει εφαρμογή σε ιστοσελίδα καταστήματος με προϊόντα καλλωπισμού προσώπου και δίνει τη δυνατότητα στην πελάτισσα να δοκιμάσει εξ αποστάσεως πάνω της το κραγιόν που την ενδιαφέρει.*

Μέρος II

Πρακτικό Μέρος

Κεφάλαιο 4

Μεθοδολογία και Υλοποίηση

Στόχος της παρούσας εργασίας είναι η ανάπτυξη ορισμένων μοντέλων βαθιάς μάθησης τελευταίας τεχνολογίας. Αυτά τα μοντέλα επεξεργάζονται και αναλύουν δεδομένα από καρέ βίντεο με σκοπό την ταξινόμηση της ανθρώπινης δράσης που αναπαρίσταται σε αυτά.

4.1 Εργαλεία

Για την προετοιμασία των συνόλων δεδομένων και την υλοποίηση των αρχιτεκτονικών, που χρησιμοποιούμε στα πειράματά μας, εργαζόμαστε με τη γλώσσα προγραμματισμού Python3 και ορισμένες βιβλιοθήκες της. Συγκεκριμένα, εισάγουμε και εκμεταλλευόμαστε συναρτήσεις από τις βιβλιοθήκες: torch, deepspeed, decord, timm, triton, moviepy, imageio και mpich.

4.2 Σύνολα Δεδομένων

Τα βίντεο που θα διαχειριστούμε για την υλοποίηση του παραπάνω συστήματος προέρχονται από δύο πολύ διαδεδομένα στο πρόβλημα της αναγνώρισης ανθρώπινης δράσης, το Kinetics 400 και το Kinetics 600.

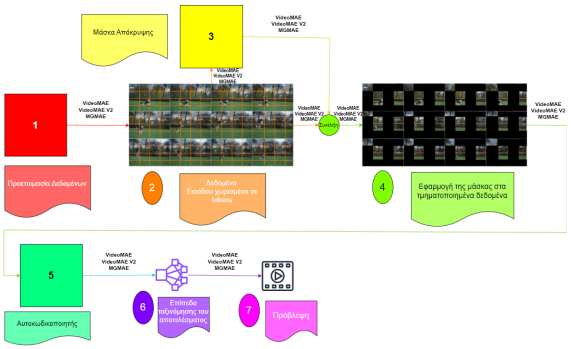
Το Kinetics 400 εμφανίστηκε πρώτη φορά στο άρθρο των Kay et al.[108] το 2017 και αποτελείται από βίντεο που αναπαριστούν 400 κατηγορίες ανθρώπινων δράσεων. Υπάρχουν 400 βίντεο διάρκειας περίπου δέκα δευτερολέπτων για κάθε κλάση δράσης, τα οποία προέρχονται από οπτικοακουστικό υλικό ανεβασμένο στην πλατφόρμα του YouTube. Το σύνολο δεδομένων καλύπτει ένα μεγάλο εύρος ανθρώπινων δράσεων, αφού περιλαμβάνει τόσο αλληλεπίδραση μεταξύ ανθρώπων και αντικειμένων (παραδείγματος χάριν να παίζουν άνθρωποι μουσικά όργανα), ανθρώπων και ανθρώπων (όπως το σφίξιμο χεριών) όσο και ανθρώπους να εκτελούν κινήσεις μόνοι τους (για παράδειγμα χειροκρότημα).

Λόγω της μεγάλης χρησιμότητας που είχε το Kinetics 400 για προεκπαίδευση μοντέλων σε προβλήματα αναγνώρισης ανθρώπινης δράσης, το 2018 οι Carreira et al.[109] δημιούργησαν μία επέκτασή του, το σύνολο δεδομένων Kinetics 600. Αυτό περιλαμβάνει τις 400 κλάσεις ανθρώπινης δράσης που περιλαμβάνει και το Kinetics 400 μαζί με επιπλέον 200 καινούριες κατηγορίες. Επίσης αποτελείται από βίντεο που διαρκούν γύρω στα δέκα δευτερόλεπτα προερχόμενα από δεδομένα διαθέσιμα στο YouTube. Όπως είναι λογικό, αυτό το σύνολο δεδομένων περιέχει μία ακόμα ευρύτερη ποικιλία δράσεων, δίνοντας έτσι τη δυ-

νατότητα στα μοντέλα που εκπαιδεύονται σε αυτό να μάθουν χαρακτηριστικά διαφόρων ανθρώπινων δραστηριοτήτων. Με αυτόν τον τρόπο, μπορούν να γενικεύσουν τις προβλέψεις σε περισσότερα δεδομένα και να μην υπερπροσαρμόζονται για την αναγνώριση συγκεκριμένων δράσεων. Το Kinetics 600 είναι ένα αρκετά μεγάλο σύνολο δεδομένων, εφόσον αποτελείται από 480,000 βίντεο κατά προσέγγιση. Χωρίζεται σε σύνολο εκπαίδευσης, 390,000 βίντεο, σύνολο επαλήθευσης, 30,000 βίντεο, και σύνολο δοκιμών, 60,000 βίντεο.

Η επιλογή μας να πειραματιστούμε με τα συγκεκριμένα σύνολα δεδομένων βασίζεται αφενός στο γεγονός ότι είναι προσβάσιμα από όλους, εφόσον είναι διαθέσιμα μέσω του YouTube, και αφετέρου στο ότι οι αρχιτεκτονικές αποκρύπτοντων αυτοκωδικοποιητών, που αναπαράγουμε, έχουν ήδη προεκπαιδευθεί σε αυτά. Με αυτόν τον τρόπο, μπορούμε να αναδείξουμε τις καλές επιδόσεις των επιλεγμένων μοντέλων σε αυτά τα δεδομένα, καθώς και, αλλάζοντας σε μικρό βαθμό τα δεδομένα εισόδου, να πειραματιστούμε χωρίς να χρειαστεί να εκπαιδεύσουμε τα μοντέλα από την αρχή. Αυτό θα ήτα αρκετά δαπανηρό τόσο σε υπολογιστική ισχύ όσο και σε χρόνο. Συνεπώς, απλώς κάνουμε ρύθμιση με ακρίβεια των μοντέλων στα δύο σύνολα δεδομένων.

Ακολουθεί μία εικόνα που περιγράφει τα βήματα των πειραμάτων μας με τα δύο σύνολα δεδομένων:



Εικόνα 4.1: *Τα βήματα των πειραμάτων μας.*

Αναλυτικότερα:

1. Σε πρώτη φάση προετοιμάζουμε τα δεδομένα, όπως θα παρουσιάσουμε αναλυτικότερα παρακάτω, για να είναι έτοιμα προς επεξεργασία από τα μοντέλα.

2. Στη συνέχεια χωρίζονται σε tokens τα δεδομένα.

3. Ακολουθεί η δημιουργία της μάσκας που αποκρύπτει τμήματα των δεδομένων, (διαφέρει ανάλογα το μοντέλο).

4. Εφαρμόζεται η μάσκα μέσω συνέλιξης στα δεδομένα (στο σχήμα φαίνεται συγκεκριμένα η εφαρμογή της μάσκας σωληνοειδούς απόκρυψης).

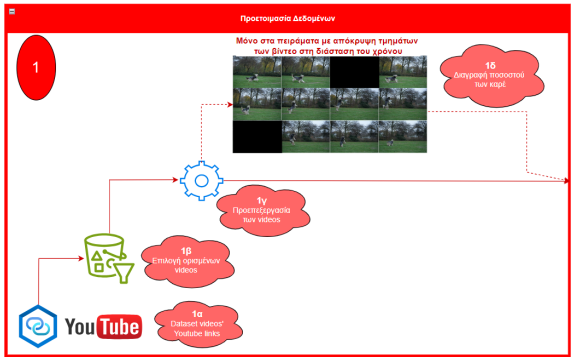
5. Τα τμήματα που δεν έχουν καλυφθεί εισάγονται στο μοντέλο.

6. Τα ανακατασκευασμένα δεδομένα οδηγούνται σε ένα επίπεδο ταξινόμησης της δράσης.

7. Από όπου τελικά προκύπτει η πρόβλεψη του μοντέλου.

4.3 Προετοιμασία Δεδομένων

Παρατίθεται ξανά μία εικόνα, όπου φαίνονται τα επιμέρους βήματα που ακολουθούνται για να βρεθούν, επιλεχθούν και προετοιμαστούν τα δεδομένα εισόδου.



Εικόνα 4.2: *Τα βήματα της προετοιμασίας των δεδομένων.*

1α. Αρχικά βρίσκουμε το αρχείο csv, που περιέχει τους συνδέσμους για τα βίντεο κάθε συνόλου δεδομένων, ώστε να μπορούμε να τα βρούμε στο YouTube, και ο αριθμός, που αντιστοιχεί στην κλάση της δράσης στο βίντεο. Επίσης, στην περίπτωση του αρχείου για το Kinetics 600 μαζί με το σύνδεσμο που αντιστοιχεί σε κάθε βίντεο, δίνεται και το διάστημα των δέκα δευτερολέπτων, στο οποίο εμφανίζεται η δράση που μας ενδιαφέρει να κατηγοριοποιήσουμε. Για τα βίντεο του Kinetics 400, λόγω της περιορισμένης μνήμης στο προγραμματιστικό περιβάλλον που δουλεύουμε και λόγω της ανάγκης ομοιομορφίας με τα δεδομένα του Kinetics 600, υποθέσαμε πως η δράση διαδραματίζεται στη μέση του βίντεο. Οπότε ορίζεται το διάστημα των δέκα δευτερολέπτων στη μέση του βίντεο, το κομμάτι που πρέπει να κρατηθεί. Στην περίπτωση βίντεο διάρκειας μικρότερης των δέκα δευτερολέπτων, διατηρείται όλο το βίντεο.

2α. Όπως αναφέρθηκε παραπάνω, τα δύο σύνολα δεδομένων, που χρησιμοποιούμε είναι πολύ μεγάλου μεγέθους. Συνεπώς η επεξεργασία και η δοκιμή όλων αυτών από τα μοντέλα μας είναι εξαιρετικά χρονοβόρα και κοστίζει πολύ σε υπολογιστική ισχύ. Επιπλέον, το γεγονός ότι όλες οι αρχιτεκτονικές βαθιάς μάθησης που χρησιμοποιούμε έχουν ήδη προεκπαιδευθεί στο Kinetics 400, που αποτελεί υποσύνολο του Kinetics600, και η μία από αυτές και στο Kinetics 600, μας δίνει τη δυνατότητα με τη χρήση μόνο ενός κομματιού των δεδομένων να κάνουμε finetuning των μοντέλων σε αυτά. Για αυτό επιλέγουμε ένα αντιπροσωπευτικό μικρό κομμάτι από τα δύο σύνολα δεδομένων, φροντίζοντας να περιλαμβάνονται βίντεο από κάθε κατηγορία δράσης. Τέλος, εξασφαλίζουμε ότι το σύνολο εκπαίδευσης περιέχει μεγαλύτερο πλήθος βίντεο από αυτά της επαλήθευσης και των δοκιμών

3α. Έχοντας επιλέξει ένα τμήμα του συνόλου εκπαίδευσης, επαλήθευσης και δοκιμών από τα Kinetics 400 και 600, προχωράμε στη λήψη των δεδομένων. Χρησιμοποιούμε συναρτήσεις της βιβλιοθήκης moviepy, προκειμένου να κατεβάσουμε τα επιλεγμένα δεδομένα μέσω συνδέσμων YouTube και στη συνέχεια να τα επεξεργαστούμε κατάλληλα, διατηρώντας το διάστημα δέκα δευτερολέπτων που έχει ήδη οριστεί για το Kinetics

600 και αυτό που ορίσαμε εμείς για το Kinetics 400. Αποθηκεύουμε τα δεδομένα σε ξεχωριστούς φακέλους και δημιουργούμε από αυτά τα έξι υποσύνολα δεδομένων που θα χρειαστούμε, από ένα σύνολο εκπαίδευσης, επαλήθευσης και δοκιμών για τα δύο μεγάλα σύνολα δεδομένων. Τέλος, δημιουργούμε ένα αρχείο με οδηγίες για το μοντέλο, ώστε αυτό να ξέρει την τοποθεσία του κάθε υποσυνόλου δεδομένων στο προγραμματιστικό μας περιβάλλον καθώς και την αριθμητική επισήμανση της δράσης, που περιέχει το καθένα. Έτσι, τα δεδομένα εισόδου είναι έτοιμα να εισαχθούν στους αποκρύπτοντες αυτοκωδικοποιητές των πειραμάτων μας και αυτοί ακολουθώντας μία διαδικασία επιβλεπόμενης μάθησης να ρυθμιστούν με ακρίβεια στα δύο datasets.

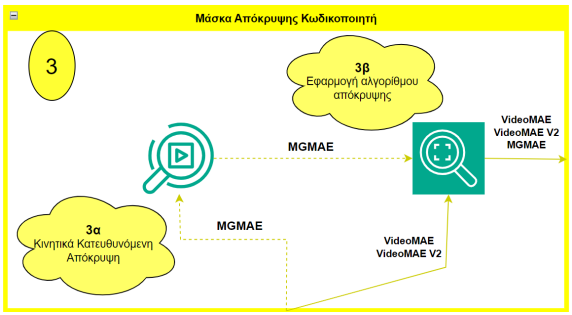
4α. Προκειμένου τα πειράματα της εργασίας να μην αποτελούν μόνο αναπαραγωγή αποτελεσμάτων παλαιότερων εργασιών, δοκιμάζουμε να εκπαιδεύσουμε τα μοντέλα στα δεδομένα με μία πρωτοτυπία. Αυτό μας το επιτρέπει το γεγονός ότι είναι προεκπαιδευμένα στα δύο σύνολα δεδομένων τα μοντέλα και επομένως, οι πειραματισμοί μας δεν απαιτούν εκ νέου εκπαίδευση αυτών των αρχιτεκτονικών πολλών παραμέτρων, αλλά αρκούν οι δοκιμές στα υποσύνολα δεδομένων, που έχουμε ήδη δημιουργήσει κατά το προηγούμενο βήμα. Έτσι, εκτός από τη ρύθμιση με ακρίβεια των αποκωδικοποιητών στα Kinetics 400 και 600, τρέχουμε και πειράματα με παραλλαγές στα δεδομένα εισόδου.

Η ιδέα του πειράματος βασίζεται στο βασικό χαρακτηριστικό των αποκρύπτοντων αυτοκωδικοποιητών, που είναι η χρήση μάσκας για απόκρυψη ορισμένων τμημάτων των δεδομένων. Αυτό όμως γίνεται μόνο σε κάθε καρέ των βίντεο, που εισάγονται στα μοντέλα, δηλαδή αποκλειστικά στη διάσταση του χώρου. Προκύπτει, λοιπόν, το ερώτημα, πώς επηρεάζεται η επίδοση των μοντέλων στη ταξινόμηση ενός βίντεο σε περίπτωση χρήσης μάσκας για απόκρυψη τμημάτων των βίντεο στη διάσταση του χρόνου. Με άλλα λόγια, δοκιμάζουμε με τυχαίο τρόπο να κρύψουμε ένα ποσοστό των καρέ των βίντεο, που εισάγονται στους αυτοκωδικοποιητές, και να παρατηρήσουμε κατά πόσο αυτό μειώνει την ορθότητα με την οποία προβλέπουν την κατηγορία δράσης, στην οποία ανήκει το εκάστοτε βίντεο.

Συνεπώς, καταλήγουμε με δύο εκδοχές πειραμάτων για κάθε μοντέλο με κάθε σύνολο δεδομένων, ένα πείραμα με τα δεδομένα, στα οποία έχει προεκπαιδευθεί το μοντέλο, και ένα πείραμα με τα προεπεξεργασμένα από εμάς δεδομένα, από τα οποία έχει αποκρυφθεί τυχαία ένας αριθμός καρέ.

4.4 Μάσκα Απόκρυψης

Στην παραπάνω εικόνα φαίνονται τα δύο βήματα, που συνιστούν τη μέθοδο δημιουργίας της μάσκας, που θα εφαρμοστεί σε κάθε καρέ των βίντεο. Για την ακρίβεια, μόνο η εφαρμογή του αλγορίθμου απόκρυψης αποτελεί αναπόσπαστο κομμάτι της διαδικασίας δημιουργίας μάσκας, καθώς, όπως υπονοούν οι διακεκομμένες γραμμές στο βελάκι προς αυτό, το βήμα 3α ακολουθείται μόνο στην περίπτωση του μοντέλου MGMAE. Γενικά, στην περίπτωση του βασικού αποκρύπτοντος αυτοκωδικοποιητή για βίντεο, ο αλγόριθμος μάσκας απόκρυψης τμημάτων των δεδομένων που εφαρμόζεται είναι αυτός της σωληνοειδούς απόκρυψης. Με



Εικόνα 4.3: *Τα βήματα κατασκευής της μάσκας απόκρυψης πριν του κωδικοποιητή.*

βάση αυτόν, δημιουργείται μάσκα κατά μήκος όλων των καρέ του βίντεο, με τέτοιο τρόπο, ώστε να αποκρύπτονται σε κάθε καρέ τμήματα, που βρίσκονται στην ίδια θέση.

Ειδικότερα, στην περίπτωση του MGMAE υιοθετείται ένας τρόπος δημιουργίας πιο αποδοτικής μάσκας, που βασίζεται στην ιδέα ότι δεν αποκρύπτονται ποτέ τμήματα του βίντεο, στα οποία διαδραματίζεται κίνηση. Ο υπολογισμός της οπτικής ροής στο βίντεο δίνει τη δυνατότητα να εντοπιστούν τα τμήματα, στα οποία μεταφέρεται η κίνηση από καρέ σε καρέ. Έτσι, δημιουργείται η μάσκα απόκρυψης πριν την εισαγωγή των δεδομένων στον κωδικοποιητή του MGMAE με έναν διαφορετικό αλγόριθμο, κατευθυνόμενο από την αλλαγή στις κινήσεις, σε σχέση με τη μάσκα που εφαρμόζεται στα VideoMAE και VideoMAE V2.

4.5 Μοντέλο

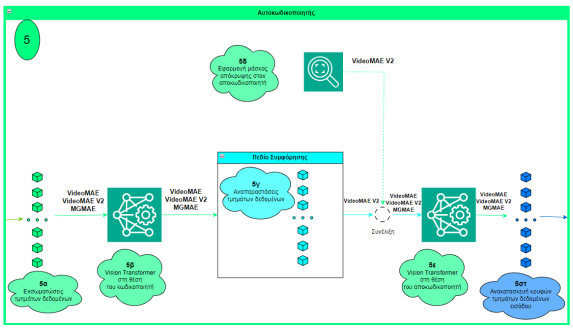
Το μεγαλύτερο κομμάτι των πειραμάτων μας καταλαμβάνει το μοντέλο απορκύπτοντος αυτοκωδικοποιητή, που έχουμε επιλέξει κάθε φορά. Αν και η μέθοδος δημιουργίας μάσκας αποτελεί κομμάτι του μοντέλου, εφόσον την παρουσιάσαμε πριν, ακολουθεί μία περιγραφή των βημάτων, που αφορούν αμιγώς τον αυτοκωδικοποιητή.

4.5.1 Προεπεξεργασία Δεδομένων

Το στάδιο της προεπεξεργασίας των δεδομένων αποτελεί κομμάτι του μοντέλου. Εφόσον στη θέση του κωδικοποιητή χρησιμοποιείται ο κλασικός Vision Transformer, οι εικόνες εισόδου, τα καρέ των βίντεο δηλαδή, πρέπει να τεμαχιστούν σε τμήματα, τα οποία θα μπορέσει να ενσωματώσει σε διανύσματα ο μετασχηματιστής. Αφού έχει χωριστεί σε τμήματα κάθε καρέ, στην εικόνα πιο πάνω προκύπτουν 16 τμήματα σε κάθε καρέ, με συνέλιξη εφαρμόζεται σε αυτά η μάσκα απόκρυψης. Έτσι, προκύπτει για το βίντεο ό,τι φαίνεται στο τέταρτο βήμα της παραπάνω εικόνας. Συγκεκριμένα σε αυτό το παράδειγμα έχει εφαρμοστεί σωληνοειδής απόκρυψη, εφόσον παρατηρείται απόκρυψη των τμημάτων ίδια θέσης σε όλα τα καρέ.

4.5.2 Αυτοκωδικοποιητής

Πιο αναλυτικά τα περιεχόμενα και οι διεργασίες, που εκτελούνται εντός του αυτοκωδικοποιητή: Ο ρόλος του αυτοκωδικοποιητή στο πεδίο της όρασης υπολογιστών, ως μοντέλο αυτο-επιβλεπόμενης μάθησης, είναι η ανακατασκευή των εικόνων εισόδου. Αυτό επιτυγχάνεται μέσω της ανάλυσης των ενσωματώσεων των τμημάτων των δεδομένων από τον κωδι-



Εικόνα 4.4: *Η επεξεργασία των δεδομένων από τον αυτοκωδικοποιητή.*

κοποιητή, ο οποίος είναι ένας οπτικός μετασχηματιστής σε αυτήν την περίπτωση, από την οποία προκύπτουν αναπαραστάσεις των χαρακτηριστικών του βίντεο εισόδου. Αυτές αφορούν πληροφορίες όπως τη μετακίνηση από το βάθος στο μπροστινό μέρος ενός καρέ ή την κίνηση από δεξιά προς τα αριστερά, πληροφορίες που εξάγονται μέσω του υπολογισμού της προσοχής μεταξύ των τμημάτων των δεδομένων στη χρονική και τη χωρική διάσταση.

Οι αναπαραστάσεις αυτές συγκεντρώνονται στο επίπεδο συμφόρησης, ένα επίπεδο του δικτύου όπου έχει συμπιεστεί όλη η πληροφορία από τον κωδικοποιητή, μειώνοντας τη διάσταση των αρχικών ενσωματώσεων των δεδομένων εισόδου. Από αυτο το επίπεδο καλείται ο αποκωδικοποιητής στη συνέχεια να μάθει τα χαρακτηριστικά του βίντεο και να το ανακατασκευάσει. Συγκεκριμένα, πρέπει να προβλέψει την τιμή των εικονοστοιχείων, και κατ' επέκταση των τμημάτων, που έχουν καλυφθεί από τη μάσκα σε κάθε καρέ.

4.5.3 Μάσκα Απόκρυψης Αποκωδικοποιητή

Όπως υπονοεί το βελάκι αποτελούμενο από διακεκομμένες γραμμές στη εικόνα, το βήμα 5δ δεν είναι συστατικό κάθε αυτοκωδικοποιητή. Αντιθέτως, αφορά μόνο την περίπτωση του μοντέλου του VideoMAE V2, το οποίο εφαρμόζει μάσκα απόκρυψης στις αναπαραστάσεις πληροφορίας του επιπέδου συμφόρησης. Αυτό, όπως αποδεικνύεται και στο άρθρο που εισήγαγε το συγκεκριμένο μοντέλο, ενισχύει την ικανότητα του αποκωδικοποιητή να μαθαίνει συσχετίσεις χαμηλού επιπέδου μεταξύ τμημάτων των καρέ και μειώνει το απαιτούμενο υπολογιστικό κόστος, επιτρέποντας τη χρήση μεγαλύτερου οπτικού μετασχηματιστή στη θέση του κωδικοποιητή (βήμα 5β). Η μάσκα σε αυτό το σημείο του δικτύου δεν αποκρύπτει με βάση την οπτική ροή του βίντεο ή με βάση την ίδια θέση των τμημάτων σε κάθε καρέ, αλλά με τυχαίο τρόπο, για να δημιουργηθεί ποικιλία περιπτώσεων καρέ που πρέπει να ανακατασκευαστούν, μέσα από τις οποίες μαθαίνει περισσότερους χαμηλού επιπέδου συσχετισμούς ο μετασχηματιστής, που αναλαμβάνει το ρόλο του αποκωδικοποιητή. Η μάσκα και σε αυτήν την περίπτωση εφαρμόζεται στα στοιχεία του επιπέδου συμφόρησης με συνέλιξη.

4.5.4 Ταξινόμηση

Μετά τη ρύθμιση των μοντέλων των αυτοκωδικοποιητών με ακρίβεια στα Kinetics 400 και 600 μέσω του υποσυνόλου εκπαίδευσης, ελέγχουμε την επίδοσή τους στην πρόβλεψη της σωστής κατηγορίας δράσης σε κάθε βίντεο. Τον αυτοκωδικοποιητή ακολουθεί στο δίκτυο του πειράματός μας ένα επίπεδο ταξινόμησης, που λαμβάνει στο βήμα 6 ως είσοδο τα ανακα-

τασκευασμένα δεδομένα και υπολογίζει με βάση τις τιμές των εικονοστοιχείων των καρέ κάθε βίντεο ποια κλάση δραστηριότητας απεικονίζεται σε αυτά. Ειδικότερα, το μοντέλο υπολογίζει το πόσο πιθανό είναι κάθε κλάση να αναπαρίσταται στο βίντεο, συνδέοντας κάθε μία από αυτές με ένα βαθμό εμπιστοσύνης στην πρόβλεψή του να ανήκει το βίντεο στη συγκεκριμένη κατηγορία δράσης.

Τελικά, το δίκτυο βγάζει στην έξοδο με φθίνουσα σειρά βαθμού εμπιστοσύνης τις κατηγορίες δραστηριοτήτων, που μπορεί να εμφανίζονται στο βίντεο. Εφόσον πρόκειται για πείραμα επιβλεπόμενης μάθησης, συγκρίνονται αυτά τα αποτελέσματα με την πραγματική επισήμανση για τη δράση στο βίντεο και επιστρέφονται δύο μετρικές. Η πρώτη αφορά την ορθότητα της πρώτης, δηλαδή αυτής με το μεγαλύτερο βαθμό εμπιστοσύνης, πρόβλεψης του μοντέλου, η οποία λέγεται "Top-1 accuracy". Η δεύτερη αφορά την ορθότητα των πρώτων πέντε προβλέψεων που επιστρέφει το μοντέλο, τη μετρική "Top-5 accuracy", η οποία ουσιαστικά μετράει την πιθανότητα η δράση που όντως διαδραματίζεται στο βίντεο να βρίσκεται ανάμεσα στις πέντε πρώτες, που έχει προβλέψει το μοντέλο.

Κεφάλαιο 5

Πειράματα και Ανάλυση Αποτελεσμάτων

Στο κεφάλαιο αυτό περιγράφονται αναλυτικά οι αρχιτεκτονικές των μοντέλων, που χρησιμοποιήσαμε στα πειράματά μας, καθώς και οι παράμετροι εκπαίδευσης που θέσαμε, για να ρυθμιστούν με ακρίβεια στα σύνολα δεδομένων Kinetics 400 και Kinetics 600. Στη συνέχεια παρουσιάζονται τα αποτελέσματα των πειραμάτων μας για κάθε μοντέλο με είσοδο τα δύο datasets στην περίπτωση που έχουμε αποκρύψει ποσοστό των καρέ των βίντεο εισόδου και στην περίπτωση που δεν έχουμε εφαρμόσει αυτή τη μάσκα χρονικής απόκρυψης τμημάτων των δεδομένων.

5.1 Πειράματα με Διαφορετικές Αρχιτεκτονικές Αποκρύπτοντων Αυτοκωδικοποιητών για Βίντεο

Για να στήσουμε σε προγραμματιστικό περιβάλλον τα πειράματα finetuning των τριών μοντέλων αποκρύπτοντων αυτοκωδικοποιητών για βίντεο, χρησιμοποιούμε τις οδηγίες που έχουν ανεβάσει στο GitHub οι ίδιοι οι συγγραφείς των άρθρων, που πρότειναν το κάθε μοντέλο. Με βάση αυτές τις οδηγίες λοιπόν, κατεβάζουμε και εγκαθιστούμε ως ραχοκοκαλιά των αρχιτεκτονικών που θα εκπαιδεύσουμε, δηλαδή στη θέση του κωδικοποιητή και του αποκωδικοποιητή των μοντέλων, το Vision Transformer-Base.

Αυτό αποτελεί το μοντέλο οπτικού μετασχηματιστή μεσαίου μεγέθους με 86 εκατομμύρια παραμέτρους και 12 βασικά επίπεδα. Κάθε ένα από αυτά τα επίπεδα αποτελείται από πολλαπλές κεφαλές υπολογισμού αυτο-προσοχής που εναλλάσσονται με επιμέρους επίπεδα πολυεπίπεδων νευρώνων, οι οποίοι ουσιαστικά αποτελούν ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης. Ανάμεσα στα επιμέρους επίπεδα υπάρχουν επίπεδα κανονικοποίησης, επίπεδα γραμμικοποίησης των διανυσμάτων που προκύπτουν καθώς και επίπεδα αποκοπής μέρους των τμημάτων επεξεργασίας του δικτύου. Επίσης, στους πολυεπίπεδους νευρώνες εφαρμόζεται στις εξόδους πολλών κόμβων η μη γραμμική συνάρτηση ενεργοποίησης GeLu.

Η διάσταση των διανυσμάτων που χειρίζεται αυτή η έκδοση του Vision Transformer είναι 768, συνεπώς τα επίπεδα κανονικοποίησης ενσωματώνουν την πληροφορία από το προηγούμενο επίπεδο του δικτύου σε διανύσματα αυτής της διάστασης. Οι αρχικές ενσωματώσεις των δεδομένων εισόδου όμως, υπολογίζονται πριν από τα 12 βασικά επίπεδα, μέσω της εφαρμογής μίας τρισδιάστατης συνέλιξης στο βίντεο εισόδου, όπου δημιουργούνται τα τμήματα του βίντεο και ενσωματώνονται σε διανύσματα σε συνδυασμό με τη θέση τους στο καρέ προ-

έλευσής τους.

Ο οπτικός μετασχηματιστής, που προσθέτουμε ως ραχοκοκαλιά στα δίκτυα των αυτοκωδικοποιητών, έχει προεκπαιδευθεί στο σύνολο δεδομένων ImageNet-21k, που περιλαμβάνει 14 εκατομμύρια εικόνες, και έχει ρυθμιστεί με ακρίβεια στο σύνολο εικόνων ImageNet 2012, που περιέχει 1 εκατομμύριο στοιχεία. Όπως είναι κατανοητό, ο μετασχηματιστής έχει μάθει ήδη σε μεγάλο βαθμό να εξάγει χαρακτηριστικά από εικόνες, οπότε είναι πολύ αποδοτικός και για την εξαγωγή πληροφοριών από τα καρέ των βίντεο, που θα χρειαστεί να αναλύσει.

Αξίζει να σημειωθεί επίσης ότι ο πειραματισμός μας όσον αφορά τη μάσκα απόκρυψης τμημάτων στη διάσταση του χρόνου δεν αποτελεί μέρος της αρχιτεκτονικής των μοντέλων που χρησιμοποιούμε. Επομένως, η απόκρυψη, ουσιαστικά διαγραφή ενός ποσοστού των καρέ που αποτελούν κάθε βίντεο, γίνεται κατά την προεπεξεργασία των δεδομένων.

Τέλος, για να επισπευσθεί η ρύθμιση με ακρίβεια των μοντέλων στα δεδομένα και για να εξάγουμε πιο γρήγορα τα αποτελέσματα από τα πειράματα, χρησιμοποιούμε τις δύο μονάδες επεξεργασίας γραφικών GPU T4, που είναι διαθέσιμες μέσα στο προγραμματιστικό περιβάλλον.

5.1.1 VideoMAE

Η ιδιαιτερότητα του VideoMAE σε σχέση με τα άλλα δύο μοντέλα είναι το γεγονός ότι δεν είχε σχεδιαστεί, για να προεκπαιδευθεί στο σύνολο δεδομένων Kinetics 600, αλλά μόνο στο 400. Συνεπώς χρειάζεται μετά την εγκατάστασή του στο περιβάλλον προγραμματισμού, όπου εκπονούμε τα πειράματα, να προσθέσουμε στα αρχεία παραμετροποίησης των συνόλων δεδομένων, που προορίζονται να εισαχθούν και να αναλυθούν από το μοντέλο.

Στη συνέχεια, ακολουθώντας πάλι τις οδηγίες των ερευνητών που δημιούργησαν το μοντέλο, κατεβάζουμε από το GitHub του άρθρου τους ένα μοντέλο VideoMAE, που έχει περάσει προεκπαίδευση 800 εποχών στο Kinetics 400. Με αυτόν τον τρόπο, είμαστε σε θέση να κάνουμε σε λίγες εποχές και με ένα μικρό ποσοστό των δεδομένων όλου του συνόλου finetuning του μοντέλου, πετυχαίνοντας υψηλές επιδόσεις.

Με βάση τις οδηγίες, ορίζουμε για τη ρύθμιση με ακρίβεια του αυτοκωδικοποιητή τις εξής μεταβλητές εκπαίδευσης:

- batch size: Ορίζει το πλήθος των δεδομένων εισόδου, στην προκειμένη περίπτωση βίντεο, που αναλύονται ταυτόχρονα σε κάθε βήμα εκπαίδευσης του μοντέλου. Ορίζουμε το μέγεθος της κάθε παρτίδας εκπαίδευσης σε 8.

- input size: Ορίζει το μέγεθος, στο οποίο μετατρέπονται οι αρχικές διαστάσεις των καρέ, για να είναι κατάλληλες προς επεξεργασία από το μοντέλο. Ορίζεται σε 224.

- num frames: Ορίζει τον αριθμό τμημάτων ανά πλευρά των καρέ κάθε βίντεο, στα οποία χωρίζεται η εικόνα. Το ορίζουμε σε 16.

- num workers: Ορίζει το πλήθος των παράλληλων διεργασιών, που μπορεί να εκτελεί το σύστημα. Στο προγραμματιστικό περιβάλλον που δουλεύουμε ο μέγιστος αριθμός πυρήνων CPU είναι 4, οπότε αυτήν την τιμή δίνουμε και στη μεταβλητή.

- opt adamw: Ορίζει τον αλγόριθμο βελτιστοποίησης των παραμέτρων του δικτύου κατά τη διάρκεια της εκπαίδευσης. Επιλέγουμε τη στοχαστική εκδοχή του βελτιστοποιητή Adam, τον AdamW.

- lr: Ορίζει το ρυθμό μάθησης του μοντέλου, δηλαδή πόσο γρήγορα ενημερώνονται οι παράμετροι του δικτύου κατά την εκπαίδευσή του. Μεγάλη τιμή ρυθμού μάθησης αφενός οδηγεί σε ταχύτερη σύγκλιση στη βέλτιστη τιμή των παραμέτρων αφετέρου εγκυμονεί τον κίνδυνο να προσπερνάει τη βέλτιστη τιμή, όταν έχει φτάσει πολύ κοντά της. Για να αποφύγουμε το ρίσκο αυτό, επιλέγουμε μία σχετικά μικρή τιμή, 10^{-3}.

- epochs: Ορίζει το πλήθος εποχών εκπαίδευσης του μοντέλου. Ουσιαστικά θέτει τον αριθμό των φορών που θα εισαχθούν τα δεδομένα εκπαίδευσης στο μοντέλο, προκειμένου αυτό να δοκιμάσει να προβλέψει τη σωστή κατηγορία, στην οποία ανήκουν, και μέσω του υπολογισμού του σφάλματός του να ενημερώσει τις τιμές των παραμέτρων του προς τη σωστή κατεύθυνση. Όσο περισσότερες εποχές αφήνεται να εκπαιδευτεί ένα μοντέλο, τόσο πιο κοντά φτάνει στις βέλτιστες τιμές των παραμέτρων του. Επειδή στην περίπτωσή μας, δεν μας ενδιαφέρει η από το μηδέν εκπαίδευση του μοντέλου, ακολουθώντας τις οδηγίες των συγγραφέων, επιλέγουμε ένα μικρό αριθμό εποχών, 10.

- warmup epochs: Ορίζει το νούμερο των εποχών προθέρμανσης εκπαίδευσης. Συχνά κατά τα πρώτα στάδια της εκπαίδευσης ενός μοντέλου, αυτό είναι ευαίσθητο στην υπερπροσαρμογή των παραμέτρων για την ταξινόμηση λίγων παραδειγμάτων του συνόλου δεδομένων, τα οποία μπορεί να αποτελούν και εξαιρέσεις, οπότε δεν θέλουμε να γενικεύσει πάνω σε αυτά. Για αυτό κατά τις εποχές προθέρμανσης είναι μικρότερος ο ρυθμός μάθησης και έτσι ενημερώνονται σε μικρότερο βαθμό οι τιμές των παραμέτρων. Ορίζουμε τη μεταβλητή αυτή σε 3, γεγονός που σημαίνει ότι από τις 10 συνολικές εποχές εκπαίδευσης, οι 3 θα είναι προθέρμανσης.

5.1.2 VideoMAE V2

Σε αντίθεση με τον προκάτοχό του, το VideoMAE V2 έχει προεκπαιδευθεί και στα δύο σύνολα δεδομένων, με τα οποία δουλεύουμε, οπότε δεν χρειάζονται προσθήκες στα αρχεία παραμετροποίησής του όσον αφορά τον τρόπο εισαγωγής των συνόλων δεδομένων σε αυτό. Ωστόσο, η εξελιγμένη αυτή εκδοχή του VideoMAE δουλεύει με ένα εκτεταμένο υβριδικό σύνολο δεδομένων, που αποτελεί μίξη των Kinetics 400, 600 και 700. Οπότε χρειάστηκαν μερικές αλλαγές στο αρχείο, που αναλαμβάνει την αντιστοίχηση των 710 κλάσεων του τεράστιου υβριδικού dataset με τις κλάσεις των μικρότερων και στην περίπτωσή μας των Kinetics 400 και Kinetics 600.

Έπειτα, κατεβάζουμε και σε αυτήν την περίπτωση από το GitHub του άρθρου ένα μοντέλο VideoMAE, που έχει περάσει προεκπαίδευση 1200 εποχών στο σύνολο δεδομένων με τις 710 κατηγορίες ανθρώπινων δράσεων. Το συγκεκριμένο μοντέλο χρησιμοποίησε ως ραχοκοκκαλιά ένα Vision Transformer-Giant, με πολύ περισσότερους παραμέτρους από τον οπτικό μετασχηματιστή, που φορτώνουμε εμείς στα πειράματά μας. Για αυτό, προσέχουμε να εγκαταστήσουμε μία προεκπαιδευμένη εκδοχή του μοντέλου προσαρμοσμένη στο μικρότερου

μεγέθους Vision Transformer-Base, στην οποία έχει γίνει απόσταξη γνώσης και μεταφορά των βελτιστοποιημένων τιμών των παραμέτρων από το μεγαλύτερο προεκπαιδευμένο μοντέλο.

Οι μεταβλητές εκπαίδευσης του VideoMAE V2 δεν διαφέρουν πολύ από αυτές του VideoMAE ούτε ως προς το είδος, ούτε ως προς την τιμή:

- batch size: Ορίζουμε το μέγεθος της κάθε παρτίδας εκπαίδευσης σε 3. Μικρότερο μέγεθος παρτίδων βελτιώνει την ταχύτητα εκπαίδευσης, χαρακτηριστικό που χρειάζεται στην περίπτωση του μεγαλύτερου και βαθύτερου δικτύου του VideoMAE V2, αλλά και μειώνει τη χρησιμοποιούμενη από το μοντέλο μνήμη του συστήματος, χαρακτηριστικό εξίσου σημαντικό στο πλαίσιο περιορισμένων πόρων του προγραμματιστικού περιβάλλοντος, όπου εργαζόμαστε. Από την άλλη, μικρότερη τιμή αυτής της μεταβλητής οδηγεί σε εσφαλμένους υπολογισμούς της κλίσης, που χρησιμοποιείται για την ενημέρωση των τιμών των μεταβλητών, εφόσον η εκπαίδευση σε περιορισμένο τμήμα του συνόλου δεδομένων προκαλεί εξειδίκευση της ικανότητας πρόβλεψης σε αυτά.

- clip grad: Ορίζει το πάνω όριο των τιμών που μπορεί να πάρει η κλίση ενημέρωσης των τιμών των παραμέτρων του μοντέλου. Αυτή η μεταβλητή μετριάζει σε ένα βαθμό την επιρροή που μπορεί να έχει η χαμηλή τιμή της προαναφερθείσας μεταβλητής στην κλίση ενημέρωσης. Έτσι αποφεύγεται η εκτόξευση της τιμής της κλίσης, η οποία μπορεί να οφείλεται στην εμφάνιση έκτοπων τιμών κατά την εκπαίδευση. Τη θέτουμε σε 5.

- input size: Ορίζεται σε 224, όπως και στο VideoMAE, γιατί έχει να κάνει με τα χαρακτηριστικά των εικόνων που είναι ικανός να επεξεργαστεί ο οπτικός μετασχηματιστής, που έχουμε θέσει ως βάση της αρχιτεκτονικής του μοντέλου.

- num frames: Το ορίζουμε σε 16, καθώς με αυτήν την τιμή τμηματοποίησης των καρέ των βίντεο λειτουργεί ο Vision Transformer-Base.

- num workers: Ορίζεται ξανά σε 4, εφόσον αυτός είναι ο μέγιστος διαθέσιμος αριθμός πυρήνων CPU στο προγραμματιστικό περιβάλλον, όπου στήνουμε τα πειράματα.

- opt adamw: Επιλέγουμε ξανά τη στοχαστική εκδοχή του βελτιστοποιητή Adam, τον AdamW.

- lr: Επιλέγουμε μία μικρότερη τιμή, 10^{-5}, από αυτήν στο VideoMAE.

- epochs: Επειδή και σε αυτήν περίπτωσή, δεν μας ενδιαφέρει η από το μηδέν εκπαίδευση του μοντέλου, ακολουθώντας τις οδηγίες των συγγραφέων, επιλέγουμε ένα μικρό αριθμό εποχών, 10.

- warmup epochs: Ορίζουμε ξανά τη μεταβλητή αυτή σε 3.

5.1.3 MGMAE

Αν και το MGMAE δεν έχει προεκπαιδευθεί στο Kinetics 400, μοιράζεται τα ίδια αρχεία παραμετροποίησης με το VideoMAE V2, συνεπώς δεν χρειάζεται να προσθέσουμε γραμμές κώδικα σε αυτά για να αποκτήσει τη δυνατότητα το δίκτυο να επεξεργαστεί τα δεδομένα του

Kinetics 400. Εφόσον, όμως, η σελίδα στο GitHub του μοντέλου περιέχει τα ίδια αρχεία με αυτή του VideoMAE V2, χρειάζεται και σε αυτήν την περίπτωση να διορθώσουμε τον τρόπο που αντιστοιχίζονται οι κλάσεις δράσεων του υβριδικού συνόλου δεδομένων με 710 κατηγορίες ανθρώπινων δραστηριοτήτων με με τις κλάσεις των των Kinetics 400 και Kinetics 600.

Πάλι χρειάζεται η εγκατάσταση ενός μοντέλου με έτοιμες βελτιστοποιημένες τιμές βαρών και συντελεστών συστημικού σφάλματος, που έχουν προκύψει από τη διαδικασία προεκπαίδευσής του, για να κάνουμε finetuning του MGMAE. Κατεβάζουμε από το GitHub του άρθρου, όπου δημοσιεύτηκε η ιδέα της συγκεκριμένης αρχιτεκτονικής αυτοκωδικοποιητή, ένα έτοιμο μοντέλο, το οποίο έχει προεκπαιδευθεί για 800 εποχές στο Kinetics 400. Ξανά προσέχουμε η εκδοχή του μοντέλου που κατεβάζουμε να χρησιμοποιεί ως ραχοκοκαλιά το Vision Transformer-Base, όπως όλες οι αρχιτεκτονικές των πειραμάτων μας, ώστε να έχουν το ίδιο μέτρο σύγκρισης.

Οι μεταβλητές εκπαίδευσης του MGMAE είναι ακριβώς ίδιες με αυτές του VideoMAE V2:

- batch size: Ορίζουμε και εδώ το μέγεθος της κάθε παρτίδας εκπαίδευσης σε 3. Όπως και στην προηγούμενη αρχιτεκτονική επιλέγουμε μικρό μέγεθος παρτίδας για λόγους εξοικονόμησης χρόνου και μνήμης.
- clip grad: Τίθεται σε 5 περιορίζοντας το πάνω όριο της τιμής της κλίσης ενημέρωσης και αποφεύγοντας έτσι τον κίνδυνο εκτόξευσής της.
- input size: Ορίζεται σε 224, εφόσον και σε αυτήν την αρχιτεκτονική βάση αποτελεί το Vision Transformer-Base, που λειτουργεί με αυτές τις διαστάσεις δεδομένων.
- num frames: Ορίζουμε σε 16 τη δημιουργία τμημάτων ανά πλευρά των καρέ των βίντεο, λόγω του Vision Transformer-Base.
- num workers: Ορίζεται ξανά σε 4, εφόσον αυτός είναι ο μέγιστος διαθέσιμος αριθμός πυρήνων CPU στο προγραμματιστικό περιβάλλον, όπου στήνουμε τα πειράματα.
- opt adamw: Επιλέγουμε κατ' επανάληψη τη στοχαστική εκδοχή του βελτιστοποιητή Adam, τον AdamW.
- lr: Επιλέγουμε την ίδια μικρή τιμή, 10^{-5}, με αυτή του VideoMAE V2.
- epochs: Χωρίς να υπάρχει ούτε σε αυτήν την περίπτωση η ανάγκη εκτενούς εκπαίδευσης του μοντέλου στα δύο σύνολα δεδομένων, αλλά απλώς θέλουμε να το ρυθμίσουμε με ακρίβεια σε αυτά αναδεικνύοντας τις ήδη γνωστές καλές του επιδόσεις, επιλέγουμε την τιμή 10.
- warmup epochs: Ακολουθούμε την ίδια τακτική με τα δύο προηγούμενα μοντέλα και θέτουμε τη μεταβλητή αυτή σε 3.

5.2 Ανάλυση Αποτελεσμάτων

Σε αυτήν την ενότητα θα παρουσιάσουμε τα αποτελέσματα των πειραμάτων μας. Όπως έχουμε αναφέρει προηγουμένως, ρυθμίζουμε με ακρίβεια τα τρία μοντέλα αποκρύπτοντων

αυτοκωδικοποιητών για βίντεο στα δύο σύνολα δεδομένων σε δύο περιπτώσεις. Στην πρώτη περίπτωση αναπαράγουμε απλώς τη διαδικασία που περιγράφεται στα άρθρα των μοντέλων και στην άλλη καλύπτουμε με μάσκα ένα ποσοστό των καρέ κάθε βίντεο εισόδου, πριν ακολουθήσουμε πάλι τη διαδικασία του finetuning. Να σημειωθεί πως στη δεύτερη περίπτωση επιλέξαμε το ποσοστό απόκρυψης των καρέ στα βίντεο να είναι 10%, ένα μικρό σχετικά νούμερο, ώστε να μην αλλοιωθεί σε μεγάλο βαθμό το περιεχόμενο των βίντεο και να έχουν τη δυνατότητα να επιτύχουν ανταγωνιστικές επιδόσεις, όπως και είχαν.

Σε πρώτη φάση παρατίθενται συγκεντρωμένες σε δύο πίνακες οι επιδόσεις Top-1 accuracy και Top-5 accuracy των μοντέλων στο σύνολο δοκιμών κάθε συνόλου δεδομένων. Στον ένα πίνακα είναι τα αποτελέσματα των πειραμάτων, πριν τα οποία έχουμε διαγράψει το 10% των καρέ, ενώ στον άλλον είναι τα πειράματα χωρίς κάποια προεπεξεργασία των δεδομένων. Επίσης, για κάθε ένα από αυτά τα πειράματα παρουσιάζονται σε πίνακες πάλι οι χρόνοι εκπαίδευσης των μοντέλων. Και αυτοί οι πίνακες διαχωρίζουν τα αποτελέσματα με βάση το αν έχει γίνει προεπεξεργασία των δεδομένων ή όχι.

Ξεκινάμε με τους πίνακες, που αφορούν τα πειράματα χωρίς προεπεξεργασία των βίντεο εισόδου:

		VideoMAE	**VideoMAE V2**	**MGMAE**
Kinetics 400	Top-1	77.54%	78.66%	75.63%
	Top-5	90.66%	94.45%	90.59%
Kinetics 600	Top-1	32.13%	81.31%	36.76%
	Top-5	52.21%	96.66%	55.85%

Πίνακας 5.1: *Πίνακας για την ορθότητα της πρώτης και των 5 πρώτων προβλέψεων των μοντέλων για τα Kinetics 400 και Kinetics 600*

Παρατηρούμε, όπως είναι λογικό, πως την καλύτερη επίδοση και στα δύο σύνολα δεδομένων σημειώνει το VideoMAE V2. Στην περίπτωση του Kinetics 400 αυτό οφείλεται στο γεγονός ότι το συγκεκριμένο μοντέλο έχει προεκπαιδευθεί με τη βοήθεια ενός τεράστιου υβριδικού συνόλου δεδομένων. Συνεπώς έχει καταφέρει να γενικεύσει σε μεγαλύτερη ποικιλία δεδομένων. Όσον αφορά την περίπτωση του Kinetics 600, υπενθυμίζουμε ότι το VideoMAE V2 είναι το μόνο από τα τρία μοντέλα, το οποίο έχει προεκπαιδευθεί στην κατηγοριοποίηση των δράσεων στο μεγαλύτερο αυτό σύνολο δεδομένων. Οι τιμές της ορθότητας είναι στα επίπεδα, αυτών που παρουσιάζονται και στο άρθρο του μοντέλου, όντας 9% και 6% χαμηλότερες από αυτές για το Kinetics 400 και Kinetics 600 αντίστοιχα, εφόσον κάνουμε finetuning σε ένα μικρό αριθμό εποχών.

Συγκρίνοντας τα αποτελέσματα των VideoMAE και MGMAE στο Kinetics 400 παρατηρούμε πως η ορθότητα πρόβλεψης του πρώτου είναι υψηλότερη κατά περίπου 1% από αυτή του δεύτερου, αν και από τις μετρικές επίδοσης των δύο μοντέλων στα αντίστοιχα άρθρα παρουσίασής τους περιμέναμε το αντίθετο. Αυτό πιθανότατα οφείλεται στο μικρό αριθμό εποχών, που αφήνουμε τα δύο δίκτυα να ρυθμιστούν με ακρίβεια στο σύνολο δεδομένων, οπότε βραχυπρόθεσμα οι επιδόσεις του VideoMAE κατάφεραν να ξεπεράσουν αυτές του MGMAE. Από την άλλη, η ορθότητα πρόβλεψης του MGMAE στο Kinetics 600, στο οποίο

κανένα από τα δύο μοντέλα δεν έχει προεκπαιδευθεί, παρουσιάζει υψηλότερη τιμή από αυτή του VideoMAE, όπως ήταν αναμενόμενο. Συγκεκριμένα, η πρωτοποριακή κινητικά κατευθυνόμενη δημιουργία μάσκας για την κάλυψη τμημάτων των δεδομένων εισόδου επικρατεί της κλασικής μεθόδου σωληνοειδούς απόκρυψης του VideoMAE και με αυτόν τον τρόπο στο finetuning στο νέο, για τα δύο μοντέλα, σύνολο δεδομένων προέβλεψε κατά 3%-4% ορθότερα την κλάση των δράσεων.

Στην περίπτωση του finetuning των VideoMAE και MGMAE στο Kinetics 600 παρατηρείται μεγάλη διαφορά στην ορθότητα μεταξύ αυτών των μοντέλων και του VideoMAE V2. Τα δύο πρώτα δεν έχουν προεκπαιδευθεί στο συγκεκριμένο σύνολο δεδομένων και το Kinetics 600 περιέχει όλες τις κλάσεις του Kinetics 400 με την προσθήκη 200 νέων. Οπότε, ο λόγος γνωστών κλάσεων, λόγω προεκπαίδευσης στο Kinetics 400, προς το σύνολο των κλάσεων είναι 2/3 για αυτά τα δύο μοντέλα. Για αυτό το λόγο, προσμέναμε αυτά τα μοντέλα να επιτύχουν το 2/3 της επίδοσης του VideoMAE V2 στα πειράματα με το Kinetics 600. Στα πειράματά μας, όμως, σημειώνεται ένας λόγος περίπου 3/5 της ορθότητας των πέντε πρώτων προβλέψεων των VideoMAE, MGMAE προς αυτής του VideoMAE V2: $52.21/96.66 \approx 55.85/96.66 \approx 0.6$, ενώ ο λόγος των μετρικών Top-1 accuracy είναι 2/5 περίπου: $32.13/81.31 \approx 36.76/81.31 \approx 0.4$. Αυτή η απόκλιση από το αναμενόμενο αποτέλεσμα οφείλεται κατά πάσα πιθανότητα ξανά στο μικρό αριθμό εποχών, κατά τις οποίες κάνουμε finetuning.

Ως γενική παρατήρηση προφανώς και οι τιμές της ορθότητας των πέντε πρώτων προβλέψεων είναι πάντα υψηλότερες από αυτές της ορθότητας της πρώτης πρόβλεψης, αφού η πρώτη μετρική λαμβάνει υπόψη την επίδοση της πρώτης πρόβλεψης συναθροίζοντάς την με των τεσσάρων επόμενων πιο πιθανών.

	VideoMAE	**VideoMAE V2**	**MGMAE**
Kinetics400	3749s	4231s	4395s
Kinetics600	4164s	5526s	5652s

Πίνακας 5.2: *Πίνακας χρόνων εκπαίδευσης κάθε μοντέλου στα δύο σύνολα δεδομένων*

Παρατηρείται πως τα αποτελέσματα, όσον αφορά το χρόνο εκπαίδευσης των τριών μοντέλων στα δύο σύνολα δεδομένων, είναι αυτά που περιμέναμε. Ειδικότερα. γενικά η ρύθμιση με ακρίβεια στο Kinetics 600 παίρνει περισσότερο χρόνο από την αντίστοιχη διαδικασία στο μικρότερο Kinetics 400. Μάλιστα, η διαφορά στους χρόνους ανάμεσα στα δύο σύνολα δεδομένων είναι μικρότερη για το VideoMAE, καθώς αποτελεί το απλούστερο μοντέλο, το οποίο δεν επιβαρύνεται στον ίδιο βαθμό με την προσθήκη επιπλέον δεδομένων προς επεξεργασία.

Σχετικά με τη σύγκριση των χρόνων εκπαίδευσης ανάμεσα στα μοντέλα, σημειώνεται ότι για το MGMAE απαιτείται μεγαλύτερο χρονικό διάστημα για να εκπαιδευθεί. Αυτό είναι λογικό, εφόσον χρησιμοποιεί τη μέθοδο εντοπισμού κίνησης από καρέ σε καρέ του βίντεο μέσω του υπολογισμού της οπτικής ροής, για να κατασκευάσει τη μάσκα απόκρυψης, μία σύνθετη και κοστοβόρα χρονικά διαδικασία. Δεύτερο σε απαιτούμενη διάρκεια εκπαίδευσης μοντέλο έρχεται το VideoMAE V2, το οποίο αφενός ακολουθεί την απλή τακτική της σωληνοειδούς απόκρυψης τμημάτων των δεδομένων εισόδου, αλλά αφετέρου παρουσιάζει μεγαλύτερη

πολυπλοκότητα από το VideoMAE. Το VideoMAE δεν περιλαμβάνει στην αρχιτεκτονική του εφαρμογή δεύτερης μάσκας απόκρυψης πριν το επίπεδο του αποκωδικοποιητή, συνεπώς χρειάζεται το λιγότερο από τις τρεις αρχιτεκτονικές χρόνο εκπαίδευσης.

Ακολουθούν οι πίνακες με τα σχετικά αποτελέσματα για τα πειράματά μας, που εκτελέσαμε με τη διαγραφή ενός 10% των καρέ κάθε βίντεο.

		VideoMAE	**VideoMAE V2**	**MGMAE**
Kinetics 400	Top-1	76.35%	78.15%	74.12%
	Top-5	89.82%	93.28%	88.91%
Kinetics 600	Top-1	28.92%	81.28%	30.39%
	Top-5	51.41%	94.44%	50.72%

Πίνακας 5.3: *Πίνακας για την ορθότητα της πρώτης και των 5 πρώτων προβλέψεων των μοντέλων για τα σύνολα δεδομένων, αφού έχουν υποστεί προεπεξεργασία*

Ως ένα πρώτο σχόλιο αναφέρουμε ότι και σε αυτήν την περίπτωση το VideoMAE V2 φαίνεται να πετυχαίνει με μεγαλύτερη ευστοχία από τα άλλα δύο μοντέλα τη σωστή κατηγορία δράσης στα βίντεο, που προέρχονται από το Kinetics 400 και από το Kinetics 600. Αυτό, όπως αναφέρθηκε και παραπάνω οφείλεται στην προεκπαίδευση του μοντέλου σε ένα εκτεταμένο σύνολο δεδομένων.

Σε αυτή την εκδοχή του πειράματος, όμως, παρατηρούμε πως η ορθότητα πρόβλεψης του MGMAE είναι χαμηλότερη από αυτή της πρόβλεψης του VideoMAE όχι μόνο στο Kinetics 400, όπως είχε σημειωθεί και στα πειράματα χωρίς απόκρυψη καρέ των βίντεο, αλλά και στο Kinetics 600. Συγκεκριμένα, στην ορθότητα πρώτης πρόβλεψης παρουσιάζεται μία διαφορά 2% στο Kinetics 400 και 4% στο Kinetics 600, ενώ στην ορθότητα των πέντε πρώτων προβλέψεων, μόνο μία διαφορά περίπου ίση με 1%. Ίσως αυτή η πτώση της ικανότητας του MGMAE να κατηγοριοποιεί σωστά τα βίντεο εισόδου με βάση τη δραστηριότητα, που επιτελείται σε αυτά, δικαιολογείται, επειδή με την τυχαία διαγραφή του 10% των καρέ σε κάθε βίντεο επηρεάζεται ο υπολογισμός της οπτικής ροής εντός του βίντεο και ο τρόπος που δημιουργείται η μάσκα απόκρυψης στο συγκεκριμένο μοντέλο. Συνεπώς, η μάσκα πιθανότατα δεν ακολουθεί την κίνηση ανθρώπων ή αντικειμένων στο βίντεο λόγω της ελλιπούς πληροφορίας διαθέσιμης στο βίντεο για την κίνησή τους.

Επιπλέον, παρατηρείται και σε αυτά τα πειράματα η χαμηλή επίδοση των VideoMAE και MGMAE στην πρόβλεψη των κλάσεων δράσης στα βίντεο του Kinetics 600, όπου δεν έχουν δεχτεί προεκπαίδευση. Ο αναμενόμενος λόγος 2/3 της τιμής ορθότητας του VideoMAE V2, που περιμένουμε να φτάσουν τα VideoMAE, MGMAE, δεν επιτυγχάνεται και μάλιστα μειώνεται ακόμα περισσότερο. Ειδικότερα, ο λόγος των επιδόσεων των VideoMAE, MGMAE προς αυτές του VideoMAE V2 πλησιάζει το 0.35, αν κοιτάξουμε το Top-1 accuracy, και είναι 0.54 ως προς τη μετρική Top-1 accuracy.

Ο λόγος που συμβαίνει αυτό είναι ότι παρατηρείται μία γενική μείωση όλων των μετρικών, σε όλα τα μοντέλα και σε κάθε dataset. Αυτό που προκαλεί ενδιαφέρον είναι πως η μεγαλύτερη πτώση της ορθότητας πρόβλεψης παρατηρείται στα μοντέλα VideoMAE και MGMAE και συγκεκριμένα στις επιδόσεις τους στο Kinetics 600. Το MGMAE λόγω της επίδρασης που

έχει η αφαίρεση καρέ από τα βίντεο στον υπολογισμό της μάσκας απόκρυψής του φαίνεται να επηρεάζεται περισσότερο από τα δύο, αφού σημειώνει μία πτώση της τάξης του 5%-6% και στις δύο μετρικές. Το VideoMAE από την άλλη εμφανίζει μία σημαντική μείωση στην ορθότητα πρώτης πρόβλεψης, κατά ένα 4%, ενώ λιγότερο σημαντική είναι η αλλαγή στην τιμή της ορθότητας των πέντε πρώτων προβλέψεων, κατά 1%. Φυσικά, αυτές οι αισθητές πτώσεις στην επίδοση παρατηρούνται στο σύνολο δεδομένων, όπου δεν έχουν προεκπαιδευθεί τα δύο μοντέλα.

Σε όλες τις άλλες περιπτώσεις παρατηρείται μία ανεπαίσθητη μείωση της ορθότητας πρόβλεψης της τάξης του 1%-2% με το VideoMAE V2 να επηρεάζεται λιγότερο από όλα τα μοντέλα στο Kinetics 400, εφόσον παρατηρείται πτώση της επίδοσής του κατά λιγότερο από 1% (0.5% στο Top-1 accuracy και 0.2% στο Top-5 accuracy). Αυτή η πτώση ήταν αναμενόμενη, δεδομένου ότι μειώθηκε η διαθέσιμη πληροφορία για τη δράση που αναπαριστάται σε κάθε βίντεο και αναπόφευκτα ελαττώνεται η ορθότητα πρόβλεψης όλων των μοντέλων.

	VideoMAE	**VideoMAE V2**	**MGMAE**
Kinetics400	3210s	3870s	3793s
Kinetics600	3295s	4815s	4732s

Πίνακας 5.4: *Πίνακας χρόνων εκπαίδευσης κάθε μοντέλου στα δύο σύνολα δεδομένων με την αφαίρεση του 10% των καρέ από κάθε βίντεο*

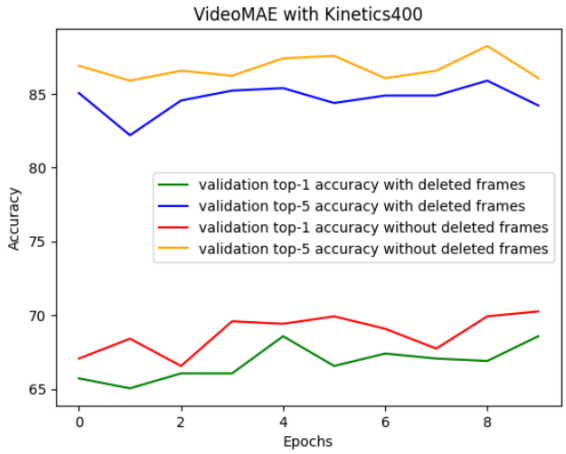
Για άλλη μία φορά, η πιο απλή αρχιτεκτονική, το VideoMAE απαιτεί τον λιγότερο χρόνο εκπαίδευσης, ωστόσο παρατηρείται μία αλλαγή όσον αφορά το ποια από τις άλλες δύο αρχιτεκτονικές δικτύου χρειάζεται να ολοκληρώσει τη διαδικασία της εκπαίδευσης σε λιγότερο χρόνο. Ενώ στην προηγούμενη εκδοχή του πειράματος είναι το MGMAE με την περίπλοκη τεχνική δημιουργίας της μάσκας απόκρυψης, σε αυτήν την εκδοχή το πιο κοστοβόρο χρονικά μοντέλο είναι το VideoMAE V2 αν και με πολύ μικρή διαφορά, της τάξης των 100 δευτερολέπτων ή 1.5 λεπτού. Μπορούμε να υποθέσουμε ότι η διαγραφή ενός 10% των δεδομένων εισόδου μειώνει σε τέτοιο βαθμό την πολυπλοκότητα της διαδικασίας υπολογισμού της οπτικής ροής, ώστε η διπλή εφαρμογή μάσκας στο VideoMAE V2, μία πριν από τον κωδικοποιητή και μία πριν το στάδιο του αποκωδικοποιητή απαιτεί περισσότερο χρόνο επεξεργασίας από αυτήν, ακόμα και αν τα δεδομένα εισόδου μειώνονται κατά το ίδιο ποσοστό και στα δύο μοντέλα.

Όσον αφορά το γενικότερο σχολιασμό των αποτελεσμάτων για τον χρόνο εκπαίδευσης, σε σύγκριση με τους χρόνους που χρειάστηκαν, για να ολοκληρωθούν τα πειράματα χωρίς προεπεξεργασία των δεδομένων εισόδου, αυτά ήταν όπως τα περιμέναμε. Συγκεκριμένα, ο χρόνος που απαιτείται για την διαδικασία εκπαίδευσης κάθε μοντέλου σε κάθε σύνολο δεδομένων είναι στα πειράματα, όπου έχει διαγραφεί το 10% των καρέ από τα βίντεο εισόδου, μικρότερος από τον αντίστοιχο στα πειράματα, όπου δεν κάναμε καμία αλλαγή στα δεδομένα εισόδου.

5.2.1 VideoMAE

Στη συνέχεια παραθέτουμε το κοινό διάγραμμα σφάλματος εκπαίδευσης και σφάλματος επαλήθευσης συναρτήσει των εποχών εκπαίδευσης του μοντέλου καθώς και το κοινό διάγραμμα της ορθότητας πρώτης πρόβλεψης και της ορθότητας των πέντε πρώτων προβλέψεων στο σύνολο επαλήθευσης συναρτήσει των εποχών εκπαίδευσης. Αυτό γίνεται για κάθε σύνολο δεδομένων, στο οποίο ρυθμίζεται το VideoMAE. Επίσης, στα κοινά διαγράμματα εμφανίζονται οι τιμές τόσο για τα πειράματα με τα αυθεντικά δεδομένα όσο και για τα πειράματα με τα προεπεξεργασμένα δεδομένα. Οι τιμές των σφαλμάτων αφορούν τιμές της συνάρτησης σφάλματος διασταυρωμένης εντροπίας.

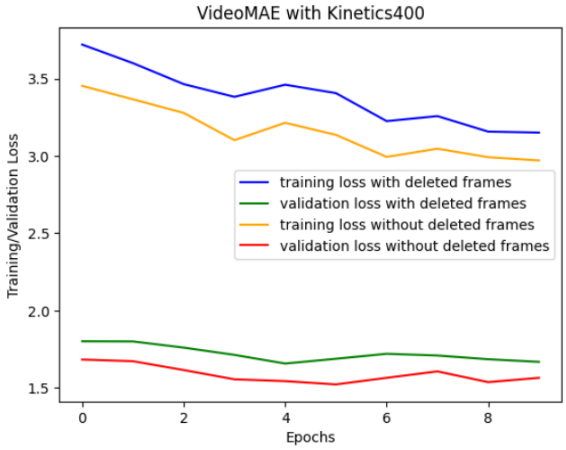
Πειράματα με το Kinetics 400



Εικόνα 5.1: *Κοινό διάγραμμα της ορθότητας πρώτης πρόβλεψης και της ορθότητας των πέντε πρώτων προβλέψεων συναρτήσει των εποχών εκπαίδευσης για το VideoMAE στο Kinetics 400.*

Όπως είναι αναμενόμενο, η τιμή της μετρικής Top-5 accuracy είναι ψηλότερη αυτής της Top-1 accuracy και οι γραφικές παραστάσεις των πειραμάτων μετά τη διαγραφή των καρέ, έχουν χρώμα μπλε και πράσινο, είναι χαμηλότερα από αυτές των πειραμάτων χωρίς επεξεργασία. Παρατηρούμε, επίσης, πως τείνουν να αυξάνονται οι τιμές των μετρικών με την πάροδο του χρόνου, αν και φαίνεται μία απότομη μείωση κατά τη 10η εποχή στην τιμή της ορθότητας των πρώτων πέντε προβλέψεων. Οι επιδόσεις του μοντέλου σε σύνολο επαλήθευσης και εκπαίδευσης είναι χαμηλότερες από αυτές στο σύνολο δοκιμών, γεγονός που σημαίνει ότι η εκπαίδευση είναι επιτυχής και μαθαίνει να γενικεύει τα συμπεράσματά του το μοντέλο και σε ποικίλα δεδομένα εισόδου. Σε πρώτη φάση εντοπίζουμε πως το σφάλμα στο στάδιο της εκπαίδευσης είναι και στις δύο εκδοχές του πειράματος μεγαλύτερο από το σφάλμα στο στάδιο της επαλήθευσης. Αυτό οφείλεται γενικά σε δύο παράγοντες:

- Η επαλήθευση γίνεται μετά τη διαδικασία εκπαίδευσης σε κάθε βήμα, οπότε λογικό το σφάλμα μετά την εκπαίδευση σε ένα σύνολο δεδομένων να είναι χαμηλότερο από το σφάλμα κατά τη διάρκεια της εκπαίδευσης.
- Προστίθενται επίπεδα κανονικοποίησης στο δίκτυο κατά τη διαδικασία της εκπαίδευσης, τα οποία συμβάλλουν στην αποφυγή της υπερπροσαρμογής στα δεδομένα. Αυτά



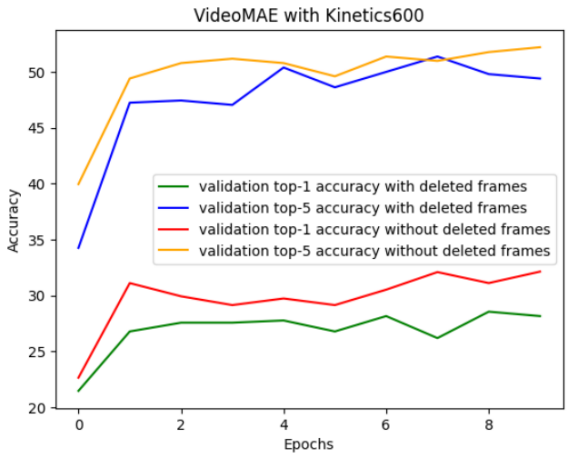
Εικόνα 5.2: *Κοινό διάγραμμα σφάλματος εκπαίδευσης και σφάλματος επαλήθευσης συναρτήσει των εποχών εκπαίδευσης για το VideoMAE στο Kinetics 400.*

τα επίπεδα, όμως, έχουν ως επίπτωση αγνόηση κάποιων δεδομένων και άρα χαμηλότερο σκορ κατά την εκπαίδευση. Αντιθέτως, κατά την επαλήθευση δεν γίνεται κανονικοποίηση του συνόλου δεδομένων.

Πάντως βλέπουμε πως μειώνεται συνεχώς το σφάλμα εκπαίδευσης. Πιθανότατα μετά από έναν αρκετά μεγάλο αριθμό εποχών αυτό να συναντούσε την τιμή σφάλματος στο σύνολο επαλήθευσης, στο οποίο δεν παρατηρούνται τόσο μεγάλες αλλαγές.

Τέλος, προφανώς και σε αυτό το διάγραμμα, το μοντέλο στο πειράματα με τα βίντεο, που έχουν γίνει αντικείμενο προεπεξεργασίας, έχει χειρότερη επίδοση, εφόσον οι γραφικές παραστάσεις που αντιστοιχούν σε αυτά (μπλε και πράσινο χρώμα) είναι ψηλότερα από τις άλλες (κίτρινο και κόκκινο αντίστοιχα).

Πειράματα με το Kinetics 600

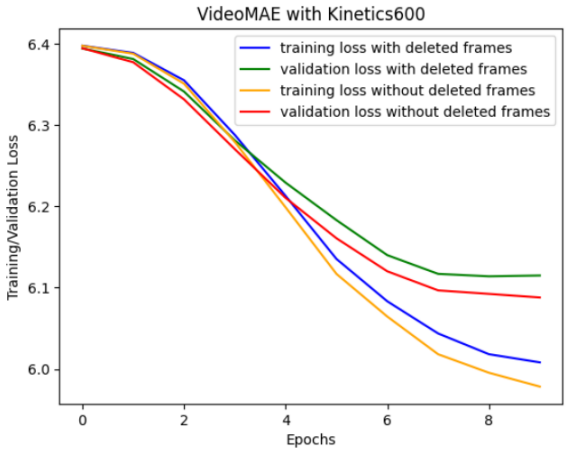


Εικόνα 5.3: *Κοινό διάγραμμα της ορθότητας πρώτης πρόβλεψης και της ορθότητας των πέντε πρώτων προβλέψεων συναρτήσει των εποχών εκπαίδευσης για το VideoMAE στο Kinetics 600.*

Το συγκεκριμένο διάγραμμα εμφανίζει μεγαλύτερο ενδιαφέρον από το αντίστοιχο για το Kinetics 600, εφόσον σε αυτό μπορούμε να δούμε τη βελτίωση της ορθότητας προβλέψεων

του μοντέλου με γρήγορο ρυθμό. Σε σχέση με το αντίστοιχο διάγραμμα για το Kinetics 400, παρατηρούμε μεγαλύτερες αλλαγές, εφόσον στο Kinetics 600 δεν έχει προεκπαιδευθεί το VideoMAE και για αυτό δεν έχει εξ αρχής καλή επίδοση σε αυτό.

Για άλλη μία φορά παρατηρείται από τη σχετική θέση των αντίστοιχων γραφικών παραστάσεων η καλύτερη επίδοση του μοντέλου στα πειράματα με τα αυθεντικά δεδομένα . Επίσης, σε σχέση με τους πίνακες παραπάνω, που συγκεντρώνουν τις επιδόσεις κάθε μοντέλου σε κάθε σύνολο δεδομένων, φτάνουν μία χαμηλότερη τιμή ορθότητας, παρ' όλα αυτά σημαντικά υψηλότερη από την τιμή ορθότητας κατά την εποχή 0 της εκπαίδευσης. Το συ-



Εικόνα 5.4: *Κοινό διάγραμμα σφάλματος εκπαίδευσης και σφάλματος επαλήθευσης συναρτήσει των εποχών εκπαίδευσης για το VideoMAE στο Kinetics 600.*

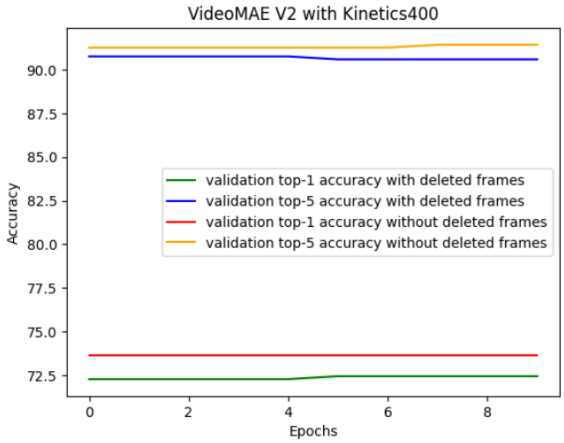
γκεκριμένο διάγραμμα, σε αντίθεση πάλι με το αντίστοιχο για το Kinetics 400, παρουσιάζει σημαντικές μεταβολές. Ειδικότερα, παρατηρείται σταδιακή μείωση του σφάλματος όσο προχωράνε οι εποχές εκπαίδευσης. Επιπλέον σε αυτήν την περίπτωση το σφάλμα επαλήθευσης είναι μεγαλύτερο από το σφάλμα εκπαίδευσης, το οποίο μάλλον οφείλεται στην υπερπροσαρμογή των δεδομένων στις νέες κλάσεις δράσεων, στην αναγνώριση των οποίων δεν έχει προεκπαιδευθεί το μοντέλο. Το VideoMAE τώρα μαθαίνει να κατηγοριοποιεί τις νέες κλάσεις και δεν παρέχεται αρκετά μεγάλος αριθμός δεδομένων στο σύνολο εκπαίδευσης, ώστε να αποφευχθεί η υπερπροσαρμογή.

5.2.2 VideoMAE V2

Παρατίθενται ξανά το κοινό διάγραμμα σφάλματος εκπαίδευσης και σφάλματος επαλήθευσης συναρτήσει των εποχών εκπαίδευσης του μοντέλου καθώς και το κοινό διάγραμμα της ορθότητας πρώτης πρόβλεψης και της ορθότητας των πέντε πρώτων προβλέψεων στο σύνολο επαλήθευσης συναρτήσει των εποχών εκπαίδευσης. Αυτό γίνεται για τα αποτελέσματα και από τις δύο εκδοχές του πειράματος και για τα δύο σύνολα δεδομένων.

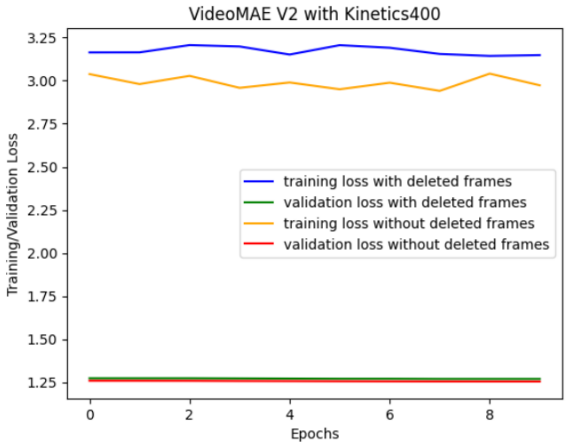
Πειράματα με το Kinetics 400

Το διάγραμμα αυτό είναι αρκετά στατικό, υπό την έννοια ότι παρατηρούμε ελάχιστες μεταβολές, και συγκεκριμένα αυξήσεις, των τιμών των δύο μετρικών στις δύο εκδοχές του πειράματος. Αυτό βασίζεται στο γεγονός ότι το μοντέλο έχει προεκπαιδευθεί στο Kinetics



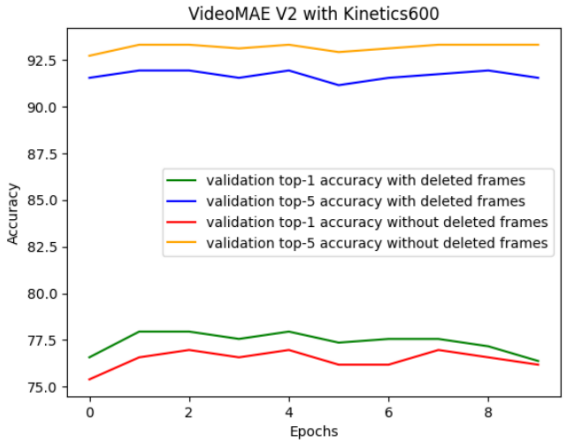
Εικόνα 5.5: *Κοινό διάγραμμα της ορθότητας πρώτης πρόβλεψης και της ορθότητας των πέντε πρώτων προβλέψεων συναρτήσει των εποχών εκπαίδευσης για το VideoMAE V2 στο Kinetics 400.*

400 και επομένως έχει από την εποχή 0 πολύ καλή επίδοση στην πρόβλεψη της σωστής κλάσης. Όπως σε όλα τα πειράματα, οι τιμές των μετρικών είναι καλύτερες (χαμηλότερο σφάλμα, υψηλότερη ορθότητα), όταν έχουμε δώσει ως είσοδο στο μοντέλο τα δεδομένα χωρίς επεξεργασία. Η από κοινού αναπαράσταση του σφάλματος επαλήθευσης και σφάλματος εκ-



Εικόνα 5.6: *Κοινό διάγραμμα σφάλματος εκπαίδευσης και σφάλματος επαλήθευσης συναρτήσει των εποχών εκπαίδευσης για το VideoMAE V2 στο Kinetics 400.*

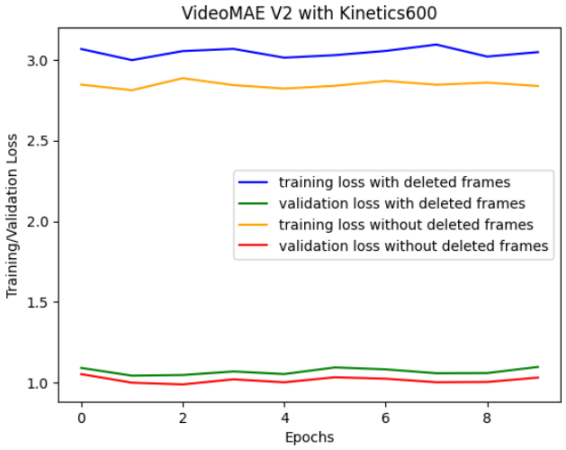
παίδευσης παρουσιάζει ελάχιστες μεταβολές όπως και το προηγούμενο διάγραμμα. Μάλιστα, οι τιμές για τη συνάρτηση σφάλματος επαλήθευσης για τις δύο εκδοχές του πειράματος είναι ακριβώς οι ίδιες, γεγονός που οφείλεται στη μικρή ακρίβεια δεκαδικών ψηφίων, με την οποία έγινε η καταγραφή των τιμών. Όσον αφορά το σφάλμα εκπαίδευσης, αυτό παρουσιάζει μία μικρή αυξομείωση και τελικά σταθεροποίηση γύρω από την αρχική τιμή, όπως αναμένεται σε σύνολο δεδομένων, όπου έχουμε από την αρχή πολύ ορθές προβλέψεις. Υπογραμμίζεται ότι το σφάλμα επαλήθευσης είναι χαμηλότερο από το σφάλμα εκπαίδευσης.



Εικόνα 5.7: *Κοινό διάγραμμα της ορθότητας πρώτης πρόβλεψης και της ορθότητας των πέντε πρώτων προβλέψεων συναρτήσει των εποχών εκπαίδευσης για το VideoMAE V2 στο Kinetics 600.*

Πειράματα με το Kinetics 600

Εφόσον το VideoMAE V2 αποτελεί τη μόνη αρχιτεκτονική δικτύου, που έχει ήδη εκπαιδευθεί με βάση το Kinetics 600, είναι το μόνο από τα τρία μοντέλα, που δεν παρουσιάζει απότομη αλλαγή των τιμών ορθότητας πρόβλεψης με την πάροδο των εποχών εκπαίδευσης. Παρατηρείται μόνο μία ελάχιστη αύξηση του Top-5 accuracy και στις δύο εκδοχές του πειράματος, ενώ μία αυξομείωση παρατηρείται στο Top-1 accuracy. Τέλος, ενδιαφέρον εμφανίζει η σύγκλιση της τιμής της μετρικής Top-1 accuracy για τις δύο εκδοχές του πειράματος μετά το πέρας της τελευταίας εποχής εκπαίδευσης. Ίσως βέβαια και σε αυτήν την περίπτωση να μην έχουν υπολογιστεί με μεγάλη ακρίβεια τα δεκαδικά ψηφία των δύο τιμών. Παρομοίως με το διάγραμμα, που αφορά την ορθότητα του μοντέλου στο Kinetics 600, και



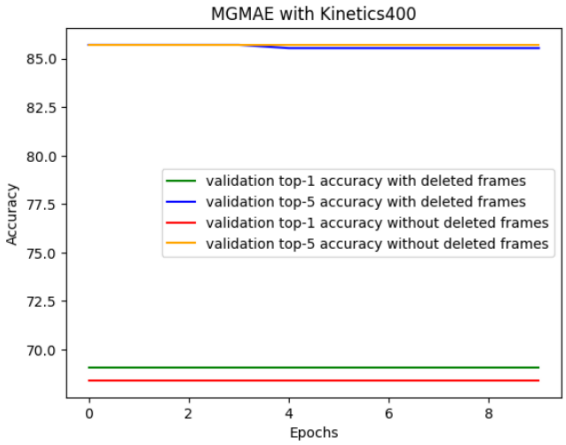
Εικόνα 5.8: *Κοινό διάγραμμα σφάλματος εκπαίδευσης και σφάλματος επαλήθευσης συναρτήσει των εποχών εκπαίδευσης για το VideoMAE V2 στο Kinetics 600.*

το διάγραμμα που αναπαριστά τις τιμές των σφαλμάτων κατά τη διάρκεια της εκπαίδευσης του μοντέλου δεν εμφανίζει αξιοσημείωτες αλλαγές στην τιμή των συναρτήσεων και φαίνεται μία σταθεροποίηση γύρω από τις τιμές 3 για το σφάλμα εκπαίδευσης και 1 για το σφάλμα επαλήθευσης.

5.2.3 MGMAE

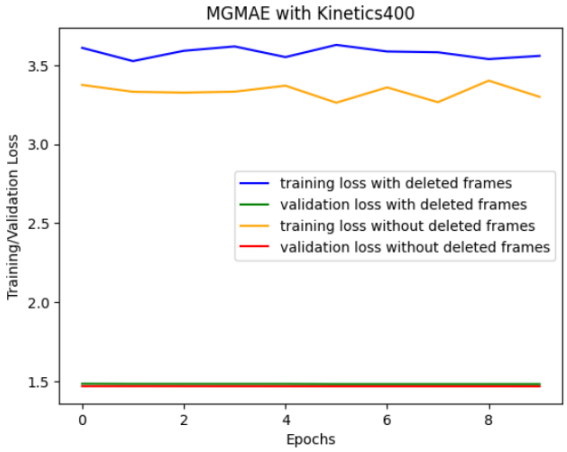
Και για το τρίτο μοντέλο, με το οποίο πειραματιστήκαμε, παρουσιάζουμε από τη μία τα κοινά διαγράμματα των συναρτήσεων σφάλματος κατά την εκπαίδευση και κατά την επαλήθευση, και από την άλλη τα κοινά διαγράμματα ορθότητας πρώτης πρόβλεψης και ορθότητας πέντε πρώτων προβλέψεων για τα Kinetics 400 και Kinetics 600. Όπως και στα προηγούμενα μοντέλα φαίνονται τα αποτελέσματα για όλες τις εκδοχές των πειραμάτων.

Πειράματα με το Kinetics 400



Εικόνα 5.9: *Κοινό διάγραμμα της ορθότητας πρώτης πρόβλεψης και της ορθότητας των πέντε πρώτων προβλέψεων συναρτήσει των εποχών εκπαίδευσης για το MGMAE στο Kinetics 400.*

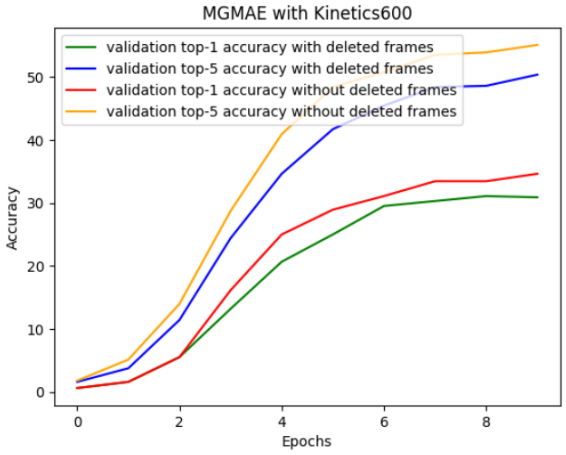
Στο διάγραμμα αυτό δεν παρατηρείται σχεδόν καμία μεταβολή των αρχικών τιμών των μετρικών Top-1 accuracy και Top-5 accuracy. Συγκεκριμένα, μόνο το Top-5 accuracy για τα πειράματα με τη διαγραφή του 10% των καρέ από τα βίντεο εισόδου μειώνεται ελάχιστα κατά την 4η εποχή εκπαίδευσης τουMGMAE, ενώ η αντίστοιχη μετρική για τα πειράματα με τα αυθεντικά δεδομένα κρατάει σταθερή τιμή λίγο πάνω από 85%. Οι τιμές του Top-1 accuracy παραμένουν σταθερές λίγο πιο κάτω από το 70%. Βλέπουμε πως και στην περίπτωση της



Εικόνα 5.10: *Κοινό διάγραμμα σφάλματος εκπαίδευσης και σφάλματος επαλήθευσης συναρτήσει των εποχών εκπαίδευσης για το MGMAE στο Kinetics 400.*

αναπαράστασης των τιμών των σφαλμάτων δεν σημειώνονται βασικές μεταβολές. Ιδιαίτερα, το σφάλμα επαλήθευσης, που είναι το μικρότερο από αυτό της εκπαίδευσης, παραμένει και στα δύο πειράματα σταθερό και ίσο με 1.5 περίπου. Η τιμή του σφάλματος εκπαίδευσης στις δύο εκδοχές του πειράματος παρουσιάζει ελάχιστες αυξομειώσεις γύρω από το 3.5, με το σφάλμα στην περίπτωση που έχουμε αφαιρέσει 10% των καρέ να είναι φυσικά υψηλότερο.

Πειράματα με το Kinetics 600



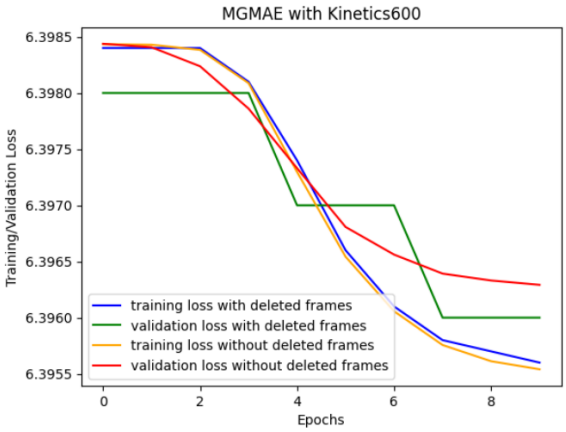
Εικόνα 5.11: *Κοινό διάγραμμα της ορθότητας πρώτης πρόβλεψης και της ορθότητας των πέντε πρώτων προβλέψεων συναρτήσει των εποχών εκπαίδευσης για το MGMAE στο Kinetics 600.*

Το ανωτέρω διάγραμμα παρουσιάζει ιδιαίτερο ενδιαφέρον για αρκετούς λόγους. Αρχικά, παρατηρούμε έντονη αλλαγή, και συγκεκριμένα αύξηση, στην τιμή της ορθότητας πρόβλεψης του μοντέλου στις δύο εκδοχές του πειράματος. Με γρήγορο ρυθμό, ειδικά ανάμεσα στις εποχές 2 και 4, και στα δύο πειράματα αυξάνεται η τιμή της ορθότητας πρώτης πρόβλεψης του MGMAE και φθάνει το 30% περίπου, ενώ η τιμή της ορθότητας των πέντε πρώτων προβλέψεων προσεγγίζει το 50%. αυτό φυσικά αποτελεί συνέπεια του ότι το MGMAE δεν έχει προεκπαιδευθεί στο Kinetics 600, οπότε δεν γνωρίζει πώς να κατηγοριοποιεί όλες τις κλάσεις του συνόλου δεδομένων από την αρχή. Απεναντίας απαιτείται η εκπαίδευσή του, για να βελτιώσει την επίδοσή του στη σωστή πρόβλεψη των δράσεων σε κάθε βίντεο.

Πρέπει να σημειωθεί πως, παρ'όλο που και το VideoMAE δεν ήταν προεκπαιδευμένο στο Kinetics 600, αυτό δεν εμφάνιζε τιμή ορθότητας ίση με 0% και στις δύο εκδοχές του πειράματος. Ξεκινούσε από μία χαμηλή τιμή (περίπου 20% για το Top-1 accuracy και 35%-40% για το Top-5 accuracy) και αυξανόταν σταδιακά. Στην περίπτωση του MGMAE, όμως, στην εποχή 0 έχουμε αρχική τιμή των μετρικών ακρίβειας πρόβλεψης ίση με 0% και στα δύο πειράματα. Μέχρι την εποχή 9 το μοντέλο καταφέρνει να βελτιώσει την επίδοσή του, όπως προαναφέρθηκε, αλλά φαίνεται να μην έχει κανένα στοιχείο για το σύνολο δεδομένων, στο οποίο ρυθμίζεται με ακρίβεια, κατά το αρχικό στάδιο της εκπαίδευσής του.

Αυτό πιθανότατα οφείλεται, όπως σχολιάστηκε και κατά την παρουσίαση των πινάκων με συγκεντρωμένα τα αποτελέσματα, στη δυσκολία που αντιμετωπίζει ο το μοντέλο κατά την κατασκευή μάσκας απόκρυψης με βάση την οπτική ροή στα βίντεο, εφόσον με την απόκρυψη 10% των καρέ μειώνεται η διαθέσιμη οπτική πληροφορία. Γίνεται δύσκολη η

παρακολούθηση της εξέλιξης της κίνησης από καρέ σε καρέ, όταν ορισμένα από αυτά έχουν διαγραφεί. Μοιάζει με το διάγραμμα απεικόνισης των συναρτήσεων σφάλματος συναρτήσει



Εικόνα 5.12: *Κοινό διάγραμμα σφάλματος εκπαίδευσης και σφάλματος επαλήθευσης συναρτήσει των εποχών εκπαίδευσης για το MGMAE στο Kinetics 600.*

των εποχών εκπαίδευσης στην περίπτωση ρύθμισης με ακρίβεια του VideoMAE στο Kinetics 600. Εδώ η μείωση του σφάλματος είναι σχεδόν ανεπαίσθητη, από το 6.3985 περίπου πέφτει στο 6.3955 το σφάλμα εκπαίδευσης για παράδειγμα.

Μία ομοιότητα με το προαναφερθέν διάγραμμα για το VideoMAE είναι ότι το σφάλμα επαλήθευσης παρουσιάζει υψηλότερες τιμές στο τέλος της εκπαίδευσης από το σφάλμα εκπαίδευσης. Και σε αυτήν την περίπτωση, ερμηνεύουμε αυτήν παρατήρηση ως απότοκο της υπερπροσαρμογής του MGMAE στα λίγα αντιπροσωπευτικά δείγματα από τις νέες κλάσεις δεδομένων κατά το στάδιο της εκπαίδευσης. Πιστεύουμε, αν εκπαιδευόταν για μεγαλύτερο χρονικό διάστημα το δίκτυο, δηλαδή σε μεγαλύτερο σύνολο δεδομένων, το πρόβλημα αυτό θα μπορούσε να ξεπεραστεί και οι τιμές των δύο σφαλμάτων θα συνέκλιναν.

Τέλος, σχολιάζουμε το γεγονός ότι η γραφική παράσταση της συνάρτησης σφάλματος επαλήθευσης στο πείραμα με την απόκρυψη καρέ από τα βίντεο εισόδου μοιάζει με βηματοσυνάρτηση. Για άλλη μία φορά, αυτό μπορούμε να το αποδώσουμε στη μικρή ακρίβεια δεκαδικών ψηφίων, με την οποία έγινε η καταμέτρηση των αποτελεσμάτων. Έτσι φαίνονται κοντινές τιμές της συνάρτησης σφάλματος ίδιες, ενώ πιθανότατα έχουν ελάχιστες διαφορές.

Κεφάλαιο 6

Συμπεράσματα και Μελλοντικές Επεκτάσεις

Στο κεφάλαιο αυτό εκθέτουμε τα συμπεράσματά μας με βάση τα αποτελέσματα των πειραμάτων, όπως αυτά παρουσιάστηκαν στο προηγούμενο κεφάλαιο.

6.1 Συμπεράσματα

Πρώτα απ' όλα, με βάση τις τιμές των μετρικών σφάλματος και ορθότητας, που παρατηρήσαμε στην εκδοχή του πειράματος με τα αυθεντικά δεδομένα, μπορούμε να πούμε πως και τα τρία μοντέλα αποτελούν αρχιτεκτονικές δικτύων βαθιάς μάθησης τελευταίας τεχνολογίας με πολύ καλή επίδοση στα δύο σύνολα δεδομένων. Επίσης χρειάστηκε ένας μικρός αριθμός εποχών, για να προσεγγίσουν τη βέλτιστη τιμή ορθότητας στα σύνολα δεδομένων, στα οποία είχε προεκπαιδευθεί το κάθε μοντέλο. Συγκεκριμένα, το VideoMAE καταφέρνει να απέχει μετά από 10 εποχές εκπαίδευσης μόνο 4% από την επίδοση που σημειώνουν στο άρθρο, όπου παρουσιάστηκε το μοντέλο, ενώ το MGMAE καταφέρνει να προσεγγίσει την επίδοση, που αναφέρουν οι δημιουργοί του, απέχοντας μόνο 6% από αυτή. Το VideoMAE V2, στο οποίο έχει γίνει προεκπαίδευση και στα δύο σύνολα δεδομένων, μετά το πέρας των 10 εποχών εκπαίδευσης σημειώνει τιμή ορθότητας πρώτης πρόβλεψης 9% μικρότερη από τη βέλτιστη στο Kinetics 400 και 6% μικρότερη από τη βέλτιστη στο Kinetics 600.

Αναμφίβολα πάντως το VideoMAE V2 είναι το καλύτερο μοντέλο και στις δύο εκδοχές του πειράματος, πετυχαίνοντας την υψηλότερη τιμή ορθότητας πρόβλεψης και στα δύο σύνολα δεδομένων. Σίγουρα, στην επιτυχία του μοντέλου στο πείραμα με τη διαγραφή του 10% των καρέ των βίντεο εισόδου στο Kinetics 600 συμβάλλει το γεγονός ότι έχει ήδη εκπαιδευθεί σε αυτό. Επίσης παρατηρούμε ότι το ποσοστό απώλειας στην τιμή της ορθότητας πρόβλεψης είναι πολύ μικρότερο από το ποσοστό μείωσης του χρόνου, που χρειάζεται το μοντέλο για να επεξεργαστεί τα δεδομένα εισόδου και να παράγει τις προβλέψεις του. Ειδικότερα, στο Kinetics 400 η τιμή του Top-1 accuracy παρουσιάζει πτώση 0.5% και η τιμή του Top-5 accuracy πέφτει κατά 1.2%, αλλά ο χρόνος εκπαίδευσης μειώνεται κατά 9%. Επίσης, στο Kinetics 600 η τιμή του Top-1 accuracy μειώνεται κατά λιγότερο από 0.1% και η τιμή του Top-5 accuracy εμφανίζει ελάττωση της τάξης του 2.2%, τη στιγμή που ο χρόνος εκπαίδευσης μειώνεται κατά 12%. Συνεπώς, με την απόκρυψη τμημάτων των δεδομένων εισόδου στη διάσταση του χρόνου μειώνεται το υπολογιστικό κόστος σε μεγαλύτερο βαθμό από ότι μειώνεται η ορθότητα των προβλέψεων.

Από την άλλη, πολύ ευαίσθητο στη διαγραφή ενός 10% των καρέ από τα βίντεο εισόδου

στο Kinetics 600 φάνηκε το MGMAE. Όπως αναλύσαμε εκτενώς στο προηγούμενο κεφάλαιο, αυτό αποδίδεται κατά κύριο λόγο στο πόσο επηρεάζεται από την έλλειψη χρονικής πληροφορίας ο αλγόριθμος υπολογισμού της οπτικής ροής, με βάση τον οποίο κατασκευάζεται η μάσκα απόκρυψης. Στην εκδοχή του πειράματος με τα αυθεντικά δεδομένα, όμως, καταφέρνει να ξεπεράσει τις επιδόσεις του VideoMAE στο σύνολο δεδομένων, όπου δεν έχουν προεκπαιδευθεί τα δύο μοντέλα. Στο Kinetics 400 οι τιμές ορθότητας πρόβλεψης του MGMAE υπολείπονται ελάχιστα, κατά λιγότερο από 2%, αυτές του VideoMAE.

6.2 Μελλοντικές Προσεγγίσεις

Όπως είδαμε, ειδικά στην περίπτωση του VideoMAE V2, η απόκρυψη τμημάτων των δεδομένων εισόδου όχι μόνο στη διάσταση του χώρου αλλά και στη διάσταση του χρόνου μπορεί να επιφέρει σημαντικές βελτιώσεις όσον αφορά το υπολογιστικό κόστος που απαιτεί το μοντέλο για την εκπαίδευσή του, χωρίς σοβαρές απώλειες στην επίδοσή του. Αν συνδυαστεί λοιπόν με κατάλληλο τρόπο η απόκρυψη με χρήση μάσκας, που θα δημιουργείται όχι με τυχαίο τρόπο, όπως σε αυτήν τη διπλωματική, αλλά λαμβάνοντας υπόψη πόσο σημαντικό για την αναπαράσταση της δράσης του βίντεο είναι κάθε καρέ, στο χρονικό επίπεδο με απόκρυψη περιττών τμημάτων κάθε καρέ, μπορεί να επιτευχθεί μεγάλη μείωση της απαίτησης σε χρόνο και υπολογιστικούς πόρους των μοντέλων βαθιάς μάθησης, χωρίς ταυτόχρονα να ελαττώνεται η ορθότητα πρόβλεψης.

Αυτό το θεωρούμε πολύ σημαντικό στη σύγχρονη περίοδο ανάπτυξης μοντέλων βαθιάς μάθησης, εφόσον τα νέα μοντέλα, ειδικά αυτά που χρησιμοποιούν μετασχηματιστές, όπως οι αποκρύπτοντες αυτοκωδικοποιητές, έχουν να εκπαιδεύσουν εκατομμύρια ή δισεκατομμύρια παραμέτρους. Επιπλέον, στον τομέα της αναγνώρισης δράσης σε βίντεο, όπου πρέπει να γίνονται αντικείμενο επεξεργασίας τα δεδομένα σε τέσσερις διαστάσεις, τις τρεις χωρικές και αυτή του χρόνου, αυξάνεται η ανάγκη για υπολογιστική ισχύ. Σε αυτά τα προβλήματα συνήθως εισάγονται γιγαντιαία σύνολα δεδομένων, ώστε να μπορέσουν μέσω της εκπαίδευσής τους σε αυτά τα μοντέλα μηχανικής μάθησης να γενικεύσουν την ικανότητά τους να ταξινομήσουν τη δραστηριότητα, που περιέχει ένα βίντεο.

Επίσης ενδιαφέρον θα παρουσίαζε η ιδέα εισαγωγής της μάσκας απόκρυψης στο ίδιο το μοντέλο, ώστε να μην αποτελεί ένα στάδιο της προεπεξεργασίας των δεδομένων, αλλά μία υπερπαράμετρο αυτού. Με αυτόν τον τρόπο, θα μπορούσε ανάλογα με το σκορ πρόβλεψης του μοντέλου να βρεθεί το βέλτιστο ποσοστό κάλυψης των καρέ κάθε βίντεο. Επιπροσθέτως, μεγάλη σημασία παρουσιάζει σήμερα η δυνατότητα επεξηγησιμότητας στα μοντέλα βαθιάς μάθησης, οπότε με την εφαρμογή μάσκας απόκρυψης διαφορετικών καρέ κάθε φορά και καταγραφής της αντίστοιχης επίδοσης του μοντέλου, μπορούν να βρεθούν τα πιο σημαντικά για την ταξινόμηση της δράσης του βίντεο στη σωστή κλάση καρέ. Αυτό έχει πολλές εφαρμογές, όπως για παράδειγμα την εύρεση του τμήματος ενός βίντεο, που συγκεντρώνει την περισσότερη πληροφορία για το τι γίνεται σε αυτό από το αλγόριθμο του YouTube, ώστε να επιλεχθεί ως thumbnail.

Βιβλιογραφία

[1] *Kismet*. http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html. Date of access: 10-4-2024.

[2] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang και Tieyan Liu. *On layer normalization in the transformer architecture*. *International Conference on Machine Learning*, σελίδες 10524–10533. PMLR, 2020.

[3] Earl J Kirkland και Earl J Kirkland. *Bilinear interpolation*. *Advanced computing in electron microscopy*, σελίδες 261–263, 2010.

[4] Jan Held, Anthony Cioppa, Silvio Giancola, Abdullah Hamdi, Bernard Ghanem και Marc Van Droogenbroeck. *Vars: Video assistant referee system for automated soccer decision making from multiple views*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 5085–5096, 2023.

[5] *Artificial Intelligence*. https://en.wikipedia.org/wiki/Artificial_intelligence#Philosophy. Date of access: 10-4-2024.

[6] John H Holland. *Genetic algorithms*. *Scientific american*, 267(1):66–73, 1992.

[7] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing Long Han και Yang Tang. *A brief overview of ChatGPT: The history, status quo and potential future development*. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136, 2023.

[8] Antonios Mamalakis, Imme Ebert-Uphoff και Elizabeth A Barnes. *Explainable artificial intelligence in meteorology and climate science: Model fine-tuning, calibrating trust and learning new science*. *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, σελίδες 315–339. Springer, 2020.

[9] *Categories of Machine Learning*. https://www.geeksforgeeks.org/types-of-machine-learning/. Date of access: 17-4-2024.

[10] Herve Abdi. *A neural network primer*. *Journal of Biological Systems*, 2(03):247–281, 1994.

[11] Gao Daqi και Ji Yan. *Classification methodologies of multilayer perceptrons with sigmoid activation functions*. *Pattern Recognition*, 38(10):1469–1482, 2005.

[12] Babak Zamanlooy και Mitra Mirhassani. *Efficient VLSI implementation of neural networks with hyperbolic tangent activation function*. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 22(1):39–48, 2013.

[13] Chaity Banerjee, Tathagata Mukherjee και Eduardo Pasiliao Jr. *An empirical study on generalizations of the ReLU activation function*. *Proceedings of the 2019 ACM Southeast Conference*, σελίδες 164–167, 2019.

[14] Peter Norvig Stuart Russel. *Τεχνητή Νοημοσύνη Μία σύγχρονη προσέγγιση*. Κλειδάριθμος, Αθήνα, 4η έκδοση, 2021.

[15] Shun ichi Amari. *Backpropagation and stochastic gradient descent method*. *Neurocomputing*, 5(4-5):185–196, 1993.

[16] Keiron O'shea και Ryan Nash. *An introduction to convolutional neural networks*. *arXiv preprint arXiv:1511.08458*, 2015.

[17] Paraskevi Antonia Theofilou, Georgios Tsatiris και Stefanos Kollias. *Automatic assessment of Parkinson's patients' dyskinesia using non-invasive machine learning methods*. *2022 International Conference on Interactive Media, Smart Systems and Emerging Technologies (IMET)*, σελίδες 1–4. IEEE, 2022.

[18] Wenpeng Yin, Katharina Kann, Mo Yu και Hinrich Schütze. *Comparative study of CNN and RNN for natural language processing*. *arXiv preprint arXiv:1702.01923*, 2017.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser και Illia Polosukhin. *Attention is all you need*. *Advances in neural information processing systems*, 30, 2017.

[20] Zhigang Dai, Bolun Cai, Yugeng Lin και Junying Chen. *Unsupervised pre-training for detection transformers*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár και Ross Girshick. *Masked autoencoders are scalable vision learners*. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, σελίδες 16000–16009, 2022.

[22] Sarath Chandar AP, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar και Amrita Saha. *An autoencoder approach to learning bilingual word representations*. *Advances in neural information processing systems*, 27, 2014.

[23] Junhai Zhai, Sufang Zhang, Junfen Chen και Qiang He. *Autoencoder and its various variants*. *2018 IEEE international conference on systems, man, and cybernetics (SMC)*, σελίδες 415–419. IEEE, 2018.

[24] Ian Goodfellow, Yoshua Bengio και Aaron Courville. *Deep learning*. MIT press, 2016.

[25] Jacob Devlin, Ming Wei Chang, Kenton Lee και Kristina Toutanova. *Bert: Pre-training of deep bidirectional transformers for language understanding*. *arXiv preprint arXiv:1810.04805*, 2018.

[26] Qiu S, Wang Z, Zhao H και Hu H. *Using Distributed Wearable Sensors to Measure and Evaluate Human Lower Limb Motions*. *IEEE Transactions on Instrumentation and Measurement*, 65(4):939 - 950, 2016.

[27] Sebastijan Sprager και Matjaz B Juric. *Inertial sensor-based gait recognition: A review*. *Sensors*, 15(9):22089–22127, 2015.

[28] S Patel, H Park, P Bonato, L Chan και M Rodgers. *" A review of wearable sensors and systems with application in rehabilitation," Journal of NeuroEngineering and Rehabilitation, vol. 9, no. 1, pp. 1-17, Apr*. 2012.

[29] Irina Spulber, Enrica Papi, Y M Chen, Salzitsa Anastasova-Ivanova, J Bergmann, Pantelis Georgiou και Alison H McGregor. *Development of a wireless multi-functional body sensing platform for smart garment integration*. *2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings*, σελίδες 157–160. IEEE, 2014.

[30] Kellen Garrison Cresswell, Yongyun Shin και Shanshan Chen. *Quantifying variation in gait features from wearable inertial sensors using mixed effects models*. *Sensors*, 17(3):466, 2017.

[31] Weijun Tao, Tao Liu, Rencheng Zheng και Hutian Feng. *Gait analysis using wearable sensors*. *Sensors*, 12(2):2255–2283, 2012.

[32] Marcello Fusca, Francesco Negrini, Paolo Perego, Luciana Magoni, Franco Molteni και Giuseppe Andreoni. *Validation of a wearable IMU system for gait analysis: Protocol and application to a new system*. *Applied Sciences*, 8(7):1167, 2018.

[33] *What is Motion Capture?* https://www.audiomotion.com/blog/what-is-motion-capture.html. Date of access: 31-3-2024.

[34] *What is motion capture and how does it work?* https://www.mo-sys.com/news/what-is-motion-capture-and-how-does-it-work/. Date of access: 31-3-2024.

[35] Linwei Fan, Fan Zhang, Hui Fan και Caiming Zhang. *Brief review of image denoising techniques*. *Visual Computing for Industry, Biomedicine, and Art*, 2(1):7, 2019.

[36] Rafael C Gonzalez. *Digital image processing*. Pearson education india, 2009.

[37] AL AMEEN Zohair, AL AMEEN Shamil και Ghazali Sulong. *Latest methods of image enhancement and restoration for computed tomography: a concise review*. *Applied Medical Informatics*, 36(1):1–12, 2015.

[38] Anil K Jain. *Fundamentals of digital image processing*. Prentice-Hall, Inc., 1989.

[39] Jacob Benesty, Jingdong Chen και Yiteng Huang. *Study of the widely linear Wiener filter for noise reduction*. *2010 IEEE international conference on acoustics, speech and signal processing*, σελίδες 205–208. IEEE, 2010.

[40] AN Venetsanopoulos και I Pitas. *Nonlinear digital filters*. *PRINCIPLES AND APPLICATIONS*. *Kluwer*, 1990.

[41] Ruikang Yang, Lin Yin, Moncef Gabbouj, Jaakko Astola και Yrjö Neuvo. *Optimal weighted median filtering under structural constraints*. *IEEE transactions on signal processing*, 43(3):591–604, 1995.

[42] Stephen Gould, Tianshi Gao και Daphne Koller. *Region-based segmentation and object detection*. *Advances in neural information processing systems*, 22, 2009.

[43] *Thresholding-Based Image Segmentation*. https://www.geeksforgeeks.org/thresholding-based-image-segmentation/. Date of access: 1-4-2024.

[44] Connor Shorten και Taghi M Khoshgoftaar. *A survey on image data augmentation for deep learning*. *Journal of big data*, 6(1):1–48, 2019.

[45] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bartter Haar Romeny, John B Zimmerman και Karel Zuiderveld. *Adaptive histogram equalization and its variations*. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.

[46] Andrzej Maćkiewicz και Waldemar Ratajczak. *Principal components analysis (PCA)*. *Computers & Geosciences*, 19(3):303–342, 1993.

[47] Xiaochen Zheng, Meiqing Wang και Joaquín Ordieres-Meré. *Comparison of data preprocessing approaches for applying deep learning to human activity recognition in the context of industry 4.0*. *Sensors*, 18(7):2146, 2018.

[48] Ervin Sejdić, Igor Djurović και Jin Jiang. *Time–frequency feature representation using energy concentration: An overview of recent advances*. *Digital signal processing*, 19(1):153–183, 2009.

[49] Daniele Ravi, Charence Wong, Benny Lo και Guang Zhong Yang. *Deep learning for human activity recognition: A resource efficient implementation on low-power devices*. *2016 IEEE 13th international conference on wearable and implantable body sensor networks (BSN)*, σελίδες 71–76. IEEE, 2016.

[50] Sean R Eddy. *Hidden markov models*. *Current opinion in structural biology*, 6(3):361–365, 1996.

[51] Zia Moghaddam και Massimo Piccardi. *Training initialization of hidden Markov models in human action recognition*. *IEEE Transactions on Automation Science and Engineering*, 11(2):394–408, 2013.

[52] Hend Basly, Wael Ouarda, Fatma Ezahra Sayadi, Bouraoui Ouni και Adel M Alimi. *CNN-SVM learning approach based human activity recognition*. *Image and Signal Processing: 9th International Conference, ICISP 2020, Marrakesh, Morocco, June 4–6, 2020, Proceedings 9*, σελίδες 271–281. Springer, 2020.

[53] Lei Wang, Yangyang Xu, Jun Cheng, Haiying Xia, Jianqin Yin και Jiaji Wu. *Human action recognition by learning spatio-temporal features with deep neural networks*. *IEEE access*, 6:17913–17922, 2018.

[54] Yong Du, Yun Fu και Liang Wang. *Skeleton based action recognition with convolutional neural network*. *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, σελίδες 579–583. IEEE, 2015.

[55] Raviteja Vemulapalli, Felipe Arrate και Rama Chellappa. *Human action recognition by representing 3d skeletons as points in a lie group*. *Proceedings of the IEEE conference on computer vision and pattern recognition*, σελίδες 588–595, 2014.

[56] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook και Richard Moore. *Real-time human pose recognition in parts from single depth images*. *Communications of the ACM*, 56(1):116–124, 2013.

[57] Zhuang Liu, Hanzi Mao, Chao Yuan Wu, Christoph Feichtenhofer, Trevor Darrell και Saining Xie. *A convnet for the 2020s*. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, σελίδες 11976–11986, 2022.

[58] Brett Koonce και Brett Koonce. *ResNet 50*. *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, σελίδες 63–72, 2021.

[59] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin και David Lopez-Paz. *mixup: Beyond empirical risk minimization*. *arXiv preprint arXiv:1710.09412*, 2017.

[60] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe και Youngjoon Yoo. *Cutmix: Regularization strategy to train strong classifiers with localizable features*. *Proceedings of the IEEE/CVF international conference on computer vision*, σελίδες 6023–6032, 2019.

[61] Ekin D Cubuk, Barret Zoph, Jonathon Shlens και Quoc V Le. *Randaugment: Practical automated data augmentation with a reduced search space*. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, σελίδες 702–703, 2020.

[62] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li και Yi Yang. *Random erasing data augmentation*. *Proceedings of the AAAI conference on artificial intelligence*, τόμος 34, σελίδες 13001–13008, 2020.

[63] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens και Zbigniew Wojna. *Rethinking the inception architecture for computer vision*. *Proceedings of the IEEE conference on computer vision and pattern recognition*, σελίδες 2818–2826, 2016.

[64] Ilya Loshchilov και Frank Hutter. *Decoupled weight decay regularization*. *arXiv preprint arXiv:1711.05101*, 2017.

[65] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto και Hartwig Adam. *Mobilenets: Efficient convolutional neural networks for mobile vision applications*. *arXiv preprint arXiv:1704.04861*, 2017.

[66] François Chollet. *Xception: Deep learning with depthwise separable convolutions*. *Proceedings of the IEEE conference on computer vision and pattern recognition*, σελίδες 1251–1258, 2017.

[67] Dan Hendrycks και Kevin Gimpel. *Gaussian error linear units (gelus)*. *arXiv preprint arXiv:1606.08415*, 2016.

[68] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly και others. *An image is worth 16x16 words: Transformers for image recognition at scale*. *arXiv preprint arXiv:2010.11929*, 2020.

[69] Jiasen Lu, Dhruv Batra, Devi Parikh και Stefan Lee. *Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks*. *Advances in neural information processing systems*, 32, 2019.

[70] Namuk Park και Songkuk Kim. *How do vision transformers work?* *arXiv preprint arXiv:2202.06709*, 2022.

[71] Ran Shao και Xiao Jun Bi. *Transformers meet small datasets*. *IEEE Access*, 10:118454–118464, 2022.

[72] Letitia Parcalabescu. *How do Vision Transformers work? - Paper explained | multi-head self-attention convolutions*.

[73] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles και Hervé Jégou. *Training data-efficient image transformers & distillation through attention*. *International conference on machine learning*, σελίδες 10347–10357. PMLR, 2021.

[74] Dianyuan Han. *Comparison of commonly used image interpolation methods*. *Conference of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)*, σελίδες 1556–1559. Atlantis Press, 2013.

[75] Anqi Mao, Mehryar Mohri και Yutao Zhong. *Cross-entropy loss functions: Theoretical analysis and applications*. *International Conference on Machine Learning*, σελίδες 23803–23828. PMLR, 2023.

[76] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu και Yunhe Wang. *Transformer in transformer*. *Advances in neural information processing systems*, 34:15908–15919, 2021.

[77] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin και Baining Guo. *Swin transformer: Hierarchical vision transformer using shifted windows*. *Proceedings of the IEEE/CVF international conference on computer vision*, σελίδες 10012–10022, 2021.

[78] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang και others. *Spvit: Enabling faster vision transformers via latency-aware soft token pruning*. *European conference on computer vision*, σελίδες 620–640. Springer, 2022.

[79] Hao Yu και Jianxin Wu. *A unified pruning framework for vision transformers*. *Science China Information Sciences*, 66(7):179101, 2023.

[80] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash και Jürgen Gall. *Adaptive token sampling for efficient vision transformers*. *European Conference on Computer Vision*, σελίδες 396–414. Springer, 2022.

[81] Daniel Bolya, Cheng Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer και Judy Hoffman. *Token merging: Your vit but faster*. *arXiv preprint arXiv:2210.09461*, 2022.

[82] Yixing Xu, Chao Li, Dong Li, Xiao Sheng, Fan Jiang, Lu Tian και Ashish Sirasao. *FDViT: Improve the Hierarchical Architecture of Vision Transformer*. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, σελίδες 5950–5960, 2023.

[83] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić και Cordelia Schmid. *Vivit: A video vision transformer*. *Proceedings of the IEEE/CVF international conference on computer vision*, σελίδες 6836–6846, 2021.

[84] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang και Luc Van Gool. *Temporal segment networks: Towards good practices for deep action recognition*. *European conference on computer vision*, σελίδες 20–36. Springer, 2016.

[85] Dirk Weissenborn, Oscar Täckström και Jakob Uszkoreit. *Scaling autoregressive video models*. *arXiv preprint arXiv:1906.02634*, 2019.

[86] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin και Han Hu. *Video swin transformer*. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, σελίδες 3202–3211, 2022.

[87] Junke Wang, Xitong Yang, Hengduo Li, Li Liu, Zuxuan Wu και Yu Gang Jiang. *Efficient video transformers with spatial-temporal token selection*. *European Conference on Computer Vision*, σελίδες 69–86. Springer, 2022.

[88] Matthew Dutson, Yin Li και Mohit Gupta. *Eventful Transformers: Leveraging Temporal Redundancy in Vision Transformers*. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, σελίδες 16911–16923, 2023.

[89] Zhan Tong, Yibing Song, Jue Wang και Limin Wang. *Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training*. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[90] Xinlei Chen, Saining Xie και Kaiming He. *An empirical study of training self-supervised vision transformers*. *Proceedings of the IEEE/CVF international conference on computer vision*, σελίδες 9640–9649, 2021.

[91] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang και Yu Qiao. *Videomae v2: Scaling video masked autoencoders with dual masking*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 14549–14560, 2023.

[92] Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Xiang Wang, Yuehuan Wang, Yiliang Lv, Changxin Gao και Nong Sang. *Mar: Masked autoencoders for efficient action recognition*. *IEEE Transactions on Multimedia*, 2023.

[93] Bingkun Huang, Zhiyu Zhao, Guozhen Zhang, Yu Qiao και Limin Wang. *Mgmae: Motion guided masking for video masked autoencoding*. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, σελίδες 13493–13504, 2023.

[94] Zachary Teed και Jia Deng. *Raft: Recurrent all-pairs field transforms for optical flow*. *Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II 16*, σελίδες 402–419. Springer, 2020.

[95] Christopher Zach, Thomas Pock και Horst Bischof. *A duality based approach for realtime tv-l 1 optical flow*. *Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29*, σελίδες 214–223. Springer, 2007.

[96] Ye Htet, Thi Thi Zin, Pyke Tin, Hiroki Tamura, Kazuhiro Kondo και Etsuo Chosa. *HMM-based action recognition system for elderly healthcare by colorizing depth map*. *International Journal of Environmental Research and Public Health*, 19(19):12055, 2022.

[97] Rimeh Jarray, Ahmed Snoun, Tahani Bouchrika και Olfa Jemai. *Deep human action recognition system for assistance of alzheimer's patients*. *Hybrid Intelligent Systems: 20th International Conference on Hybrid Intelligent Systems (HIS 2020), December 14-16, 2020*, σελίδες 484–493. Springer, 2021.

[98] Xiaoyang Wang, Yilin Wang, Mingjie Zhou, Baobin Li, Xiaoqian Liu και Tingshao Zhu. *Identifying psychological symptoms based on facial movements*. *Frontiers in Psychiatry*, 11:607890, 2020.

[99] Bilge Soran, Ali Farhadi και Linda Shapiro. *Action recognition in the presence of one egocentric and multiple static cameras*. *Computer Vision-ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part V 12*, σελίδες 178–193. Springer, 2015.

[100] Ruihua Yu. *Computer-aided english pronunciation accuracy detection based on lip action recognition algorithm*. *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, σελίδες 1009–1012. IEEE, 2022.

[101] Sajjan Kiran, Umesh Patil, P Siddarth Shankar και Poonam Ghuli. *Subtitle generation and video scene indexing using recurrent neural networks*. *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, σελίδες 847–854. IEEE, 2021.

[102] Anthony Sicilia, Konstantinos Pelechrinis και Kirk Goldsberry. *Deephoops: Evaluating micro-actions in basketball using deep feature representations of spatio-temporal data*. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, σελίδες 2096–2104, 2019.

[103] Joachim Gudmundsson και Michael Horton. *Spatio-temporal analysis of team sports*. *ACM Computing Surveys (CSUR)*, 50(2):1–34, 2017.

[104] Wei Wang, Liu Lv, Qiuchen Huang και Zun Liu. *Image processing-based Athletes' injury prevention mechanism*. *Internet Technology Letters*, σελίδα e474.

[105] Jeong Hun Kim, Jong Hyeok Choi, Young Ho Park και Aziz Nasridinov. *Abnormal situation detection on surveillance video using object detection and action recognition*. *Journal of Korea Multimedia Society*, 24(2):186–198, 2021.

[106] Wen Huang Cheng, Sijie Song, Chieh Yun Chen, Shintami Chusnul Hidayati και Jiaying Liu. *Fashion meets computer vision: A survey*. *ACM Computing Surveys (CSUR)*, 54(4):1–41, 2021.

[107] Ying Xue, Jianshan Sun, Yezheng Liu, Xin Li και Kun Yuan. *Facial expression-enhanced recommendation for virtual fitting rooms*. *Decision Support Systems*, 177:114082, 2024.

[108] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev και others. *The kinetics human action video dataset*. *arXiv preprint arXiv:1705.06950*, 2017.

[109] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier και Andrew Zisserman. *A short note about kinetics-600*. *arXiv preprint arXiv:1808.01340*, 2018.

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

βλπ	βλέπε
κ.λπ.	και λοιπά
κ.ο.κ	και ούτω καθεξής
π.χ.	παραδείγματος χάριν
BPF	Band Pass Filter
CNN	Convolutional Neural Network
SVM	Support Vector Machine
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory
HMM	Hidden Markov Model
AI	Artificial Intelligence
ChatGPT	Chat Generative Pretrained Transformer
MRI	Magnetic Resonance Imaging
ML	Machine Learning
FFNN	Feed Forward Neural Network
ANN	Artificial Neural Network
SNN	Simulated Neural Network
ReLU	Rectified Linear Unit
NLP	Natural Language Processing
MAE	Masked Auto-Encoder
IMU	Inertial Measurement Unit
MOCAPS	Motion Capture Systems
PCA	Principal Components Analysis
STFT	Short-Time Fourier Transform
IAPR	International Association for Pattern Recognition
Adam	Adaptive Moment Estimation
GELU	Gaussian Error Linear Unit
ViT	Vision Transformer
DEiT	Data Efficient Image Transformer
SGD	Stochastic Gradient Decent
TnT	Transformer in Transformer
HD	High Definition
Swin	shifted Windows
ToMe	Token Merging
FDViT	Flexible Downsampling Vision Transformer

ViVit	Video Vision Transformer
STTS	Spatial Temporal Token Selection
VideoMAE	Video Masked Autoencoder
VideoMAE V2	Video Masked Autoencoder Version2
MGMAE	Motion Guided Masked Autoencoder
RAFT	Recurrent all-pairs Field Transforms
VAR	Virtual Assistant Referee
VARS	Virtual Assistant Referee System
MLP	Multilayer Perceptron

Απόδοση ξενόγλωσσων όρων

Απόδοση	Ξενόγλωσσος όρος
τεχνητή νοημοσύνη	artificial intelligence
όραση υπολογιστών	computer vision
αναγνώριση ανθρώπινης πράξης	human action recognition
ικανότητα επίλυσης προβλημάτων	problem solving
επεξεργασία φυσικής γλώσσας	natural language processing
νευρωνικό δίκτυο	neural network
μηχανική μάθηση	machine learning
βαθιά μάθηση	deep learning
υλικό	hardware
βαθμός εμπιστοσύνης	degree of confidence
επιβλεπόμενη μάθηση	supervised learning
μη επιβλεπόμενη μάθηση	unsupervised learning
ενισχυτική μάθηση	reinforcement learning
επισήμανση	labeling
συσταδοποίηση	clustering
συνάρτηση επιβράβευσης	reward function
πράκτορας	agent
πρότυπο	pattern
σύνολο δεδομένων	dataset
προεπεξεργασία	preprocessing
συνάρτηση ενεργοποίησης	activation function
συντελεστής συστημικού σφάλματος	bias
βάρος	weight
διαβάθμιση κλίσης	gradient descent
ζυγισμένο άθροισμα	weighted sum
νευρωνικά δίκτυα εμπρόσθιας τροφοδότησης	feed forward neural networks
χαρακτηριστικό	feature
ταξινόμηση	classification
συνελικτικό νευρωνικό δίκτυο	convolutional neural network
συγκέντρωση	pooling
συγκέντρωση τοπικού μέσου	average pooling
συγκέντρωση τοπικού μεγίστου	max pooling
πλήρως συνδεδεμένο επίπεδο	fully connected layer
διαχωρισμός	segmentation

ορθότητα	accuracy
ακρίβεια	precision
μετασχηματιστής	transformer
προσοχή	attention
μηχανισμός προοσχής	attention mechanism
μονάδα επεξεργασίας	token
ενσωμάτωση	embedding
αυτο-προσοχή	self-attention
απο-ενσωμάτωση	un-embedding
κανονικοποίηση	normalization
βάρη αιτημάτων	query weights
βάρη κλειδιών	key weights
βάρη τιμών	value weights
διάνυσμα αιτήματος	query vector
διάνυσμα κλειδιού	key vector
διάνυσμα τιμής	value vector
προσοχή πολλαπλών κεφαλών	multi head attention
κεφαλή προσοχής	attention head
αυτο-επιβλεπόμενη μάθηση	self-supervised learning
προεκπαίδευση	pretraining
ρύθμιση με ακρίβεια	finetuning
αυτοκωδικοποιητής	autoencoder
αποκρύπτων αυτοκωδικοποιητής	masked autoencoder
αυτοκωδικοποιητής	autoencoder
επίπεδο συμφόρησης	bottleneck
πολλαπλότητα των δεδομένων	data manifold
εικονοστοιχείο	pixel
επαλήθευση	validation
αναγνώριση ανθρώπινης δράσης	human action recognition
τελευταίας τεχνολογίας	state of the art
μονάδα μέτρησης αδράνειας	inertial measurement unit
επιταχυσνιόμετρο	accelerometer
μαγνητόμετρο	magnetometer
μέσο φιλτράρισμα	mean filtering
διάμεσο φιλτράρισμα	median filtering
επαύξηση	augmentation
εξομοίωση ιστογράμματος	histogram equalization
ανάλυση κύριων συνιστωσών	principal components analysis
φασματογράφημα	spectrogram
βραχυπρόθεσμος μετασχηματισμός Fourier	short-time Fourier transform
τυχαίο δάσος	random forest
κρυφό μοντέλο Markov	hidden Markov model
μηχανή διανυσμάτων υποστήριξης	support vector machine

επαναλαμβανόμενο νευρωνικό δίκτυο	recurrent neural network
δίκτυο μακράς βραχύχρονης μνήμης	long short term memory network
δέντρο απόφασης	decision tree
συνάρτηση πυρήνα	kernel function
οπισθοδιάδοση	backpropagation
πρόβλημα εκμηδένισης της τιμής της κλίσης	vanishing gradient problem
πρόβλημα εκτόξευσης της τιμής της κλίσης	exploding gradient problem
βελτιστοποιητής	optimizer
τμήμα	patch
συνέλιξη κατά βάθος	depthwise convolution
βήμα	stride
μείωση μεγέθους	downsampling
εντοπισμός αντικειμένων	object detection
πολυτροπικό	multimodal
ισαλλοίωτα μεταφράσιμο	translation equivariant
φίλτρο διέλευσης χαμηλών συχνοτήτων	low pass filter
φίλτρο διέλευσης υψηλών συχνοτήτων	high pass filter
οπτικός μετασχηματιστής	vision transformer
δικυβική παρεμβολή	bicubic interpolation
μονάδα απόσταξης γνώσης	distillation token
συνάρτηση σφάλματος διασταυρωμένης εντροπίας	cross entropy loss function
μετασχηματιστής μετατοπιζόμενων παραθύρων	shifted window transformer
συγχώνευση μονάδων επεξεργασίας	token merging
χάρτης χαρακτηριστικών	feature map
διγραμμική παρεμβολή	bilinear interpolation
ωμή βία	brute force
σωληνοειδής τμηματοποίηση	tubelet embedding
ύστερη συγχώνευση	late fusion
παραγοντοποιημένη αυτο-προσοχή	factorized self-attention
αυτο-προσοχή	self-attention
ορθότητα πρώτης πρόβλεψης	top-1 accuracy
ορθότητα πέντε πρώτων προβλέψεων	top-5 accuracy
υπερπροσαρμογή	overfitting
απόκρυψη κινούμενης κυψελίδας	running cell masking
οπτική ροή	optical flow
οπίσθια αλλοίωση	backward warping
κινητικά καθοδηγούμενος αποκρύπτων αυτοκωδικοποιητής	motion guided masked autoencoder
επεξεργασία	post-production
εικονικός βοηθός διαιτητή	virtual assistant referee
σύστημα εικονικού βοηθού διαιτητή	virtual assistant referee system

ανάκληση	recall
σύνολο εκπαίδευσης	train set
σύνολο επαλήθευσης	validation set
σύνολο δοκιμών	test set
πολυεπίπεδος νευρώνας	multilayer perceptron
αποκοπή	dropout
παρτίδα	batch
εποχή προθέρμανσης εκπαίδευσης	warmup epoch
κλίση ενημέρωσης	gradient
έκτοπη τιμή	outlier
βηματοσυνάρτηση	step function
επεξηγησιμότητα	explainability
μικρογραφία βίντεο	thumbnail