Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

Επαύξηση δεδομένων: Εξετάζοντας την αποτελεσματικότητα της τεχνικής **mixup** σε προβλήματα συναισθηματικής αναγνώρισης του πραγματικού κόσμου

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ανδρέα Ψαρουδάκη

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής ΕΜΠ

Αθήνα, Φεβρουάριος 2022

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

Επαύξηση δεδομένων: Εξετάζοντας την αποτελεσματικότητα της τεχνικής **mixup** σε προβλήματα συναισθηματικής αναγνώρισης του πραγματικού κόσμου

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ανδρέα Ψαρουδάκη

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 2η Φεβρουαρίου 2022

.....	.....	.....
Στέφανος Κόλλιας Καθηγητής ΕΜΠ	Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής ΕΜΠ	Γιώργος Στάμου Καθηγητής ΕΜΠ

Αθήνα, Φεβρουάριος 2022

.....
Ανδρέας Ψαρουδάκης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών ΕΜΠ

© Ανδρέας Ψαρουδάκης, 2022. Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Το ανθρώπινο συναίσθημα αποτελεί μια συνειδητή υποκειμενική εμπειρία η οποία μπορεί να εκδηλωθεί με ποικίλους τρόπους. Την τελευταία δεκαετία, με την ραγδαία εξέλιξη στον τομέα της Τεχνητής Νοημοσύνης, έχουν πραγματοποιηθεί πολυάριθμες μελέτες για την ανάπτυξη συστημάτων και ρομπότ που θα είναι σε θέση να αντιλαμβάνονται αυτόματα ανθρώπινα συναισθήματα και συμπεριφορές. Απώτερος στόχος είναι η δημιουργία ψηφιακών βοηθών που θα εμφανίζουν ανθρωποκεντρικό χαρακτήρα και θα αλληλεπιδρούν με τους χρήστες με όσο το δυνατόν πιο φυσικό τρόπο. Πρόκειται για ένα πολύ σύνθετο και απαιτητικό εγχείρημα μιας και η συναισθηματική αναγνώριση σε συνθήκες πραγματικού κόσμου εμπεριέχει πολλούς αστάθμητους παράγοντες.

Στο πλαίσιο αυτό φαίνεται πως μπορούν να συνεισφέρουν ιδιαίτερα τα Νευρωνικά Δίκτυα Βαθιάς Μάθησης, σύγχρονα υπολογιστικά εργαλεία που έχουν την ικανότητα να διαχειρίζονται αποτελεσματικά μεγάλα σύνολα πληροφοριών. Παρά τη σημαντική ισχύ τους, τα δίκτυα αυτά είναι επιρρεπή στο φαινόμενο της υπερεκπαίδευσης. Αυτό σημαίνει πως συχνά καταλήγουν να απομνημονεύουν τα δεδομένα εισόδου, αδυνατώντας να γενικεύσουν επιτυχημένα το εξεταζόμενο πρόβλημα. Μια καλή λύση είναι η επέκταση του συνόλου εκπαίδευσης, με την προσθήκη νέων δειγμάτων. Ωστόσο, σε πολλές εφαρμογές η συγκέντρωση νέων εικόνων και η αντίστοιχη επισημείωσή τους αποτελούν μια αρκετά χρονοβόρα και δαπανηρή διαδικασία.

Για το λόγο αυτό έχουν προταθεί διάφορες τεχνικές επαύξησης δεδομένων, δηλαδή μέθοδοι οι οποίες παράγουν τεχνητά νέα δείγματα, αξιοποιώντας αυτά που υπάρχουν ήδη διαθέσιμα. Μια αρκετά πρόσφατη τέτοια τεχνική, η οποία έχει συνεισφέρει θετικά σε διάφορα προβλήματα κατηγοριοποίησης, είναι η mixup. Σύμφωνα με αυτή, η εκπαίδευση ενός δικτύου πραγματοποιείται πάνω σε κυρτούς συνδυασμούς των δεδομένων εκπαίδευσης και των αντίστοιχων ετικετών τους. Με τον τρόπο αυτό επεκτείνεται η κατανομή των διαθέσιμων δεδομένων και το δίκτυο παρουσιάζει καλύτερη ικανότητα γενίκευσης.

Στα πλαίσια της παρούσας διπλωματικής, εξετάζουμε την αποτελεσματικότητα της τεχνικής mixup στο πρόβλημα της συναισθηματικής αναγνώρισης σε πραγματικές, μη-ελεγχόμενες συνθήκες (in-the-wild). Συγκεκριμένα, εκπαιδεύουμε δίκτυα Βαθιάς Μάθησης με την εν λόγω τεχνική για κατηγοριοποίηση εικόνων εκφράσεων προσώπου στα 7 βασικά συναισθήματα. Παράλληλα, προτείνουμε και μια παραλλαγή της mixup, την Addmixup, με βάση την οποία το δίκτυο εκπαιδεύεται ταυτόχρονα πάνω σε εικονικά και πραγματικά παραδείγματα. Συγκρίνουμε τις δύο αυτές μεθόδους με την κλασική αρχή Ελαχιστοποίησης Εμπειρικού Ρίσκου ενώ παράλληλα εξετάζουμε και την επίδραση του dropout, μιας μορφής κανονικοποίησης του δικτύου, σε όλες τις προαναφερθείσες τεχνικές. Από την πειραματική μας μελέτη εξάγονται χρήσιμα συμπεράσματα ενώ παράλληλα τίθενται οι βάσεις για πολλές ακόμα μελλοντικές επεκτάσεις.

— ***Λέξεις-Κλειδιά:*** Αναγνώριση συναισθήματος in-the-wild, Κατηγοριοποίηση στα βασικά συναισθήματα, Νευρωνικά Δίκτυα Βαθιάς Μάθησης, Επαύξηση δεδομένων, mixup

Abstract

The human emotion constitutes a conscious subjective experience that can be expressed in various ways. During the past decade, with the rapid development in the field of Artificial Intelligence, scientists have conducted numerous studies to develop systems and robots that will be capable of perceiving automatically people's feelings and behaviors. The ultimate goal is the creation of digital assistants that will display a human-centered character and interact with users in the most natural way possible. It is a very complex and demanding task since the emotional recognition in real world conditions contains many imponderables.

It seems that some modern computer tools, named Deep Neural Networks, can contribute significantly towards this direction, since they are capable of managing effectively large datasets. Despite their considerable power, these networks are prone to overfitting. This means that they often tend to memorize the input data, thus failing to generalize successfully the problem under consideration. One good solution would be to expand the training set, by adding new samples. However, in multiple applications, the collection of new images and their corresponding annotation is quite a time consuming and costly process.

Because of this, various data augmentation techniques have been proposed; methods that produce artificially new samples, by utilising those already available. A fairly recent technique of this kind, which has positively contributed to various classification tasks, is called mixup. According to it, a Neural Network is trained on convex combinations of pairs of examples and their corresponding labels. By doing so, the distribution of the available data is extended and the generalisation ability of the network improves.

In this diploma thesis, we examine the effectiveness of mixup in affective computive tasks in real, uncontrolled, conditions (in-the-wild). Specifically, we train Deep Neural Networks using this technique to classify facial images in 7 basic expressions. Meanwhile, we propose a variation of mixup, named Addmixup, according to which the network is trained concurrently on virtual and real examples. We compare these two methods with the classic Empirical Risk Minimization principle, while at the same time, we examine the effect of dropout, a form of network normalization, in all the aforementioned techniques. From our experimental study, useful conclusions are drawn. In addition, the foundations are laid for many further future extensions.

— ***Keywords:*** Human affect recognition in-the-wild, Basic expression classification, Deep Neural Networks, Data augmentation, mixup

Ευχαριστίες

Η παρούσα διπλωματική εργασία αποτελεί το τελευταίο κεφάλαιο της φοίτησής μου στο προπτυχιακό πρόγραμμα σπουδών της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου. Με την ολοκλήρωσή της κλείνει ένας πολυετής κύκλος σπουδών, ο οποίος μου προσέφερε σημαντικές γνώσεις και εμπειρίες. Με αυτή την αφορμή, αισθάνομαι την ανάγκη να ευχαριστήσω όλους όσους συνέβαλαν στην σταδιοδρομία μου όλα αυτά τα χρόνια.

Πρωτίστως, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή της διπλωματικής μου, κύριο Στέφανο Κόλλια, ο οποίος μου παρείχε τη δυνατότητα να ασχοληθώ με το αντικείμενο που με ενδιέφερε, προσφέροντάς μου παράλληλα διαρκή υποστήριξη. Χωρίς τη βοήθεια και τις πολύτιμες συμβουλές του, η συγγραφή της παρούσας διπλωματικής δεν θα ήταν εφικτή.

Στη συνέχεια, θα ήθελα να ευχαριστήσω ιδιαίτερα τον κύριο Δημήτριο Κόλλια, αναπληρωτή καθηγητή του Queen Mary University of London, ο οποίος με καθοδήγησε αναλυτικά κατά τη διάρκεια εκπόνησης της εργασίας, έτσι ώστε να επιτύχω μια πολύπλευρη προσέγγιση. Υπήρξε πάντα πρόθυμος να απαντήσει άμεσα σε οποιαδήποτε απορία μου, βοηθώντας με έτσι να διευρύνω του ορίζοντες μου αλλά και να αγαπήσω το αντικείμενο της έρευνάς μου.

Τέλος, θα ήθελα να ευχαριστήσω ιδιαίτερα την οικογένειά μου και τους φίλους μου, οι οποίοι στάθηκαν δίπλα μου όλα αυτά τα χρόνια, προσφέροντας μου στήριξη με κάθε τρόπο.

Ανδρέας Ψαρουδάκης
Φεβρουάριος 2022

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Το ανθρώπινο συναίσθημα

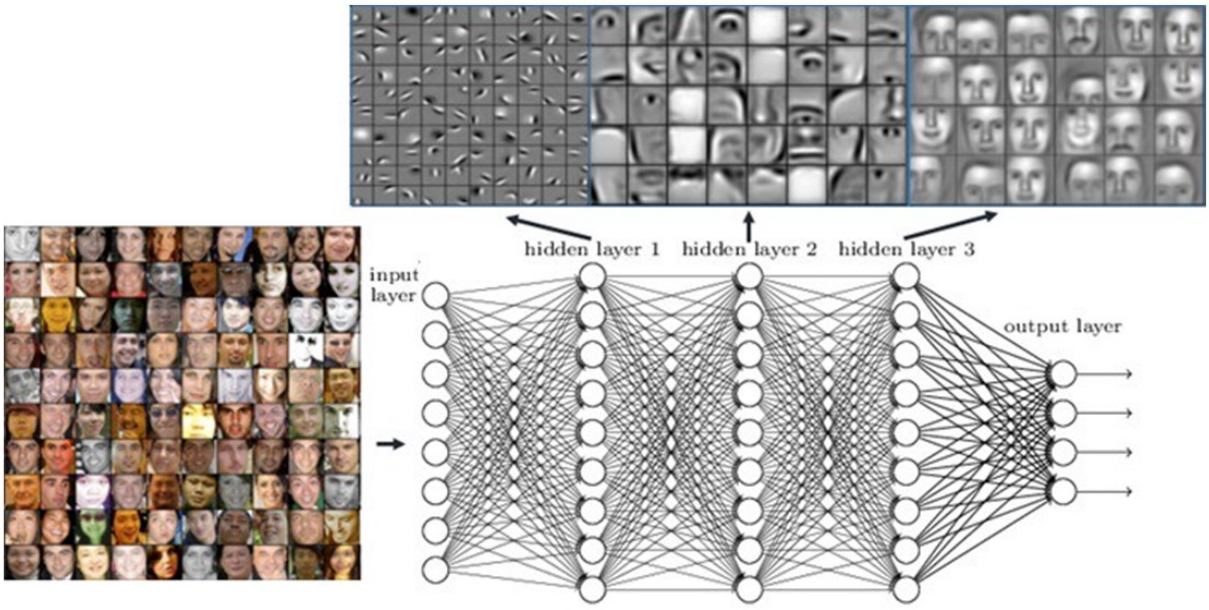
Από τα πρώτα κιόλας στάδια της ζωής του, ο άνθρωπος είναι σε θέση να βιώνει μια πληθώρα συναισθημάτων, τα οποία παίζουν καθοριστικό ρόλο στην επικοινωνία του με τον εξωτερικό κόσμο. Κάθε συναίσθημα ορίζεται ως μια σύνθετη υποκειμενική συνειδητή εμπειρία, πρόκειται δηλαδή για έναν συνδυασμό νοητικών και ψυχοσωματικών καταστάσεων. Η ανάλυση και ερμηνεία των συναισθημάτων αποτελεί αντικείμενο μελέτης των κλάδων της Ψυχολογίας και της Φυσιολογίας.

Με την πάροδο του χρόνου έχουν πραγματοποιηθεί πολυάριθμες μελέτες για την αποτελεσματική ανίχνευση και μοντελοποίησή τους. Μάλιστα την τελευταία δεκαετία, με την εξέλιξη της Τεχνητής Νοημοσύνης, έχουν γίνει πολλαπλές προσπάθειες, από την πλευρά των επιστημόνων, για την ανάπτυξη συστημάτων και ρομπότ που θα είναι σε θέση να αντιλαμβάνονται ανθρώπινα συναισθήματα και συμπεριφορές. Απώτερος στόχος είναι η δημιουργία ψηφιακών βοηθών που θα εμφανίζουν ανθρωποκεντρικό χαρακτήρα και θα αλληλεπιδρούν με τους χρήστες με όσο το δυνατόν πιο φυσικό τρόπο.

Στο παρελθόν κάτι τέτοιο δεν ήταν εφικτό, λόγω του περιορισμένου αριθμού δεδομένων αλλά και της χαμηλής υπολογιστικής ισχύος. Ωστόσο, η ραγδαία ανάπτυξη του διαδικτύου και των μέσων κοινωνικής δικτύωσης έχουν οδηγήσει στη δημιουργία ενός τεράστιου όγκου δεδομένων. Παράλληλα, έχουν ανακαλυφθεί ισχυρά υπολογιστικά εργαλεία, όπως τα Νευρωνικά Δίκτυα Βαθιάς Μάθησης, τα οποία έχουν συντελέσει στην αποτελεσματική αντιμετώπιση πολυάριθμων σύνθετων προβλημάτων.

1.2 Βαθιά Μάθηση για αναγνώριση συναισθημάτων

Τα Νευρωνικά Δίκτυα Βαθιάς Μάθησης (DNNs) μπορούν να συμβάλουν στην αντιμετώπιση σημαντικών προβλημάτων ανάλυσης και κατηγοριοποίησης στον τομέα της Όρασης Υπολογιστών, καθώς είναι σε θέση να διαχειρίζονται αποτελεσματικά μεγάλα σύνολα πληροφοριών. Πρόκειται για πολυεπίπεδες δομές που αξιοποιούν κατάλληλα συνελικτικά φίλτρα για την αυτόματη ανίχνευση χωρικών και χρονικών εξαρτήσεων των δεδομένων εισόδου. Τα πρώτα κρυφά επίπεδα των δικτύων αυτών κωδικοποιούν γενικά χαρακτηριστικά, ωστόσο καθώς προχωράμε σε μεταγενέστερα επίπεδα, εξάγονται αναπαραστάσεις υψηλού επιπέδου, οι οποίες σχετίζονται με το πρόβλημα που εξετάζεται. Ένα χαρακτηριστικό παράδειγμα αναγνώρισης προσώπου με DNN φαίνεται στο Σχήμα 1.1. Τα πρώτα κρυφά στρώματα κωδικοποιούν χαρακτηριστικά χωρίς κάποιο φυσικό νόημα, σε αντίθεση με τα βαθύτερα επίπεδα τα οποία διαμορφώνουν αναπαραστάσεις που προσεγγίζουν πολύ ανθρώπινα πρόσωπα.



Σχήμα 1.1: Εκπαίδευση Αρχιτεκτονικής Δικτύου Βαθιάς Μάθησης με εικόνες προσώπων

Παρά τις πολλαπλές δυνατότητες που παρέχουν τα Δίκτυα Βαθιάς Μάθησης, το πρόβλημα της ανίχνευσης συναισθηματικών καταστάσεων είναι ιδιαίτερα απαιτητικό, καθώς όπως αναφέρθηκε, το συναίσθημα συνιστά μια υποκειμενική και αφηρημένη έννοια που μπορεί να εκδηλωθεί με ποικίλους τρόπους. Αναλυτικότερα, ένα συναίσθημα αποτυπώνεται κατά βάση στην έκφραση του προσώπου, ωστόσο μερικές φορές μπορεί να γίνει καλύτερα αντιληπτό μέσω της γλώσσας του σώματος (π.χ. από χειρονομίες) ή μέσω της φυσικής ομιλίας (π.χ. έντονη λεκτική αντιπαράθεση).

Αν λοιπόν, κατά την προσέγγιση του προβλήματος, θέλουμε να λάβουμε υπόψη μας παραπάνω από μια ενδείξεις (π.χ. εικόνα της έκφρασης του προσώπου και ομιλία) τότε αυξάνεται σημαντικά η υπολογιστική πολυπλοκότητα του προβλήματος. Από την άλλη, αξιοποιώντας μόνο κάποια συγκεκριμένη πληροφορία (π.χ. το οπτικό σήμα) έχουμε μικρότερες πιθανότητες επιτυχημένης αναγνώρισης.

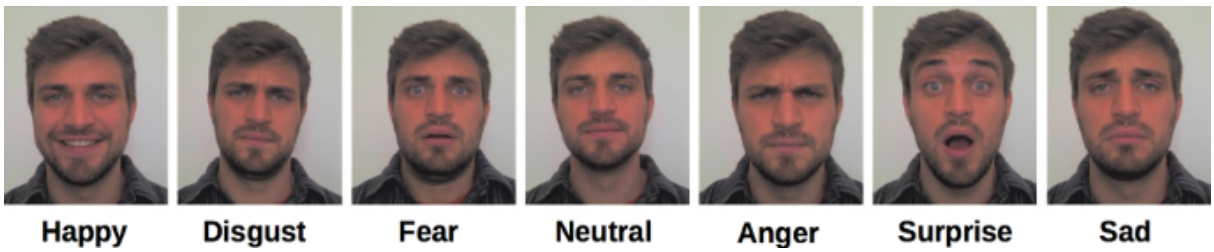
Εκτός αυτών, δεν πρέπει να ξεχνάμε ότι συνήθως δεν μελετάμε ένα στατικό πρόβλημα (π.χ. μια ανεξάρτητη εικόνα) αλλά ένα χρονικά μεταβαλλόμενο (π.χ. χρονική εξέλιξη συναισθημάτων κατά τη διάρκεια ενός βίντεο), το οποίο μάλιστα εξαρτάται από το χρήστη αλλά και τις συνθήκες του περιβάλλοντος. Με άλλα λόγια, άτομα διαφορετικού φύλου,ηλικίας ή εθνικότητας είναι πιθανό να εκφράζονται με διαφορετικό τρόπο ενώ, παράλληλα, το περιβάλλον στο οποίο βρίσκονται επηρεάζει άμεσα ή έμμεσα τη συναισθηματική τους κατάσταση. Είναι λοιπόν αντιληπτό πως πρόκειται για ένα αρκετά σύνθετο και δυσεπίλυτο πρόβλημα που θα απασχολεί για αρκετά ακόμα χρόνια την επιστημονική κοινότητα.

1.3 Αναπαράσταση συναισθημάτων

Η αναπαράσταση και μοντελοποίηση των ανθρωπίνων συναισθημάτων αποτελεί ένα πεδίο που συγκεντρώνει ιδιαίτερο ερευνητικό ενδιαφέρον. Με την πάροδο του χρόνου έχουν προταθεί διάφορα μοντέλα συναισθηματικής απεικόνισης. Στη ενότητα αυτή αναλύουμε τρία εξ αυτών, τα οποία έχουν επικρατήσει και χρησιμοποιούνται ευρέως μέχρι και σήμερα στο πρόβλημα της αναγνώρισης συναισθημάτων.

1.3.1 Κατηγορικό μοντέλο

Η πιο ευρέως διαδεδομένη τεχνική αναπαράστασης των ανθρωπίνων συναισθημάτων είναι αυτή της ταξινόμησης σε βασικές διακριτές κατηγορίες. Πιο συγκεκριμένα, το 1970, ο ψυχολόγος Paul Ekman πραγματοποίησε μια διαπολιτισμική έρευνα [24] για να εξετάσει αν υπάρχουν συναισθηματικές καταστάσεις που μπορούν να θεωρηθούν καθολικές. Πράγματι, κατέληξε στο συμπέρασμα πως οι άνθρωποι, ανεξαρτήτως φύλου, ηλικίας, κουλτούρας και μορφωτικού επιπέδου, αντιλαμβάνονται και εκφράζουν σχεδόν με τον ίδιο τρόπο 6 βασικά συναισθήματα: τη χαρά, την απέχθεια, το φόβο, το θυμό, την έκπληξη και τη λύπη.



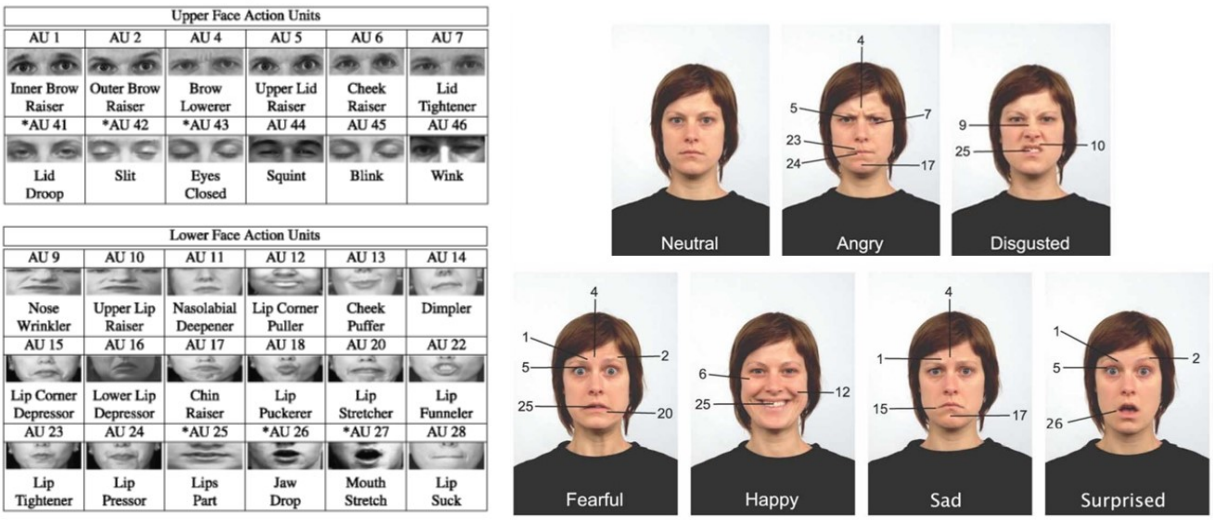
Σχήμα 1.2: Τα 6 βασικά συναισθήματα που όρισε ο Ekman και η ουδέτερη κατάσταση [74]

Παρόλα αυτά, σύγχρονες προηγμένες μελέτες στις Νευροεπιστήμες και την Ψυχολογία έρχονται σε σύγκρουση με τη συγκεκριμένη διατύπωση καθώς υποστηρίζουν πως το μοντέλο των 6 συναισθημάτων δεν είναι καθολικό αλλά εξαρτάται άμεσα από τον πολιτισμό κάθε λαού. Επίσης, μία ακόμα αδυναμία του κατηγορικού μοντέλου είναι η απλότητά του, καθώς αδυνατεί να αναπαραστήσει σύνθετες συναισθηματικές καταστάσεις της καθημερινής ζωής.

Ωστόσο, παρά τους περιορισμούς που εισάγει, εξακολουθεί να αποτελεί την πιο δημοφιλή προσέγγιση στον τομέα της αναγνώρισης συναισθήματος από εικόνες εκφράσεων προσώπου, γεγονός που οφείλεται στο ότι είναι καλά ορισμένη και γίνεται εύκολα αντιληπτή. Δεν είναι άλλωστε τυχαίο που οι περισσότερες επισημειωμένες βάσεις δεδομένων που χρησιμοποιούνται στην αναγνώριση συναισθήματος περιέχουν ετικέτες για τα 7 βασικά συναισθήματα (τα 6 που όρισε ο Ekman και την ουδέτερη κατάσταση). Για το λόγο αυτό, στα πλαίσια της παρούσας διπλωματικής, θα αξιοποιήσουμε το συγκεκριμένο πρότυπο συναισθηματικής αναπαράστασης.

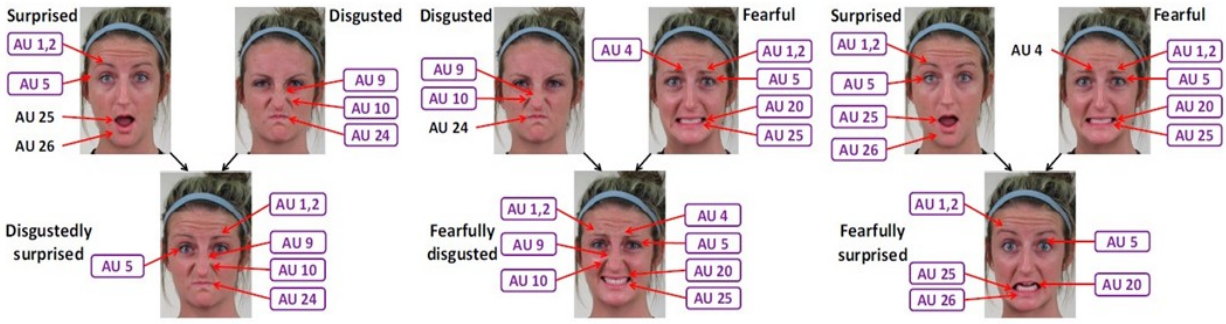
1.3.2 Ανίχνευση Μονάδων Δράσης στο πρόσωπο

Η διακριτή αναπαράσταση των συναισθημάτων μπορεί να επιτευχθεί και με τη χρήση του Συστήματος Κωδικοποίησης Δράσεων Προσώπου (Facial Action Coding System - FACS), το οποίο προτάθηκε από τον Paul Ekman και τους συναδέλφους του το 1978, ενώ εμπλουτίστηκε περαιτέρω το 2002 [23]. Με βάση αυτή την προσέγγιση, κάθε ανθρώπινη έκφραση αποτελεί μια παραμόρφωση του ουδέτερου (ανέκφραστου) προσώπου, η οποία οφείλεται στην ενεργοποίηση ενός ή περισσοτέρων μυών. Οι μυς αυτοί που δρουν και προκαλούν κάθε στοιχειώδη παραμόρφωση διαμορφώνουν τις λεγόμενες μονάδες δράσης (Action Units). Τα AUs δεν εμφανίζονται μεμονωμένα, αλλά ως στοιχειώδεις μονάδες εκφράσεων προσώπου. Έτσι, ορισμένα από αυτά συχνά εκδηλώνονται ταυτόχρονα, ενώ άλλα είναι αμοιβαίως αποκλειόμενα, δηλαδή δεν ενεργοποιούνται ποτέ την ίδια χρονική στιγμή.



Σχήμα 1.3: Αριστερά απεικονίζονται τα βασικά Action Units καθώς και οι κινήσεις μυών του προσώπου στις οποίες αντιστοιχούν. Δεξιά φαίνονται τα Action Units που ενεργοποιούνται στα 7 βασικά συναισθήματα [67].

Δεδομένου ότι οποιαδήποτε έκφραση προσώπου, όσο σύνθετη και αν είναι, μπορεί να αναπαρασταθεί από έναν συνδυασμό επιλεγμένων μονάδων δράσης, είναι φανερό ότι το σύστημα FAC μπορεί να χρησιμοποιηθεί ως ένα κοινό πρότυπο για την ανάλυση και κατηγοριοποίηση ακόμα και ιδιαίτερα περίπλοκων συναισθημάτων [21], αφού ο συνολικός αριθμός συνδυασμών διαφορετικών Action Units που μπορεί να προκύψει είναι εκθετικά υψηλός. Πρόκειται, επομένως, για ένα ιδιαίτερα σημαντικό εύρημα στον τομέα της Όρασης Υπολογιστών, που ανοίγει το δρόμο για την μελέτη ενός ευρύτερου φάσματος εφαρμογών.



Σχήμα 1.4: Στις εικόνες προσώπου της κάτω σειράς απεικονίζονται τα Action Units που ενεργοποιούνται σε τρία σύνθετα συναισθήματα. Αυτά φαίνεται πως προκύπτουν με κατάλληλο συνδυασμό των AUs βασικών συναισθημάτων [21].

Παρόλο που τα Action Units φαίνεται να αναπαριστούν αποτελεσματικά τόσο απλές όσο και σύνθετες συναισθηματικές καταστάσεις, η ενεργοποίησή τους είναι συνήθως ανεπαίσθητη και διαρκεί πολύ λίγο χρόνο. Για το λόγο αυτό, η ανίχνευσή τους κατά τη διάρκεια μιας χρονικά μεταβαλλόμενης ακολουθίας εικόνων (π.χ. βίντεο) είναι ιδιαίτερα απαιτητική και προϋποθέτει τη διεξοδική παρατήρηση πιθανών μικρών μεταβολών του προσώπου σε κάθε νέο καρέ. Έτσι, μια τέτοια διαδικασία καταγραφής μονάδων δράσης, είναι αρκετά χρονοβόρα αλλά και κοστοβόρα εφόσον απαιτεί την πρόσληψη εξειδικευμένων επιστημόνων. Αυτός είναι και ο βασικός λόγος που η αναπαράσταση συναισθηματικών καταστάσεων με Action Units δεν είναι τόσο ευρέως διαδεδομένη. Εδώ αξίζει να αναφερθεί πως για την αντιμετώπιση

του προβλήματος έχει πραγματοποιηθεί μεγάλη προσπάθεια για τη δημιουργία εργαλείων αυτόματης ανίχνευσης AUs [6, 28].

Σε μια ψυχολογική μελέτη [21] που συμμετείχαν συνολικά 230 επιλεγμένοι υποψήφιοι, καταγράφηκαν τα Action Units που ενεργοποιούνται σε κάθε ένα από τα 6 βασικά συναισθήματα. Τα αποτελέσματα της έρευνας έδειξαν πως τα ενεργά AUs για ένα δεδομένο συναίσθημα δεν είναι κοινά σε όλους του ανθρώπους.

	Ψυχολογική Μελέτη [21]		Εμπειρικές Αποδείξεις, Aff-Wild2 [50]
Συναίσθημα	Πρωτότυπα AUs	Παρατηρούμενα AUs (με βάρη w)	AUs (με βάρη w)
Χαρά	12, 25	6 (0.51)	12 (0.82), 25 (0.7), 6 (0.57), 7 (0.83), 10 (0.63)
Λύπη	4, 15	1 (0.6), 6 (0.5), 11 (0.26), 17(0.67)	4 (0.53), 15 (0.42), 1 (0.31), 7 (0.13), 17 (0.1)
Φόβος	1, 4 , 20, 25	2 (0.57), 5 (0.63), 26 (0.33)	1 (0.52), 4 (0.4), 25 (0.82), 5 (0.38), 7 (0.57), 10 (0.57)
Θυμός	4, 7, 24	10 (0.26), 17 (0.52), 23 (0.29)	4 (0.65), 7 (0.45), 25 (0.4), 10 (0.33), 9 (0.15)
Έκπληξη	1, 2, 25, 26	5 (0.66)	1 (0.38), 2 (0.37), 25 (0.85), 26 (0.3), 5 (0.5), 7 (0.2)
Απέχθεια	9, 10, 17	4 (0.31), 24 (0.26)	9 (0.21), 10 (0.85), 17 (0.23), 4 (0.6), 7 (0.75), 25 (0.8)

Πίνακας 1.1: Σύσχέτιση Action Units με τα 6 βασικά συναισθήματα

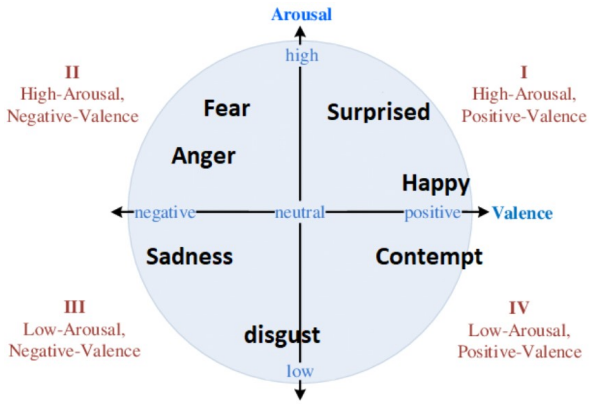
Συγκεκριμένα, όπως φαίνεται στον Πίνακα 1.1, σε κάθε μία από τις 6 συναισθηματικές κατηγορίες, υπάρχουν ορισμένα AUs που εμφανίστηκαν σε όλους τους υποψηφίους (πρωτότυπα) και κάποια τα οποία ενεργοποιήθηκαν μόνο σε έναν συγκεκριμένο αριθμό αυτών (παρατηρούμενα). Για τα παρατηρούμενα AUs καταγράφονται εντός παρενθέσεων τα ποσοστά των ανθρώπων που είχαν ενεργό το αντίστοιχο AU για το εκάστοτε συναίσθημα. Για παράδειγμα, μπορούμε να δούμε πως στο αίσθημα της χαράς το AU6 ήταν ενεργό μόνο στο 51% των υποψηφίων. Στην δεξιά στήλη του ίδιου πίνακα μπορούμε να παρατηρήσουμε τα αποτελέσματα μιας αυτόματης ανίχνευσης AUs που πραγματοποιήθηκε πάνω στη βάση Aff-Wild2 [49, 55, 56, 58], ένα database που περιέχει 558 βίντεο από το διαδίκτυο (συνολικά 2.786.201 καρέ). Σε αντιστοιχία με την προαναφερθείσα έρευνα, παρατηρούμε πως και εδώ, τα Action Units που ενεργοποιούνται δεν συμπίπτουν για όλα τα καρέ ενός δεδομένου συναισθήματος.

Τα συμπεράσματα αυτά καταδεικνύουν πως τα Action Units, παρόλο που είναι ιδιαίτερα χρήσιμα στην μοντελοποίηση και ανάλυση ανθρώπινων συναισθημάτων, δεν εμφανίζουν κάποια αμφιμονοσήμαντη αντιστοιχία με τις 6 διακριτές συναισθηματικές κατηγορίες. Το γεγονός αυτό σε συνδυασμό με τη μεγάλη δυσκολία δημιουργίας επισημειώσεων καθιστούν το πεδίο έρευνας σε αυτή την κατεύθυνση σχετικά περιορισμένο.

Αξίζει να σημειωθεί, πως η μοναδική βάση δεδομένων από εικόνες προσώπου κάτω από μη-ελεγχόμενες συνθήκες (in-the-wild), η οποία περιέχει labels από Action Units είναι η Emotionet [8]. Ωστόσο, ακόμα και σε αυτή, η συντριπτική πλειοψηφία των AUs (950.000) έχει ανιχνευθεί αυτόματα με χρήση κατάλληλου αλγορίθμου ενώ μόνο ένα μικρό ποσοστό (50.000) έχει επισημειωθεί χειροκίνητα από έμπειρους υπομνηματιστές.

1.3.3 Διανυσματικό μοντέλο Valence-Arousal

Μια άλλη αρκετά διαδεδομένη προσέγγιση αναπαράστασης συναισθημάτων είναι η διανυσματική, με βάση την οποία κάθε συναισθηματική κατάσταση μπορεί να περιγραφεί από έναν αριθμό συνεχών διαστάσεων [94, 111]. Στο πλαίσιο αυτό, το πρότυπο Valence-Arousal, που προτάθηκε το 1980 από τον James Russel [95], φαίνεται να είναι το πιο διαδεδομένο μέχρι και σήμερα. Σύμφωνα με αυτό, κάθε συναίσθημα μπορεί να μοντελοποιηθεί ως ένα σημείο στο δισδιάστατο χώρο, με συντεταγμένες τις συνεχείς συνιστώσες Valence και Arousal.

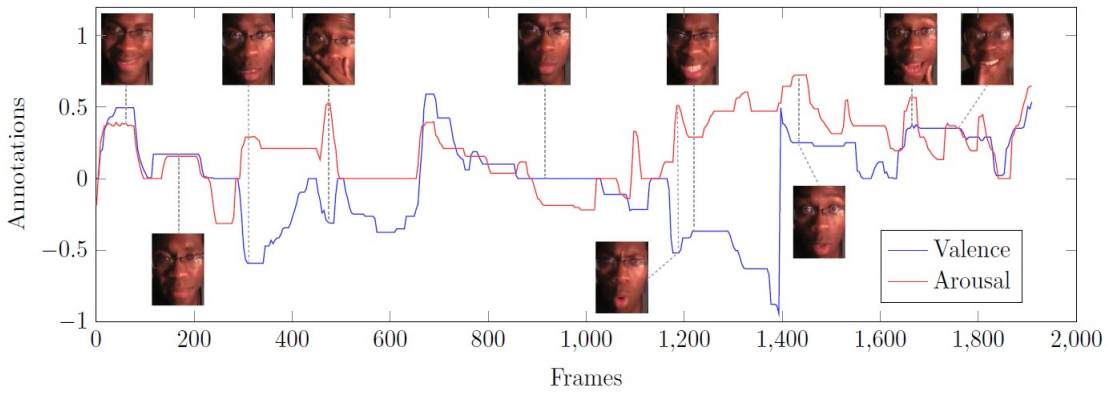


Σχήμα 1.5: Μοντέλο Valence-Arousal [9]

Η πρώτη απεικονίζεται στον οριζόντιο άξονα και εκφράζει το πόσο θετικό (ευχάριστο) ή αρνητικό (δυσάρεστο) είναι το συναίσθημα. Η δεύτερη απεικονίζεται στον κατακόρυφο άξονα και καθορίζει τη ένταση του συναισθήματος, δηλαδή πόσο ενεργό ή παθητικό είναι. Και οι δύο συνιστώσες αναπαριστώνται συνήθως στο εύρος [-1,1]. Στο Σχήμα 1.5 παρατηρούμε πως για δεδομένες τιμές των συνιστωσών προκύπτουν οι 6 συναισθηματικές καταστάσεις που όρισε ο Ekman καθώς και το αίσθημα της περιφρόνησης ("contempt") που εισήχθη αργότερα στα βασικά συναισθήματα [75].

Αξίζει να αναφέρουμε πως οι συνιστώσες Valence και Arousal σχετίζονται άμεσα με συγκεκριμένες λειτουργίες περιοχών του ανθρώπινου εγκεφάλου. Αναλυτικότερα, ο βρεγματικός λοβός του δεξιού ημισφαιρίου φαίνεται να παίζει έναν ειδικό ρόλο στη διαμόρφωση της συνιστώσας Arousal, ενώ οι μετωπιαίοι λοβοί στον καθορισμό της τιμής της Valence. Ωστόσο, παρά τις διαφορές που εμφανίζουν, υπάρχουν ισχυρές επιστημονικές πειραματικές ενδείξεις που καταδεικνύουν ότι οι δύο αυτές συνιστώσες είναι αλληλένδετες, δηλαδή παρουσιάζουν σημαντική αλληλεξάρτηση [2, 39, 69, 86].

Το πρότυπο Valence-Arousal έχει αποδειχθεί ιδιαίτερα αποτελεσματικό καθώς μπορεί να χρησιμοποιηθεί αρκετά επιτυχημένα στην μοντελοποίηση χρονικά μεταβαλλόμενων συναισθηματικών καταστάσεων (π.χ. για την αναπαράσταση συναισθημάτων κατά τη διάρκεια ενός βίντεο), όπως φαίνεται στο Σχήμα 1.6. Παρατηρούμε πως οι δύο συνιστώσες σκιαγραφούν αποτελεσματικά τις απότομες εναλλαγές συναισθημάτων και κατά συνέπεια ενσωματώνουν τη δυναμική του προβλήματος.



Σχήμα 1.6: Οι τιμές των συνιστωσών Valence και Arousal κατά τη διάρκεια ενός βίντεο της βάσης Aff-Wild [53], μαζί με τα αντίστοιχα καρέ.

Ωστόσο, παρά την σημαντική ισχύ του συγκεκριμένου μοντέλου, δεν πρέπει να παραλείψουμε να πούμε ότι η συγκέντρωση ετικετών για τις συνεχείς συνιστώσες Valence και Arousal αποτελεί μια πολύ χρονοβόρα διαδικασία που πραγματοποιείται από ειδικούς. Εκτός αυτού, υπάρχει επίσης και μεγάλη πιθανότητα να εμφανιστούν αισθητές αποκλίσεις μεταξύ των επισημειώσεων διαφορετικών υπομνηματιστών, καθώς οι τιμές που ανατίθενται στις δύο συνιστώσες για κάθε εικόνα, βασίζονται κατά ένα βαθμό στην υποκειμενική τους κρίση. Για

το λόγο αυτό, μετά το πέρας της διαδικασίας συγκέντρωσης των ετικετών, απαιτείται περαιτέρω στατιστική ανάλυση, έτσι ώστε να παραλειφθούν τυχόν έκτοπες μετρήσεις (outliers). Λόγω αυτών των προβλημάτων, το μοντέλο της διανυσματικής αναπαράστασης συγκεντρώνει μικρότερο ερευνητικό ενδιαφέρον από αυτό της κατηγορικής προσέγγισης.

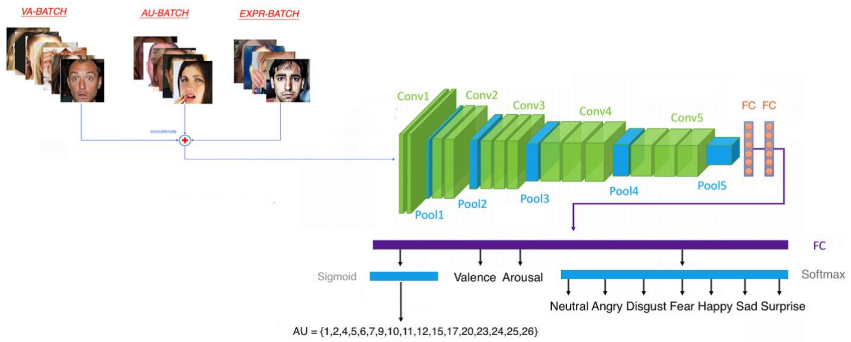
1.3.4 Διασύνδεση συναισθηματικών προσεγγίσεων

Αναλύσαμε λοιπόν τις 3 βασικές προσεγγίσεις που χρησιμοποιούνται για την αναπαράσταση συναισθηματικών καταστάσεων:

1. το κατηγορικό μοντέλο με την ταξινόμηση στις βασικές κατηγορίες συναισθημάτων
2. την ανίχνευση Μονάδων Δράσης (Action Units) στο πρόσωπο και
3. τη διανυσματική προσέγγιση με τις συνεχείς συνιστώσες Valence-Arousal

Μέχρι και σήμερα τα τρία αυτά tasks μελετώνται κυρίως ανεξάρτητα, με την εκπαίδευση δικτύων πάνω σε datasets που έχουν επισημειωθεί με έναν από τους παραπάνω τρόπους (single-task learning). Ωστόσο, όπως είδαμε, οι εναλλακτικές αυτές μέθοδοι είναι άρρηκτα διασυνδεδεμένες και όλες χρησιμοποιούνται για την προσέγγιση του ίδιου προβλήματος, την αναπαράσταση των ανθρώπινων συναισθημάτων. Για παράδειγμα, αναφέραμε πως το Σύστημα Κωδικοποίησης Δράσεων Προσώπου (Facial Action Coding System) ορίστηκε για την καταγραφή των πρωτότυπων σημείων δράσης που ενεργοποιούνται σε κάθε ένα από τα βασικά συναισθήματα της κατηγορικής προσέγγισης [23, 25], ενώ μπορεί επίσης να χρησιμοποιηθεί για την μοντελοποίηση σύνθετων συναισθηματικών εκφράσεων [21]. Στο [44], μια ομάδα ερευνητών απέδειξε πως Νευρωνικά Δίκτυα που εκπαιδεύονται πάνω στην αναγνώριση συναισθημάτων, μαθαίνουν εμμέσως να αναγνωρίζουν και Action Units. Επίσης, με βάση το [78], οι συνεχείς συνιστώσες Valence και Arousal μπορούν να ερμηνευτούν με χρήση των AUs. Για παράδειγμα, το AU12 (τράβηγμα της γωνίας των χειλιών) σχετίζεται με θετική τιμή Valence. Στο [37], Ο Stephan Hamann προσπάθησε να αντιστοιχήσει τα διακριτά συναισθήματα με τις συνεχείς συνιστώσες Valence-Arousal, αξιοποιώντας τεχνολογία νευροαπεικόνισης, για τον εντοπισμό των περιοχών του εγκεφάλου που σχετίζονται με τα δύο είδη αναπαραστάσεων. Στο ίδιο πλαίσιο, οι Sven Buechel και Udo Hahn [11] πρότειναν ένα νευρωνικό δίκτυο που μετατρέπει τις συνιστώσες Valence-Arousal στα βασικά συναισθήματα για την κατασκευή ενός λεξικού συναισθημάτων.

Είναι λοιπόν εμφανές πως, στον τομέα της αναγνώρισης και ανάλυσης ανθρωπίνων συναισθημάτων, θα μπορούσε να χρησιμοποιηθεί και μια συνδυαστική προσέγγιση, κατά την οποία θα αξιοποιούνται παράλληλα δύο ή περισσότερες συναισθηματικές αναπαραστάσεις. Η τεχνική του multi-task learning μελετήθηκε για πρώτη φορά στο [13], όπου προτάθηκε η ταυτόχρονη εκμάθηση πολλαπλών διασυνδεδεμένων εργασιών για την επίτευξη καλύτερων αποτελεσμάτων. Με βάση την μέθοδο αυτή, μέρος της γνώσης που αποκομίζεται από την επίλυση ενός από τα συσχετιζόμενα tasks μπορεί να χρησιμοποιηθεί για την βελτίωση της εκπαίδευσης των υπολοίπων. Στο πλάισο αυτό, το 2020, πραγματοποιήθηκε η πρώτη και μεγαλύτερη έρευνα για multi-task learning, πάνω και στις τρεις προσεγγίσεις συναισθηματικής αναπαράστασης [51,58]. Συγκεκριμένα, παρουσιάστηκε ένα Συνελικτικό Νευρωνικό Δίκτυο, το FaceBehaviorNet, το οποίο εκπαιδεύτηκε πάνω σε όλα τα δημόσια διαθέσιμα σύνολα δεδομένων που περιέχουν εικόνες και βίντεο ανθρώπινων εκφράσεων προσώπου, κάτω από μη-ελεγχόμενες συνθήκες (πάνω από 5 εκατομμύρια εικόνες). Τα αποτελέσματα έδειξαν πως το δίκτυο σημείωσε καλύτερες επιδόσεις από αντίστοιχα state-of-the-art μοντέλα που είχαν εκπαιδευτεί πάνω σε ένα μόνο task.



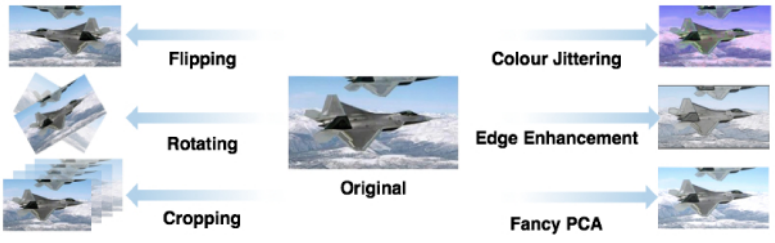
Σχήμα 1.7: Multi-task learning στο δίκτυο FaceBehaviorNet, αξιοποιώντας εικόνες με labels και από τις 3 συναισθηματικές προσεγγίσεις [58].

1.4 Επαύξηση δεδομένων και mixup

Όπως έχει ήδη αναφερθεί, τα Δίκτυα Βαθιάς Μάθησης αποτελούν πολύ ισχυρά υπολογιστικά εργαλεία που μπορούν να χρησιμοποιηθούν για την επίλυση σύνθετων προβλημάτων ανάλυσης και κατηγοριοποίησης. Ωστόσο, παρά την ισχύ τους, είναι ιδιαίτερα επιρρεπή στο φαινόμενο της υπερεκπαίδευσης (overfitting). Σύμφωνα με αυτό, το μοντέλο που εκπαιδεύεται εμφανίζει πολύ μικρό σφάλμα πάνω στα δεδομένα εκπαίδευσης (training data) αλλά υψηλό σφάλμα στα δεδομένα ελέγχου (test data). Διαπιστώνεται δηλαδή μεγάλο χάσμα στην επίδοση του δικτύου πάνω στα δύο αυτά σύνολα.

Η υπερεκπαίδευση εκδηλώνεται γενικά όταν υπάρχει μικρός αριθμός δειγμάτων εκπαίδευσης, οπότε το δίκτυο απομνημονεύει άψογα τα δεδομένα εισόδου αλλά αδυνατεί να γενικεύσει επιτυχημένα το γενικότερο πρόβλημα κατηγοριοποίησης, όταν του παρουσιάζονται καινούριες εικόνες. Ένας καλός τρόπος για να διορθώσουμε αυτό το πρόβλημα είναι να μεγαλώσουμε το σύνολο δεδομένων εκπαίδευσης, προσθέτοντας επιπλέον δεδομένα. Ωστόσο, κάτι τέτοιο δεν είναι πάντα εύκολο, ειδικά σε ένα πρόβλημα επιβλεπόμενης μάθησης όπου η συλλογή εικόνων και ετικετών μπορεί να είναι δαπανηρή αλλά και χρονοβόρα. Το σύνολο δεδομένων εκπαίδευσης μπορεί να μεγαλώσει και με τεχνητά δεδομένα που δημιουργούνται μέσω 3D τεχνικών όρασης [46,47] ή μέσω των Αναγεννητικών Ανταγωνιστικών Δικτύων [31,32,33,59] (Generative Adversarial Networks - GANs). Ωστόσο το τεχνητά δεδομένα θα πρέπει να είναι ρεαλιστικά (να μοιάζουν με πραγματικά στο ανθρώπινο και όχι μόνο μάτι), πράγμα που δεν είναι ιδιαίτερα εύκολο και αποτελεί ακόμη αντικείμενο έρευνας.

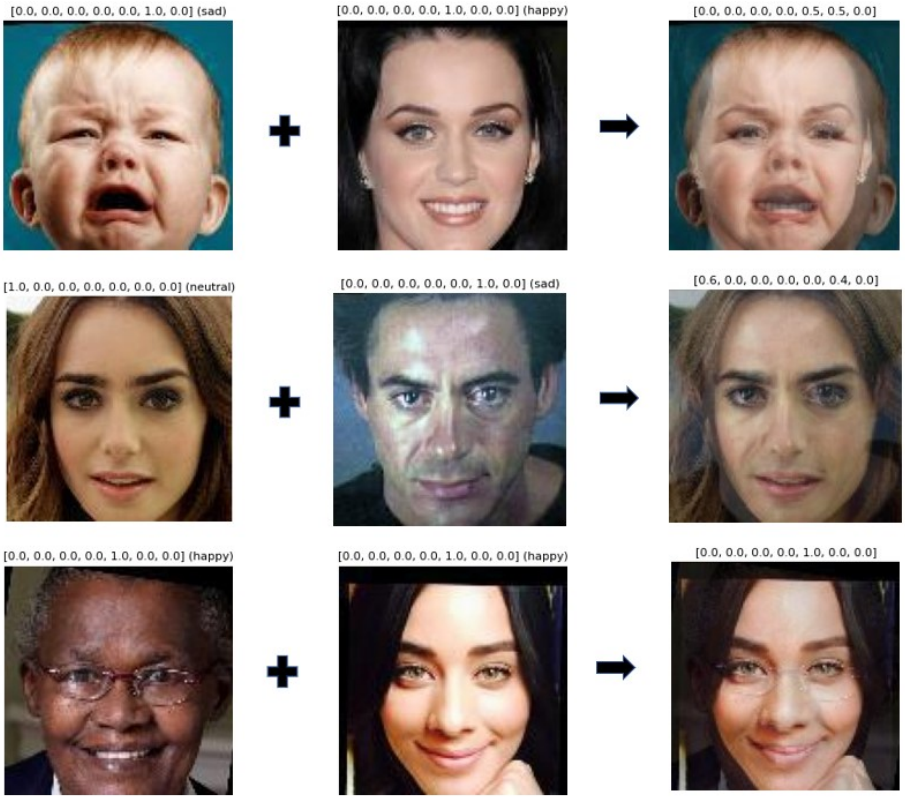
Για το λόγο αυτό, συχνά εφαρμόζονται τεχνικές επαύξησης δεδομένων (data augmentation), δηλαδή μέθοδοι οι οποίες παράγουν τεχνητά νέες εικόνες, αξιοποιώντας αυτές που υπάρχουν ήδη διαθέσιμες στο dataset. Οι πιο κλασικές από αυτές τις τεχνικές εφαρμόζουν αφινικούς μετασχηματισμούς (περιστροφές, μετατοπίσεις, κλιμακώσεις κτλ.) πάνω στις αρχικές εικόνες. Με τον τρόπο αυτό το δίκτυο εκτίθενται σε περισσότερες εκδόσεις των δεδομένων και αποκτά μια καλύτερη ικανότητα γενίκευσης.



Σχήμα 1.8: Τεχνικές επαύξησης δεδομένων για εικόνες [102]

Παρόλο που οι τυπικές αυτές μέθοδοι επαύξησης δεδομένων έχουν στεφθεί με ιδιαίτερη επιτυχία σε ένα μεγάλο αριθμό προβλημάτων, υπάρχουν περιπτώσεις στις οποίες δεν μπορούν να συνδράμουν αποτελεσματικά στην αντιμετώπιση του προβλήματος της υπερεκπαίδευσης. Για αυτό τα τελευταία χρόνια έχουν αναπτυχθεί πολλές νέες τεχνικές data augmentation.

Μια πρόσφατη και αρκετά απλή τεχνική επαύξησης δεδομένων, η οποία φαίνεται να δίνει ιδιαίτερα καλά αποτελέσματα σε προβλήματα κατηγοριοποίησης εικόνων, είναι η mixup [117]. Σύμφωνα με αυτή, η εκπαίδευση του δικτύου πραγματοποιείται πάνω σε κυρτούς συνδυασμούς από τυχαία ζεύγη των αρχικών εικόνων καθώς και των αντίστοιχων ετικετών τους. Με τον τρόπο αυτό, το δίκτυο κανονικοποιείται και παρουσιάζει καλύτερη ικανότητα γενίκευσης στα δεδομένα ελέγχου. Παράλληλα εμφανίζει μεγαλύτερη ευστάθεια στις προβλέψεις του καθώς και καλύτερη ανθεκτικότητα απέναντι σε ανταγωνιστικά παραδείγματα [87].



Σχήμα 1.9: Παραδείγματα εφαρμογής mixup σε εικόνες προσώπου της βάσης AffectNet [83]

Στο Σχήμα 1.9 μπορούμε να δούμε ορισμένα παραδείγματα από την εφαρμογή της mixup πάνω σε εικόνες εκφράσεων προσώπου της βάσης AffectNet. Όπως παρατηρούμε, οι νέες συνθετικές εικόνες που παράγονται (τελευταία στήλη), αποτελούν την ανάμειξη δύο τυχαίων αρχικών δειγμάτων με κατάλληλους συντελεστές βαρύτητας. Για παράδειγμα, στην πρώτη σειρά διακρίνουμε την ισόποση ανάμειξη μιας εικόνας ενός λυπημένου προσώπου με μία ενός χαρούμενου. Στην δεύτερη σειρά έχουμε την παρεμβολή της εικόνας ενός ουδέτερου προσώπου με μια άλλη ενός λυπημένου, σε αναλογία 60%-40%, ενώ στην τελευταία σειρά το συνδυασμό δύο χαρούμενων εκφράσεων.

Τα συνθετικά δείγματα που προκύπτουν δεν έχουν πάντοτε κάποια ιδιαίτερη φυσική σημασία, ωστόσο έχει αποδειχτεί πως η εκπαίδευση πάνω σε αυτά συνδράμει στην βελτίωση της γενίκευσης του μοντέλου. Πρόκειται για μια διαδικασία που εισάγει ελάχιστη υπολογιστική επιβάρυνση (overhead) και μπορεί να υλοποιηθεί σε λίγες μόνο γραμμές κώδικα.

1.5 Συνεισφορές

Στα πλαίσια της παρούσας διπλωματικής, εξετάζουμε την αποτελεσματικότητα της τεχνικής mixup στο πρόβλημα της συναισθηματικής αναγνώρισης σε πραγματικές, μη-ελεγχόμενες συνθήκες (in-the-wild). Συγκεκριμένα, πραγματοποιούμε διεξοδικές πειραματικές μελέτες, οι οποίες μας οδηγούν σε σημαντικές συνεισφορές:

- Εφαρμόζουμε τη μέθοδο mixup σε δύο διαφορετικές εκδόσεις της βάσης AffectNet [83]. Μάλιστα σε μία από αυτές οι εικόνες προσώπων δεν είναι ευθυγραμμισμένες, οπότε παρουσιάζουμε μια μέθοδο ευθυγράμμισης που βασίζεται στην εφαρμογή μετασχηματισμού ομοιότητας (similarity transformation).
- Εφαρμόζουμε τη μέθοδο mixup σε δύο εκδόσεις της βάσης RAF-DB [70], στις εικόνες των οποίων έχει εφαρμοσθεί ανίχνευση προσώπου με χρήση διαφορετικών detectors.
- Εξετάζουμε την επίδραση του dropout, μιας μεθόδου κανονικοποίησης του δικτύου, στην εφαρμογή της mixup και στις δύο βάσεις.
- Προτείνουμε μια παραλλαγή της κλασικής μεθόδου mixup, την AddMixup, την οποία και εφαρμόζουμε στην RAF-DB. Σύμφωνα με αυτή, σε κάθε batch εικόνων που τροφοδοτείται στο δίκτυο για εκπαίδευση δεν περιλαμβάνονται μόνο αναμειγμένες εικόνες αλλά και αντίστοιχες εικόνες του συνόλου εκπαίδευσης.
- Παραθέτουμε αναλυτική σύγκριση των τεχνικών mixup και Addmixup για τη βάση RAF-DB.
- Εξάγουμε στατιστικά στοιχεία για τις προβλέψεις των καλύτερων μοντέλων μας.

1.6 Δομή Διπλωματικής

Η παρούσα διπλωματική εργασία οργανώνεται συνολικά σε 6 κεφάλαια. Το Κεφάλαιο 1 είναι εισαγωγικό και αποσκοπεί στην κάλυψη ορισμένων βασικών εννοιών γύρω από τα Δίκτυα Βαθιάς Μάθησης και το πρόβλημα της αναγνώρισης συναισθήματος. Αναλυτικότερα, σε πρώτη φάση, δίνεται ο ορισμός του ανθρώπινου συναισθήματος καθώς και οι τρόποι με τους οποίους αυτό μπορεί να εκδηλωθεί. Στη συνέχεια, γίνεται αναφορά στα Δίκτυα Βαθιάς Μάθησης και στη συνεισφορά τους στο πρόβλημα της αυτόματης ανίχνευσης συναισθηματικών καταστάσεων. Αναλύονται τα τρία βασικά μοντέλα συναισθηματικής αναπαράστασης καθώς και η σχέση που υπάρχει μεταξύ τους. Παρουσιάζεται συνοπτικά η λογική της επαύξησης δεδομένων ενώ γίνεται και μια σύντομη αναφορά στην τεχνική mixup και στην εφαρμογή της στο πρόβλημα της αναγνώρισης συναισθήματος. Σαν τελευταίο μέρος παρουσιάζονται συνοπτικά οι συνεισφορές της παρούσας διπλωματικής.

Τα υπόλοιπα κεφάλαια της διπλωματικής οργανώνονται ως εξής:

- **Κεφάλαιο 2**: Στο κεφάλαιο αυτό πραγματοποιείται μια βιβλιογραφική ανασκόπηση γύρω από το πρόβλημα της κατηγοριοποίησης συναισθήματος στις 7 βασικές κατηγορίες καθώς και γύρω από εφαρμογές που αξιοποιούν την ιδέα της τεχνικής mixup.
- **Κεφάλαιο 3**: Στο κεφάλαιο αυτό γίνεται αρχικά μια αναφορά στις κυριότερες κατηγορίες βάσεων δεδομένων. Στη συνέχεια, παρουσιάζονται συνοπτικά οι πιο γνωστές βάσεις που περιλαμβάνουν εικόνες ή βίντεο εκφράσεων προσώπου. Αυτές διακρίνονται

σε δύο υποενότητες. Στην πρώτη, συμπεριλαμβάνονται βάσεις που έχουν διαμορφωθεί σε περιβάλλον εργαστηρίου, κάτω από αυστηρά καθορισμένες οδηγίες. Στη δεύτερη, αναλύονται βάσεις που έχουν δημιουργηθεί από εικόνες και βίντεο σε πραγματικές, μη-ελεγχόμενες, συνθήκες περιβάλλοντος.

- **Κεφάλαιο 4**: Στο κεφάλαιο αυτό αναλύεται σε βάθος η τεχνική mixup καθώς και μια παραλλαγή της, την οποία προτείνουμε, η Addmixup. Στη συνέχεια γίνεται μια αναφορά στην χρησιμότητα της Μεταφοράς Μάθησης, την οποία και αξιοποιούμε, ενώ τέλος, παρουσιάζεται η αρχιτεκτονική του δικτύου που χρησιμοποιούμε για τα πειράματα καθώς και οι υπερπαράμετροι που επιλέγονται.

- **Κεφάλαιο 5**: Στο κεφάλαιο αυτό παρουσιάζονται οι μετρικές αξιολόγησης που χρησιμοποιούνται, οι τεχνικές προεπεξεργασίας δεδομένων που εφαρμόζονται, τα αποτελέσματα της πειραματικής μας μελέτης καθώς και σχολιασμός αυτών.

- **Κεφάλαιο 6**: Στο κεφάλαιο αυτό γίνεται μια σύντομη ανακεφαλαίωση της πειραματικής μας μελέτης ενώ παρουσιάζονται και διάφορες μελλοντικές επεκτάσεις.

Κεφάλαιο 2

Σχετική Βιβλιογραφία

2.1 Αναγνώριση συναισθήματος in-the-wild

Η αυτόματη αναγνώριση συναισθηματικών καταστάσεων κάτω από μη ελεγχόμενες συνθήκες, in-the-wild, αποτελεί ένα πολύ βασικό χαρακτηριστικό των σύγχρονων συστημάτων αλληλεπίδρασης ανθρώπου-μηχανής και βρίσκεται στο επίκεντρο της επιστημονικής έρευνας. Γενικότερα, ευφυή ρομποτικά συστήματα, όπως ψηφιακοί βοηθοί, θα πρέπει να είναι σε θέση να ανιχνεύουν και να ερμηνεύουν άμεσα και με μεγάλη ακρίβεια οπτικά και ακουστικά σήματα που εκφράζουν κάποιο συναίσθημα, έτσι ώστε να αντιλαμβάνονται την ψυχική κατάσταση του χρήστη και να επικοινωνούν μαζί του με όσο το δυνατόν πιο φυσικό τρόπο.

Ωστόσο, μια τέτοια υλοποίηση είναι ιδιαίτερα απαιτητική καθώς σε περιβάλλον πραγματικού κόσμου υπάρχουν πολλοί αστάθμητοι παράγοντες, όπως το φύλο, η ηλικία, η εθνικότητα του χρήστη, η πόζα του κεφαλιού, οι συνθήκες φωτισμού, το υπόβαθρο κτλ. Πρόκειται δηλαδή για ένα αρκετά σύνθετο πρόβλημα, στο οποίο πρέπει να λάβει κανείς υπόψη του ποικίλες μεταβλητές παραμέτρους.

2.2 Κατηγοριοποίηση στα βασικά συναισθήματα

Τα τελευταία χρόνια έχουν πραγματοποιηθεί πολυάριθμες μελέτες για αναγνώριση συναισθηματικών καταστάσεων σε πραγματικές συνθήκες, οι οποίες εστιάζουν κυρίως σε εικόνες και βίντεο του διαδικτύου που διαθέτουν επισημειώσεις ως προς μία ή περισσότερες συναισθηματικές αναπαραστάσεις. Στην ενότητα αυτή επικεντρωνόμαστε στο πρόβλημα της κατηγοριοποίησης εκφράσεων προσώπου in-the-wild στα 7 βασικά συναισθήματα (κατηγορικό μοντέλο), παρουσιάζοντας διάφορες πρόσφατες πειραματικές μελέτες που έχουν πραγματοποιηθεί. Συγκεκριμένα, αναλύουμε ορισμένες τεχνικές που αξιοποιούν μόνο την εικόνα της έκφρασης του προσώπου, άλλες που εκμεταλλεύονται παράλληλα και την ακουστική πληροφορία, καθώς και άλλες που συνδυάζουν την εικόνα προσώπου με πληροφορία από τον περιβάλλοντα χώρο και την στάση του σώματος.

2.2.1 Αξιοποίηση μόνο έκφρασης προσώπου

Μια σημαντική πρόκληση που υπάρχει στην κατηγοριοποίηση συναισθήματος σε μη-ελεγχόμενες συνθήκες, σχετίζεται με το γεγονός ότι πρόκειται για ένα πρόβλημα μη-ισορροπημένης ταξινόμησης. Αυτό σημαίνει πως στα περισσότερα in-the-wild databases που υπάρχουν ετικέτες με τα βασικά συναισθήματα επικρατεί άνιση κατανομή στις επιμέρους κλάσεις. Στο πλαίσιο αυτό, στο [18], προτάθηκε μια αρχιτεκτονική δικτύου Βαθιάς Μάθησης, η DeepEmo,

η οποία κωδικοποιεί αναπαραστάσεις χαρακτηριστικών υψηλού επιπέδου για την αναγνώριση εκφράσεων προσώπου. Το δίκτυο αυτό εκπαιδεύτηκε πάνω στην μη-ισορροπημένη βάση δεδομένων RAF-DB [70] σημειώνοντας καλύτερη απόδοση από αντίστοιχες state-of-the-art μεθόδους αναπαράστασης δεδομένων και κατηγοριοποίησης ενώ προσέγγισε αρκετά την ανθρώπινη αντίληψη.

Αντίστοιχα, στο [84], μια ομάδα ερευνητών εκπαίδευσε δύο αρχιτεκτονικές DNNs, τις AlexNet [65] και WACV-Net [82] πάνω σε ένα άλλο σύνολο εικόνων εκφράσεων προσώπου από το διαδίκτυο. Οι εικόνες αυτές συγκεντρώθηκαν αξιοποιώντας τις μηχανές αναζήτησης Google, Bing και Yahoo με χρήση 1250 λέξεων-κλειδιά σε 6 διαφορετικές γλώσσες, ενώ επισημειώθηκαν ως προς τα 7 βασικά συναισθήματα από 2 υπνομνηματιστές. Τα αποτελέσματα των πειραμάτων έδειξαν πως τα δίκτυα αυτά είναι σε θέση να κατηγοριοποιούν αυθόρμητες εκφράσεις προσώπων, που έχουν συγκεντρωθεί κάτω από μη-ελεγχόμενες συνθήκες, πολύ καλύτερα από άλλες παραδοσιακές μεθόδους μηχανικής μάθησης (που χρησιμοποιούν περιγραφητές HOG [76] ή Gabor φίλτρα [71] για την εξαγωγή των σημαντικών χαρακτηριστικών).

Εκτός από τη μεγάλη ανομοιομορφία που υπάρχει στην έκφραση των συναισθημάτων από διαφορετικούς ανθρώπους (λόγω φύλου, ηλικίας, κοινωνικής θέσης κτλ), εμφανίζονται σημαντικές αποκλίσεις και στη δημιουργία ετικετών από διαφορετικούς υπομνηματιστές. Για το σκοπό αυτό είναι ιδιαίτερα χρήσιμος ο σχεδιασμός Δικτύων Βαθιάς Μάθησης τα οποία θα μπορούν να εκμεταλλεύονται, κατά την εκπαίδευσή τους, γνώσεις όχι μόνο από το σύνολο εκπαίδευσής αλλά και από κάποιο άλλο παρόμοιο facial dataset, για την βελτίωση της απόδοσής τους πάνω σε νέα δεδομένα.

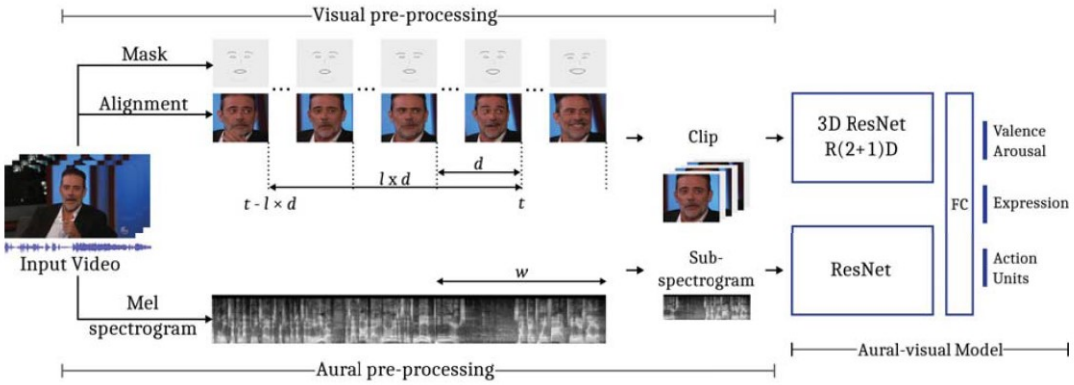
Μια τέτοια μέθοδος προτείνεται στο [57], όπου οι συγγραφείς αξιοποιούν το EmotiW 2017 training set καθώς και κάποια γνώση που εξάγουν από το αντίστοιχο validation set για το σχεδιασμό ενός DNN με βελτιωμένη ικανότητα γενίκευσης πάνω στο validation set. Αναλυτικότερα, προτείνεται μια μέθοδος domain adaptation, σύμφωνα με την οποία εξάγονται αντίστοιχες αναπαραστάσεις και από τα δύο σύνολα δεδομένων (train και validation) και έπειτα επιδιώκεται το ταίριασμα των στατιστικών κατανομών τους. Οι αναπαραστάσεις αυτές προκύπτουν εξάγοντας και χρησιμοποιώντας εσωτερικά χαρακτηριστικά από δύο Βαθιά Συνελικτικά Αναδρομικά Νευρωνικά Δίκτυα (CNN-RNN), τα οποία εκπαιδεύονται ξεχωριστά πάνω στο training και validation set αντίστοιχα. Έτσι λοιπόν, προτείνεται μια νέα συνάρτηση σφάλματος για την εκπαίδευση του δικτύου πάνω στα train δεδομένα του EmotiW dataset, η οποία περιλαμβάνει την ελαχιστοποίηση της διαφοράς των στατιστικών που παράγονται από το δίκτυο αυτό και το αντίστοιχο δίκτυο που έχει εκπαιδευτεί στα val δεδομένα.

2.2.2 Αξιοποίηση έκφρασης προσώπου και ήχου

Οι μέθοδοι αυτόματης αναγνώρισης συναισθήματος που εξετάσαμε προηγουμένως βασίζονται στην εκπαίδευση Βαθιών Νευρωνικών Δικτύων αποκλειστικά με εικόνες εκφράσεων προσώπου. Παρόλο που τέτοιες προσεγγίσεις μπορεί να σημειώσουν ιδιαίτερα καλά αποτελέσματα, τα τελευταία χρόνια έχουν αρχίσει να αξιοποιούνται multimodal τεχνικές, οι οποίες εκμεταλλεύονται παράλληλα τόσο την οπτική όσο και την ακουστική πληροφορία από οπτικοακουστικές βάσεις. Σε αυτό το πλαίσιο, το 2020 πραγματοποιήθηκε ο πρώτος διαγωνισμός αυτόματης ανίχνευσης και των τριών συναισθηματικών αναπαραστάσεων (κατηγοριοποίηση βασικών συναισθημάτων, ανίχνευσης AUs, εκτίμηση Valence-Arousal) in-the-wild, γνωστός ως ABAW [49]. Ως database χρησιμοποιήθηκε η Aff-Wild2 [49, 55, 56, 58, 60], μια βάση δεδομένων με συνολικά 558 βίντεο από το διαδίκτυο.

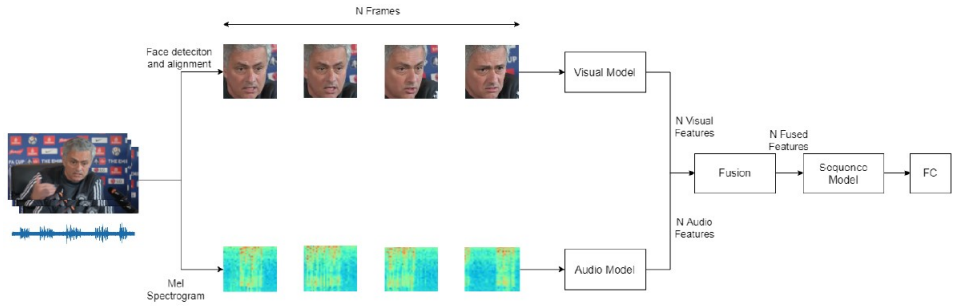
Στο [66] παρουσιάζεται το δίκτυο που σημείωσε την καλύτερη επίδοση στο διαγωνισμό

για το πρόβλημα της κατηγοριοποίησης στις 7 βασικές κλάσεις. Συγκεκριμένα, αναλύεται η αρχιτεκτονική ενός 3D Συνελικτικού Νευρωνικού Δικτύου δύο ροών, το οποίο επεξεργάζεται ξεχωριστά την οπτική και την ηχητική πληροφορία των δεδομένων. Αρχικά, κάθε βίντεο της βάσης που δίνεται ως είσοδος στο δίκτυο χωρίζεται σε εικόνες και ήχο. Στη συνέχεια, για κάθε καρέ πραγματοποιείται ανίχνευση και ευθυγράμμιση προσώπου ενώ εξάγεται και μια μάσκα με το περίγραμμα των ματιών, τη μύτη, το πιγούνι, τα φρύδια και το περίγραμμα των χειλιών. Η εικόνα μαζί με την αντίστοιχη μάσκα της δίνονται ως είσοδος σε ένα 3D-CNN υποδίκτυο ενώ το αντίστοιχο φασματογράφημα που περιέχει την ακουστική πληροφορία τροφοδοτείται σε ένα δεύτερο CNN υποδίκτυο. Στο τελευταίο επίπεδο της ενιαίας αρχιτεκτονικής οι αναπαραστάσεις από τις δύο ξεχωριστές ροές συνδέονται μεταξύ τους προτού πραγματοποιηθεί η τελική ταυτόχρονη πρόβλεψη και για τα 3 tasks, όπως φαίνεται στο Σχήμα 2.1.



Σχήμα 2.1: 3D οπτικοακουστικό δίκτυο με δύο ανεξάρτητες ροές. Κάθε βίντεο χωρίζεται σε εικόνες και ήχο, με τις δύο αυτές πληροφορίες να κωδικοποιούνται ανεξάρτητα από ένα υποδίκτυο η καθεμία. Στο τελευταίο επίπεδο της αρχιτεκτονικής οι δύο ροές συνδέονται μεταξύ τους προτού πραγματοποιηθεί η τελική πρόβλεψη [66].

Στο ίδιο πλαίσιο, στο διαγωνισμό ABAW2 [61], ο οποίος αποτελεί συνέχεια του πρώτου ABAW διαγωνισμού, πολλές από τις ομάδες που έλαβαν μέρος, πρότειναν εξαιρετικές μεθόδους αναγνώρισης συναισθήματος και για τις τρεις διαφορετικές αναπαραστάσεις. Με βάση το [43], μια εκ των καλύτερων ομάδων πρότεινε μια πολυτροπική μέθοδο η οποία αξιοποιεί οπτική αλλά και ακουστική πληροφορία από τα βίντεο της βάσης, ενώ εκπαιδεύεται παράλληλα πάνω σε δύο διαφορετικά tasks. Συγκεκριμένα, παρουσιάστηκε ένα οπτικοακουστικό δίκτυο το οποίο διαθέτει δύο ροές πληροφορίας, μια για την κωδικοποίηση των εικόνων προσώπου και μια για την κωδικοποίηση του ήχου. Σε πρώτο στάδιο, πραγματοποιείται εκπαίδευση του οπτικού μοντέλου ξεχωριστά, αξιοποιώντας τις εικόνες που είναι επισημειωμένες με AUs και με τα 7 βασικά συναισθήματα. Σε δεύτερο στάδιο, πραγματοποιείται πάγωμα των παραμέτρων του οπτικού μοντέλου και παράλληλα προστίθεται η ροή ακουστικής πληροφορίας για την εξαγωγή χρήσιμων ηχητικών αναπαραστάσεων. Τέλος, τα παραγόμενα χαρακτηριστικά ήχου και εικόνας συγχωνεύονται και τροφοδοτούνται σε ένα μετασχηματιστή κωδικοποίησης για περαιτέρω εξαγωγή χρονικών χαρακτηριστικών, όπως φαίνεται στο Σχήμα 2.2.

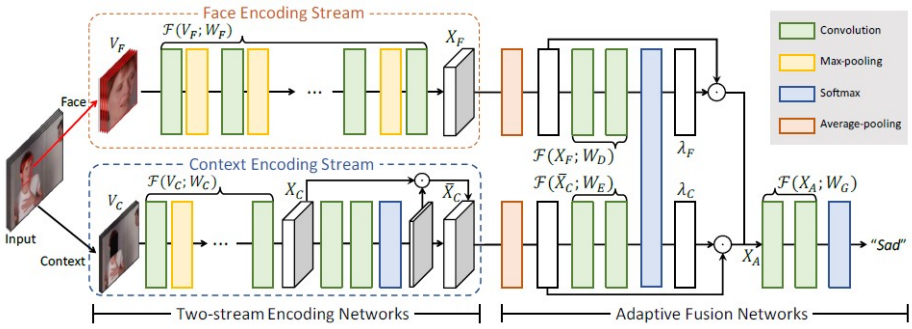


Σχήμα 2.2: Οπτικοακουστικό δίκτυο με δύο ανεξάρτητες ροές πληροφορίας. Το βίντεο εισόδου χωρίζεται σε εικόνες και ήχο και κάθε πληροφορία κωδικοποιείται ξεχωριστά από ένα υποδίκτυο. Στο τελευταίο επίπεδο της αρχιτεκτονικής οι δύο ροές συνδέονται πριν την τελική πρόβλεψη [43].

2.2.3 Αξιοποίηση έκφρασης προσώπου και περιβάλλοντος

Όπως έχουμε ήδη αναφέρει, το ανθρώπινο συναίσθημα δεν μπορεί πάντοτε να γίνει πλήρως αντιληπτό μόνο από την έκφραση του προσώπου. Στην καθημερινή μας ζωή, για την κατανόηση και ερμηνεία των συναισθημάτων, συχνά λαμβάνουμε υποσυνείδητα υπόψη μας και το εξωτερικό περιβάλλον, δηλαδή το χώρο που περιβάλλει κάποιο άτομο καθώς και πιθανές αλληλεπιδράσεις του με άλλα άτομα.

Ορισμένες μέθοδοι [15, 62] έχουν αποδείξει πως η ακρίβεια αυτόματης αναγνώρισης συναισθημάτων μπορεί να αυξηθεί σημαντικά αν, εκτός από το πρόσωπο, ληφθεί υπόψη και πληροφορία από το υπόλοιπο τμήμα της εικόνας. Στο πλαίσιο αυτό, στο [68], οι συγγραφείς πρότειναν μια αρχιτεκτονική Βαθιού Νευρωνικού Δικτύου, τη CAER-Net, η οποία πέρα από την έκφραση του προσώπου αξιοποιεί και πληροφορία από τον περιβάλλοντα χώρο με ένα συνδυαστικό και ενισχυτικό τρόπο. Αναλυτικότερα, το εν λόγω δίκτυο αποτελείται από δύο ξεχωριστά υποδίκτυα, ένα για την κωδικοποίηση του προσώπου και ένα για την κωδικοποίηση του πλαισίου που το περιβάλλει. Στόχος είναι η αναζήτηση και εξαγωγή χρήσιμων πληροφοριών και από τις δύο αυτές περιοχές ενδιαφέροντος. Για την εξαγωγή πληροφορίας από το υπόβαθρο της εικόνας, αποκρύπτεται το πρόσωπο και αναζητούνται κυρίαρχα χαρακτηριστικά στο υπόλοιπο μέρος, με χρήση ενός μηχανισμού προσοχής. Οι δύο ανεξάρτητες κωδικοποιήσεις συνδυάζονται μέσω ενός δικτύου προσαρμοστικής σύντηξης (adaptive fusion network) για την συγχώνευση των χαρακτηριστικών πριν από την τελική πρόβλεψη, όπως φαίνεται στο Σχήμα 2.3. Το CAER-Net δοκιμάστηκε πάνω σε διάφορα in-the-wild facial datasets και αποδείχθηκε πως αποδίδει καλύτερα σε σχέση με αντίστοιχα CNNs που εκμεταλλεύονται μόνο την οπτική πληροφορία από την έκφραση του προσώπου.

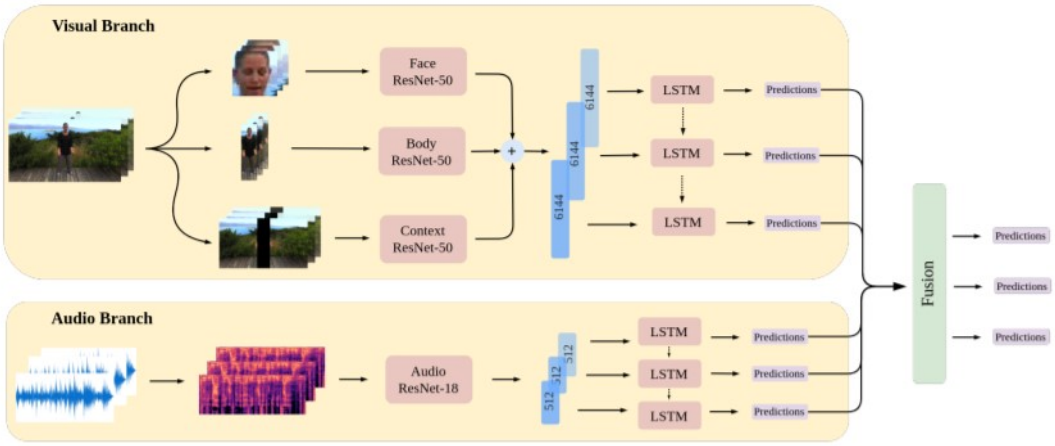


Σχήμα 2.3: Αρχιτεκτονική δικτύου CAER-Net, η οποία αποτελείται από δύο υποδίκτυα κωδικοποίησης καθώς και ένα προσαρμοστικό δίκτυο σύντηξης [68].

2.2.4 Αξιοποίηση προσώπου,σώματος,περιβάλλοντος και ήχου

Σε όλες τις προαναφερθείσες μεθόδους παρουσιάσαμε δίκτυα που εκμεταλλεύονται είτε μόνο την οπτική πληροφορία των εικόνων προσώπου είτε αυτή σε συνδυασμό με τον περιβάλλοντα χώρο ή την ηχητική πληροφορία. Στο πλαίσιο του διαγωνισμού ABAW2 [61], μια ομάδα ερευνητών από το Εθνικό Μετσόβιο Πολυτεχνείο [4] πρότεινε μια αρχιτεκτονική δικτύου η οποία αξιοποιεί παράλληλα πληροφορία από την εικόνα προσώπου, το σώμα, το περιβάλλον και τον ήχο.

Αναλυτικότερα, πρόκειται για ένα δίκτυο που χωρίζεται σε δύο επιμέρους υποδίκτυα, ένα για την εκμετάλλευση της RGB οπτικής πληροφορίας της ακολουθίας εικόνων και ένα για την επεξεργασία των αντίστοιχων ηχητικών αποσπασμάτων. Το οπτικό υποδίκτυο περιλαμβάνει 3 ανεξάρτητες ροές, μία για την εξαγωγή χαρακτηριστικών από την έκφραση του προσώπου, μία για την επεξεργασία σωματικών εκφράσεων και μία για την ανάλυση του περιβάλλοντα χώρου. Από την άλλη, το ακουστικό υποδίκτυο είναι υπεύθυνο για την κωδικοποίηση των αντίστοιχων φασματογραφημάτων. Και στα δύο υποδίκτυα χρησιμοποιούνται CNNs για την εξαγωγή χρήσιμων χαρακτηριστικών και στη συνέχεια LSTMs για την αντιστοίχηση των χαρακτηριστικών αυτών σε ετικέτες. Μια οπτικοποίηση της συνολικής αρχιτεκτονικής φαίνεται στο Σχήμα 2.4.



Σχήμα 2.4: Αρχιτεκτονική δικτύου με ένα οπτικό και ένα ακουστικό υποδίκτυο. Το πρώτο είναι υπεύθυνο για την κωδικοποίηση του προσώπου, του σώματος και του περιβάλλοντος ενώ το δεύτερο για την κωδικοποίηση της αντίστοιχης ακουστικής πληροφορίας [4].

Το δίκτυο χρησιμοποιήθηκε για κατηγοριοποίηση συναισθημάτων στις 7 βασικές κλάσεις αλλά και για εκτίμηση των συναισθηματικών συνιστωσών Valence-Arousal. Τα αποτελέσματα έδειξαν πως η σύντηξη διαφορετικών ροών πληροφορίας οδηγεί σε σημαντική βελτίωση της απόδοσης, συγκριτικά με αρχιτεκτονικές που αξιοποιούν μόνο μια εκ των ροών αυτών.

2.3 Επαύξηση δεδομένων με mixup και παραλλαγές

Στην προηγούμενη ενότητα παρουσιάστηκαν διάφορες πρόσφατες πειραματικές μελέτες για κατηγοριοποίηση συναισθημάτων στις 7 βασικές κλάσεις. Τώρα πραγματοποιούμε μια βιβλιογραφική ανασκόπηση για εφαρμογές της τεχνικής mixup [117], καθώς και για διάφορες παραλλαγές της, στην κατηγοριοποίηση εικόνων, στην σημασιολογική κατάτμηση ιατρικών εικόνων, στην επεξεργασία φυσικής γλώσσας αλλά και στην κατηγοριοποίηση ηχητικών καταγραφών.

2.3.1 Mixup στην κατηγοριοποίηση εικόνων

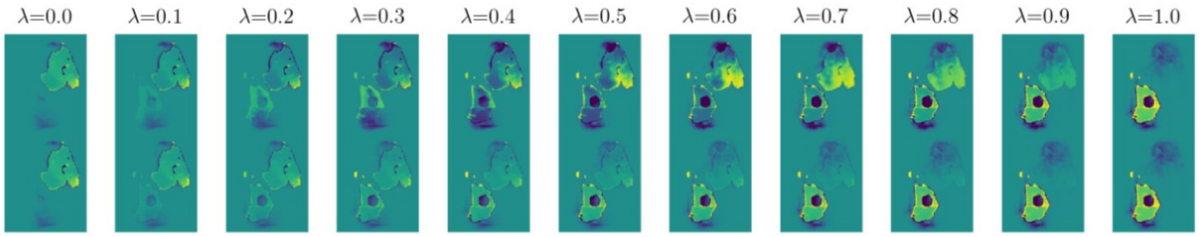
Η τεχνική mixup είναι περισσότερο διαδεδομένη σε προβλήματα κατηγοριοποίησης εικόνων. Η ομάδα ερευνητών που πρότεινε την εν λόγω τεχνική, στο [117], διεξήγαγε μια σειρά από πειράματα ταξινόμησης πάνω σε μεγάλες στατικές βάσεις δεδομένων, όπως η ImageNet [93], η CIFAR-10 [63] και η CIFAR-100 [64]. Τα αποτελέσματα έδειξαν πως και στα 3 classification tasks, η εφαρμογή της mixup οδηγεί σε καλύτερη ικανότητα γενίκευσης των μοντέλων, συγκριτικά με την κλασική εκπαίδευση με Ελαχιστοποίηση Εμπειρικού Ρίσκου. Μάλιστα διαπιστώθηκε πως η mixup συνδράμει ακόμα περισσότερο όταν εφαρμόζεται σε πιο βαθιές αρχιτεκτονικές Νευρωνικών Δικτύων (π.χ. ResNet-101 αντί για ResNet-50 [38]) καθώς επίσης και όταν η διαδικασία της εκπαίδευσης διαρκεί περισσότερες εποχές.

Το 2021, στο πλαίσιο αντιμετώπισης της πανδημίας του COVID-19, διεξήχθη ο διαγωνισμός MIA-COV19D [45, 47, 52, 54], με στόχο την αυτόματη διάγνωση της νόσου από τρισδιάστατες αξονικές τομογραφίες θώρακα. Αναλυτικότερα, οι διαγωνιζόμενοι κλήθηκαν να προτείνουν αρχιτεκτονικές Νευρωνικών Δικτύων που θα μπορούν να αποφανθούν με μεγάλη ακρίβεια αν μια δεδομένη τομογραφία προέρχεται από άνθρωπο που πάσχει από COVID-19 ή όχι. Η νικήτρια ομάδα [40], χρησιμοποιώντας ένα προεκπαιδευμένο μοντέλο ResNet-50, ανέπτυξε μια προσαρμοστική συνδυαστική στρατηγική εκπαίδευσης, η οποία ενσωματώνει τη μέθοδο επαύξηση δεδομένων mixup. Η χρήση της mixup ευνόησε την εξαγωγή χρήσιμων χαρακτηριστικών που σχετίζονται με την ασθένεια και κατά συνέπεια συνέβαλε στην βελτίωση της ακρίβειας δυαδικής κατηγοριοποίησης.

Στο πρόβλημα της κατηγοριοποίησης συναισθημάτων από εικόνες εκφράσεων προσώπου, το οποίο και θα εξετάσουμε στο Κεφάλαιο 5, η μοναδική εφαρμογή της mixup παρατηρείται στο [91], όπου μια ομάδα ερευνητών αξιοποίησε τη συγκεκριμένη μέθοδο για τη βελτίωση της απόδοσης ενός Βαθιού Νευρωνικού Δικτύου. Αναλυτικότερα, οι συγγραφείς πρότειναν μια αρχιτεκτονική CNN, την eXnet, την οποία εκπαίδευσαν και αξιολόγησαν πάνω στα σύνολα δεδομένων FER-2013 [33], CK+ [72] και RAF-DB [70]. Παρατηρήθηκε πως με χρήση της mixup σημειώνεται αύξηση του Accuracy της τάξεως του 1% σε όλες τις βάσεις.

2.3.2 Mixup στην σημασιολογική κατάτμηση εικόνων

Όπως είπαμε, η τεχνική mixup χρησιμοποιείται κυρίως για κατηγοριοποίηση "ολόκληρων" εικόνων σε κλάσεις. Στο [22], φαίνεται πως αξιοποιείται για πρώτη φορά στο πρόβλημα της σημασιολογικής κατάτμησης ιατρικών εικόνων. Πρόκειται για μια ιδιαίτερη πρόκληση αν αναλογιστεί κανείς τον περιορισμένο αριθμό επισημειωμένων datasets στον τομέα της Ιατρικής. Στο πλαίσιο αυτό, πέραν από την εφαρμογή της mixup, οι συγγραφείς προτείνουν και μια παραλλαγή της, την οποία ονομάζουν mixmatch. Σύμφωνα με αυτή, προτού επιλεγούν οι εικόνες που πρόκειται να αναμειχθούν, λαμβάνεται υπόψη η κατανομή των δειγμάτων στις κλάσεις, σε αντίθεση με την παραδοσιακή μέθοδο, όπου η επιλογή πραγματοποιείται τυχαία. Συγκεκριμένα, συνδυάζονται μεταξύ τους εικόνες που εμφανίζουν υψηλή συγκέντρωση στο προσκήνιο με άλλες που εμφανίζουν χαμηλή. Η ιδέα αυτή προήλθε από το γεγονός ότι οι ιατρικές εικόνες είναι συνήθως μη ισορροπημένες. Για παράδειγμα στο BraTS2017 [5, 79] (επισημειωμένο σύνολο εικόνων μαγνητικής τομογραφίας που απεικονίζουν γλοιώματα του εγκεφάλου), το οποίο οι συγγραφείς αξιοποιούν για τα πειράματα, η πλειοψηφία των pixels μιας εικόνας δεν αποτελούν τμήμα κάποιου όγκου. Για την εκτέλεση των πειραμάτων χρησιμοποιήθηκε το δίκτυο [109]. Οι τεχνικές mixup και mixmatch εμφάνισαν παρόμοια επίδοση, ωστόσο και οι δύο απέδωσαν καλύτερα σε σχέση με την εκπαίδευση χωρίς επαύξηση δεδομένων.

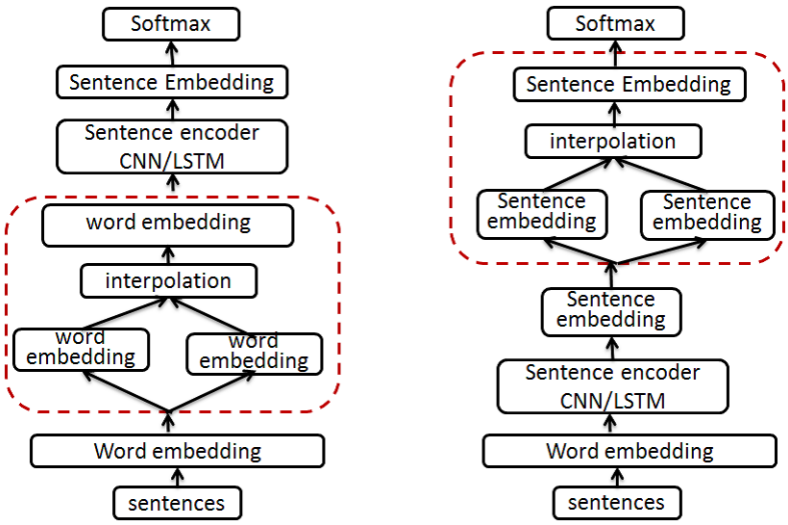


Σχήμα 2.5: Παράδειγμα μίξης ιατρικών εικόνων με αναλογία ανάμειξης λ μεταξύ 0 και 1 [22].

2.3.3 Mixup στην Επεξεργασία Φυσικής Γλώσσας

Η κατηγοριοποίηση κειμένων αποτελεί ένα από τα πιο θεμελιώδη προβλήματα στον τομέα της Επεξεργασίας Φυσικής Γλώσσας (NLP). Σε αντίθεση με τις εικόνες που αποτελούνται από pixels, οι προτάσεις αποτελούνται από ακολουθίες λέξεων, για τις οποίες δεν ορίζονται κανόνες διάταξης ή αλγεβρικές πράξεις. Για το λόγο αυτό, για να μπορέσει να εφαρμοστεί η τεχνική mixup απαιτείται σε πρώτο στάδιο η διανυσματική απεικόνιση όλων των προτάσεων. Σε ένα τυπικό CNN ή LSTM μοντέλο, μια πρόταση αρχικά αναπαρίσταται από μια ακολουθία διανυσματικών αναπαραστάσεων λέξεων (word embeddings) και στη συνέχεια τροφοδοτείται σε έναν κωδικοποιητή προτάσεων. Οι πιο δημοφιλείς τέτοιοι κωδικοποιητές είναι CNN και LSTM. Τα embeddings των προτάσεων που παράγονται από το CNN ή το LSTM στη συνέχεια τροφοδοτούνται σε ένα επίπεδο softmax για την δημιουργία της κατανομής πιθανότητας πάνω σε όλες τις υποψήφιες κλάσεις.

Στο [36] μια ομάδα ερευνητών πρότεινε δύο παραλλαγές της mixup που βρίσκουν εφαρμογή στην κατηγοριοποίηση προτάσεων: Η πρώτη πραγματοποιεί παρεμβολή δειγμάτων στο χώρο των word embeddings και ονομάζεται wordMixup, ενώ η δεύτερη πραγματοποιεί την ανάμειξη στο τελευταίο κρυφό επίπεδο του δικτύου, μόλις πριν το επίπεδο softmax, δηλαδή στο χώρο των sentence embeddings, και ονομάζεται senMixup.



Σχήμα 2.6: Αριστερά, εντός του κόκκινου ορθογωνίου που σημειώνεται με διακεκομμένες γραμμές, απεικονίζεται η τεχνική wordMixup ενώ δεξιά η senMixup [36].

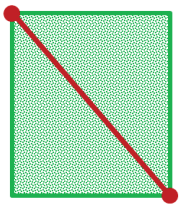
Στην τεχνική wordMixup, πραγματοποιείται zero padding σε όλες τις προτάσεις ώστε να αποκτήσουν το ίδιο μήκος και έπειτα η παρεμβολή πραγματοποιείται για κάθε διάσταση κάθε μιας από τις λέξεις μέσα σε μια πρόταση. Στην τεχνική senMixup, τα κρυφά embeddings (με την ίδια διάσταση) για τις δύο προτάσεις παράγονται από έναν κωδικοποιητή, όπως CNN ή LSTM, και έπειτα συνδυάζονται με γραμμικό τρόπο.

Οι δύο αυτές παραλλαγές της mixup εφαρμόστηκαν σε 5 διαφορετικά προβλήματα κατηγοριοποίησης προτάσεων. Τα αποτελέσματα των πειραμάτων έδειξαν πως οδηγούν σε σημαντική βελτίωση της ακρίβειας κατηγοριοποίησης τόσο σε CNN όσο και σε LSTM μοντέλα.

Παρά τη μεγάλη αποτελεσματικότητα των δύο αυτών των μεθόδων, η ισχύς τους παραμένει σχετικά περιορισμένη λόγω της γραμμικής τους φύσης. Συγκεκριμένα, όπως αναφέραμε και προηγουμένως, τόσο η wordMixup όσο και η senMixup αποτελούν πιστή εφαρμογή της παραδοσιακής τεχνικής mixup [117] στον τομέα του NLP. Εφαρμόζουν δηλαδή κυρτούς συνδυασμούς στα δείγματα εισόδου και στις αντίστοιχες ετικέτες τους. Αυτό περιορίζει σημαντικά τον χώρο των παραγόμενων εικονικών δειγμάτων και κατά συνέπεια και τον βαθμό κανονικοποίησης κατά την εκπαίδευση του δικτύου.

Με αφορμή την αδυναμία αυτή, στο [35], ο Guo εξελίσσει περαιτέρω την τεχνική wordMixup, προτείνοντας μια μη-γραμμική προσέγγιση, η οποία αναμειγνύει επίσης ζεύγη δειγμάτων (προτάσεις) στο επίπεδο των word embeddings. Ωστόσο, σε αντίθεση με την wordMixup, όπου όλες οι λέξεις σε ένα ζεύγος προτάσεων συνδυάζονται με μια βαθμωτή σταθερά αναλογίας λ, στη μη-γραμμική μέθοδο εφαρμόζεται διαφορετική ανάμειξη σε κάθε μία από τις διαστάσεις κάθε λέξης.

Ας πάρουμε το παράδειγμα ενός ζεύγους τυχαίων σημείων σε ένα δισδιάστατο σύνολο δεδομένων εκπαίδευσης. Η παραδοσιακή εκδοχή της mixup μπορεί να παράξει συνθετικά παραδείγματα κατά μήκος της γραμμής που ενώνει τα σημεία. Από την άλλη, η μη-γραμμική προσέγγιση παρέχει για κάθε συνιστώσα του ζεύγους σημείων μια ανεξάρτητη αναλογία ανάμειξης, επεκτείνοντας έτσι τον χώρο των παραγόμενων εικονικών δειγμάτων σε μια ευρεία περιοχή μεταξύ των δύο σημείων, όπως φαίνεται δεξιά στο Σχήμα 2.7.

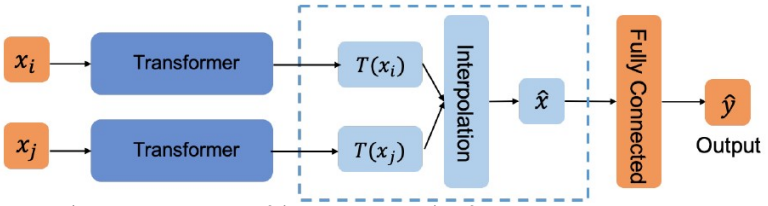


Σχήμα 2.7: Απεικόνιση του χώρου των παραγόμενων δειγμάτων για την παραδοσιακή τεχνική mixup αλλά και για τη μη-γραμμική προσέγγιση. Στην mixup, τα συνθετικά παραδείγματα προκύπτουν κατά μήκος της κόκκινης γραμμής που ενώνει το ζεύγος σημείων που αναμειγνύονται ενώ στην μη-γραμμική υλοποίηση μπορούν να παραχθούν σημεία οπουδήποτε εντός του πράσινου ορθογωνίου. [35]

Αξίζει να σημειωθεί οτι, σε αντίθεση με την κλασική εκδοχή της mixup, όπου τα δεδομένα εισόδου και τα αντίστοιχα labels συνδυάζονται με τον ίδιο ακριβώς τρόπο, η προσέγγιση αυτή περιλαμβάνει διαφορετική μη-γραμμική παρεμβολή για τις ετικέτες. Συγκεκριμένα, πραγματοποιείται μια προσαρμοστική εκμάθηση του τρόπου ανάμειξής τους, η οποία εξαρτάται από την παρεμβολή των προτύπων εισόδου.

Η μη-γραμμική mixup τεχνική εφαρμόστηκε στα πέντε προβλήματα κατηγοριοποίησης προτάσεων που εξετάστηκαν και οι μέθοδοι wordMixup και senMixup. Τα αποτελέσματα των πειραμάτων κατέδειξαν πως η συγκεκριμένη προσέγγιση υπερτερεί έναντι της mixup και των παραλλαγών της, επιτυγχάνοντας ακόμα καλύτερη ακρίβεια ταξινόμησης.

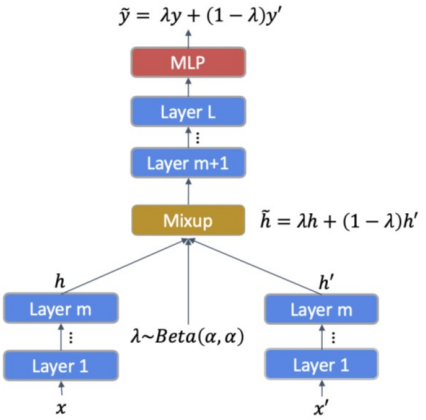
Μια βελτίωση της τεχνικής του senMixup παρουσιάζεται στο [99], όπου οι συγγραφείς, δεν εφαρμόζουν κάποια από τις παραδοσιακές μεθόδους κωδικοποίησης, όπως bag-of-words ή embeddings προερχόμενα από CNN/LSTM. Αντίθετα, για την εκμάθηση των αναπαραστάσεων των δεδομένων εισόδου (προτάσεων), χρησιμοποιούν προεκπαιδευμένα γλωσσικά μοντέλα μετασχηματιστών. Η διαδικασία της mixup πραγματοποιείται ταυτόχρονα με την ρύθμιση των παραμέτρων του δικτύου και έτσι προκύπτουν δυναμικά εικονικές κρυφές αναπαραστάσεις κατά τη διάρκεια της εκπαίδευσης.



Σχήμα 2.8: Οι προτάσεις x_i και x_j δίνονται ως είσοδος στον μετασχηματιστή T, οπότε στην έξοδο λαμβάνονται οι διανυσματικές κωδικοποιήσεις $T(x_i)$ και $T(x_j)$, οι οποίες στη συνέχεια παρεμβάλλονται γραμμικά, ώστε να προκύψει η mixed πρόταση $\hat{x} = \lambda T(x_i) + (1-\lambda)T(x_j)$ [99].

Η μέθοδος αυτή εφαρμόστηκε πάνω σε οκτώ διαφορετικά NLP tasks του GLUE [108]. Τα αποτελέσματα καταδεικνύουν πως επιφέρει μια βελτίωση της τάξεως του 1% σε όλες τις περιπτώσεις.

Στο ίδιο πλαίσιο, ο Chen πρότεινε μια παρόμοια τεχνική επαύξησης δεδομένων, την TMix [16], η οποία επίσης παρεμβάλει αναπαραστάσεις κρυφών επιπέδων ενός κωδικοποιητή.



Σχήμα 2.9: Τεχνική TMix [16]: Τα δείγματα κειμένου x και x' δίνονται ως είσοδο σε έναν κωδικοποιητή L στρωμάτων. Στο επίπεδο m, όπου $m \in [0, L]$ πραγματοποιείται παρεμβολή των κρυφών αναπαραστάσεων και έπειτα η αναμειγμένη αναπαράσταση τροφοδοτείται στα ανώτερα επίπεδα του μοντέλου.

Η διαφορά της συγκεκριμένης τεχνικής σε σχέση με την προηγούμενη [99] έγκειται στο στάδιο κωδικοποίησης που εκτελείται η mixup. Αναλυτικότερα, στη μέθοδο TMix, πραγματοποιείται ανάμειξη κρυφών αναπαραστάσεων των δεδομένων εισόδου σε κάποιο ενδιάμεσο επίπεδο του κωδικοποιητή, σε αντίθεση με την [99] όπου η παρεμβολή λαμβάνει χώρα στο τελευταίο κρυφό επίπεδο, ακριβώς πριν από την softmax. Θα μπορούσαμε να πούμε ότι πρόκειται για μια γενίκευση της προηγούμενης προσέγγισης, κατά την οποία η ανάμειξη μπορεί να πραγματοποιηθεί σε κάθε ένα από τα κρυφά επίπεδα κωδικοποίησης.

Με την μέθοδο TMix είναι δυνατόν πρακτικά να παραχθεί απεριόριστος αριθμός νέων δειγμάτων, γεγονός που μπορεί να συντελέσει σημαντικά στην αντιμετώπιση του προβλήματος της υπερεκπαίδευσης, ειδικά όταν το σύνολο δεδομένων αποτελείται από μικρό αριθμό δειγμάτων. Επίσης, η συγκεκριμένη τεχνική μπορεί να αξιοποιηθεί για ημι-επιβλεπόμενη μάθηση (semi-supervised learning) στο πρόβλημα της κατηγοριοποίησης κειμένου. Αυτό είναι ιδιαίτερα σημαντικό αν αναλογιστεί κανείς το αυξημένο κόστος και τη δυσκολία επισημείωσης ενός μεγάλου όγκου πληροφοριών.

2.3.4 Mixup σε δεδομένα ακουστικών σημάτων

Η ανάθεση ετικετών σε ηχητικές ηχογραφήσεις (audio tagging) έχει προσελκύσει ιδιαίτερο ενδιαφέρον την τελευταία δεκαετία καθώς βρίσκει εφαρμογή σε πολλούς τομείς της καθημερινής ζωής, όπως π.χ. στα συστήματα προτάσεων [12]. Ο αριθμός επισημειωμένων ακουστικών κλιπ που υπάρχει αυτή τη στιγμή διαθέσιμος είναι αρκετά περιορισμένος, μιας και η διαδικασία δημιουργίας labels είναι ιδιαίτερα χρονοβόρα και κουραστική όταν εκτελείται από τον άνθρωπο. Σε αυτή την κατεύθυνση μπορούν να συνδράμουν σημαντικά τα Δίκτυα Βαθιάς Μάθησης, τα οποία έχουν την ικανότητα να ανιχνεύουν αυτόματα σύνθετα χαρακτηριστικά των ακουστικών σημάτων. Ωστόσο, επειδή τα περισσότερα datasets περιορίζονται σε διάρκεια μόνο ορισμένων ωρών [81], η εκπαίδευση DNNs πάνω σε αυτά παρουσιάζει συχνά το φαινόμενο της υπερεκπαίδευσης.

Για την αντιμετώπιση του προβλήματος, έχουν εφαρμοστεί αρκετές κλασικές μέθοδοι επαύξησης δεδομένων, όπως χρονική επέκταση, προσθήκη θορύβου καθώς και συμπίεση δυναμικού εύρους, ωστόσο καμία δεν φαίνεται να επιφέρει σημαντική άνοδο στην απόδοση της ακουστικής μοντελοποίησης [89]. Με αφορμή το γεγονός αυτό, στο [110] οι συγγραφείς εφαρμόζουν την τεχνική mixup καθώς και δύο παραλλαγές της, την μέθοδο SamplePairing [41] και την Extrapolation [19], σε δεδομένα οικιακών ηχητικών καταγραφών.

Σε αναλογία με την mixup, στην SamplePairing λαμβάνεται ο μέσος όρος δύο τυχαίων δεδομένων εισόδου για τη δημιουργία ενός νέου συνθετικού δείγματος. Η βασική διαφοροποίηση της μεθόδου έγκειται στον ορισμό της ετικέτας του αναμειγμένου δείγματος. Συγκεκριμένα, αυτή ταυτίζεται με την ετικέτα του πρώτου από τα δείγματα που παρεμβάλλονται, δηλαδή έχουμε:

$$\hat{x} = 0.5x_i + 0.5x_j$$

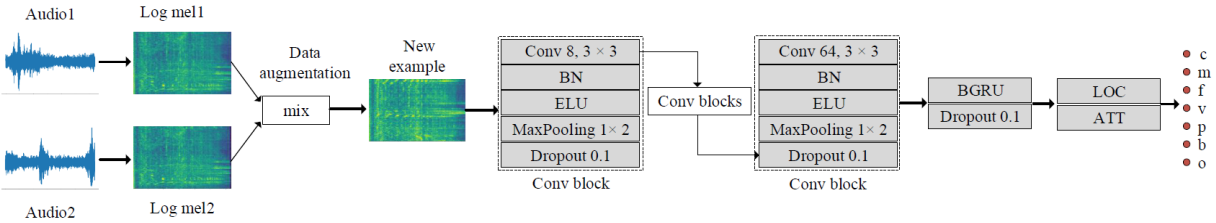
$$\hat{y} = y_i$$

Αντίστοιχα, η Extrapolation λειτουργεί σαν μια διαφορετική εναλλακτική του γραμμικού συνδυασμού δειγμάτων που εφαρμόζεται στην τυπική μέθοδο mixup. Η ετικέτα του νέου δείγματος ταυτίζεται με αυτή του πρώτου, όπως συμβαίνει και στην SamplePairing:

$$\hat{x} = (1 + \lambda)x_i - \lambda x_j$$

$$\hat{y} = y_i$$

Οι τεχνικές αυτές αξιολογήθηκαν στο Chime-home dataset [29], οι επισημειώσεις του οποίου αφορούν 7 διαφορετικές κλάσεις: ομιλία παιδιού, ομιλία ενήλικου άνδρα, ομιλία ενήλικης γυναίκας, ηλεκτρονικό παιχνίδι/τηλεόραση, κρουστικούς ήχους, ευρυζωνικό θόρυβο και άλλους αναγνωρίσιμους ήχους. Οι κατηγορίες αυτές συμβολίζονται ως c,m,f,v,p,b και o αντίστοιχα.

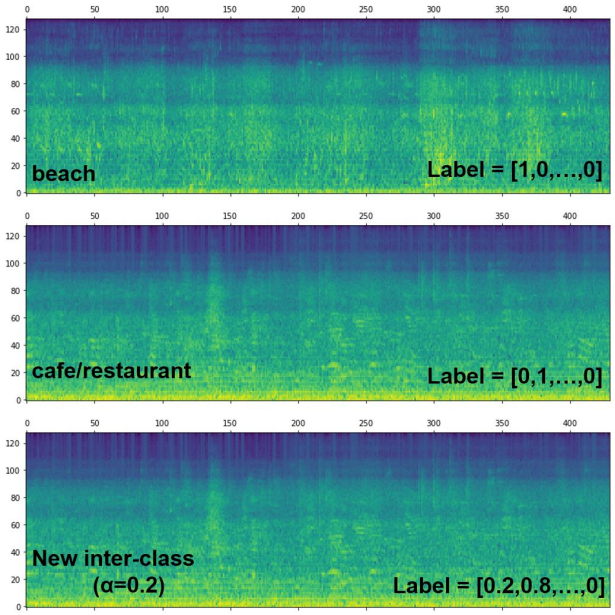


Σχήμα 2.10: Αρχιτεκτονική CNN-RNN για πρόβλεψη ετικετών σε ηχητικές καταγραφές [110]. Η επαύξηση δεδομένων πραγματοποιείται στο επίπεδο εισόδου.

Τα αποτελέσματα των πειραμάτων αποδεικνύουν πως η τεχνική mixup καθώς και οι δύο παραλλαγές της (SamplePairing και Extrapolation) συμβάλλουν στην βελτίωση της ακρίβειας κατηγοριοποίησης των ακουστικών σημάτων. Μάλιστα η μέθοδος mixup οδηγεί σε μεγαλύτερη ευστάθεια του εκπαιδευμένου μοντέλου καθώς και σε καλύτερη ικανότητα γενίκευσης.

Ένα άλλο πρόβλημα, το οποίο εντάσσεται στο γενικότερο πλαίσιο επισημείωσης ακουστικών σημάτων, είναι αυτό της κατηγοριοποίησης ακουστικής σκηνής (acoustic scene classification). Σύμφωνα με αυτό, στόχος είναι η αναγνώριση του περιβάλλοντος μέσα στο οποίο εξελίσσονται οι ήχοι μιας καταγραφής. Πρόκειται για μια ιδιαίτερα σημαντική εφαρμογή της μηχανικής ακρόασης καθώς επιτρέπει στα συστήματα να αντιλαμβάνονται αυτόματα το χώρο που τα περιβάλλει. Χάρη στην ιδιαίτερη ικανότητα τους να ανιχνεύουν αυτόματα περίπλοκα χαρακτηριστικά των δεδομένων εισόδου, τα Δίκτυα Βαθιάς Μάθησης φαίνεται πως έχουν συμβάλει δραματικά στην αποτελεσματική προσέγγιση και αυτού του προβλήματος. Ωστόσο, εξαιτίας του μεγάλου αριθμού παραμέτρων που διαθέτουν, είναι επιρρεπή στο φαινόμενο της υπερεκπαίδευσης.

Όπως, έχει αναφερθεί, ο πιο εύκολος και ευρέως διαδεδομένος τρόπος μείωσης του overfitting είναι η χρήση μεγαλύτερων datasets. Επειδή, ωστόσο, κάτι τέτοιο δεν είναι εύκολο, σαν εναλλακτική, μπορούν να χρησιμοποιηθούν μέθοδοι επαύξησης δεδομένων, με στόχο την τεχνητή επέκταση του training set. Στο [112], μια ομάδα ερευνητών εκπαίδευσε ένα πολυκάναλο CNN πάνω στο σύνολο δεδομένων DCASE 2017 audio scene classification [80], εφαρμόζοντας την τεχνική mixup. Πρόκειται για να ένα πρόβλημα κατηγοριοποίησης σε 15 διακριτικές κλάσεις (όπως παραλία, αυτοκίνητο, εμπορικό κέντρο κτλ). Τα αποτελέσματα έδειξαν πως η mixup συμβάλει στη βελτίωση της ακρίβεια κατηγοριοποίησης καθώς και στην μείωση του σφάλματος πάνω στα δεδομένα ελέγχου.



Σχήμα 2.11: Εφαρμογή της mixup πάνω σε φασματογραφήματα ακουστικών καταγραφών [112].

Κεφάλαιο 3

Βάσεις δεδομένων εκφράσεων προσώπου

3.1 Κατηγοριοποίηση βάσεων δεδομένων

Μέχρι πρόσφατα, οι περισσότερες πειραματικές μελέτες γύρω από την αναγνώριση ανθρώπινων συναισθημάτων πραγματοποιούνταν κάτω από ελεγχόμενες συνθήκες περιβάλλοντος [34, 72, 103, 113, 114], με περιορισμένο αριθμό συμμετεχόντων [10, 73, 88, 92, 104] και με προκαθορισμένα σενάρια τα οποία οι χρήστες έπρεπε να ακολουθούν πιστά [1, 77]. Ωστόσο, η ραγδαία εξάπλωση του διαδικτύου και των μέσων κοινωνικής δικτύωσης οδήγησε στη δημιουργία ενός τεράστιου όγκου δεδομένων και κατά συνέπεια ευνόησε τη δημιουργία datasets με μεγάλη ποικιλομορφία.

Συγκεκριμένα, τα τελευταία χρόνια έχουν κατασκευαστεί αρκετά σύνολα δεδομένων από εικόνες και βίντεο που απεικονίζουν ανθρώπους διαφορετικής ηλικίας, εθνικότητας αλλά και κοινωνικής θέσης σε ποικίλες συναισθηματικές καταστάσεις της καθημερινότητας. Το γεγονός αυτό έχει δώσει σημαντική ώθηση στον τομέα της αυτόματης αναγνώρισης συναισθημάτων αφού επιτρέπει την εκπαίδευση δικτύων με δεδομένα που αντικατοπτρίζουν συνθήκες του πραγματικού κόσμου.

Ανεξαρτήτως βέβαια του τρόπου δημιουργίας τους (σε ελεγχόμενο ή μη περιβάλλον), οι βάσεις, ανάλογα με το είδος δεδομένων που περιλαμβάνουν, διακρίνονται σε τρεις βασικές κατηγορίες:

- **Στατικές**: Οι στατικές βάσεις αποτελούν την πιο δημοφιλή κατηγορία βάσεων δεδομένων. Περιλαμβάνουν μόνο εικόνες, δηλαδή διαθέτουν μόνο στατική πληροφορία. Χαρακτηριστικά παραδείγματα αποτελούν βάσεις όπως η FER-2013 [33], η AffectNet [83] και η RAF-DB [70].

- **Δυναμικές**: Πρόκειται για βάσεις που περιλαμβάνουν βίντεο χωρίς ήχο, δηλαδή χρονικά μεταβαλλόμενες ακολουθίες εικόνων χωρίς ηχητική πληροφορία. Παραδείγματα τέτοιων βάσεων είναι οι DISFA [76], BP4D-Spontaneous [118] και AFEW [20].

- **Οπτικοακουστικές**: Πρόκειται για βάσεις που περιλαμβάνουν οπτικοακουστικά βίντεο, δηλαδή περιέχουν τόσο οπτική όσο και ακουστική πληροφορία για χρονικά μεταβαλλόμενες ακολουθίες εικόνων. Χαρακτηριστικά παραδείγματα αποτελούν οι βάσεις OMG-Emotion [7], RECOLA [92] και Aff-Wild2 [49, 55, 56, 58].

3.2 Βάσεις σε ελεγχόμενες συνθήκες

Στον τομέα της αναγνώρισης συναισθήματος, η πλειοψηφία των διαθέσιμων βάσεων έχει δημιουργηθεί μέσα σε ελεγχόμενο περιβάλλον εργαστηρίου, με τους συμμετέχοντες να ακολουθούν αυστηρά καθορισμένες οδηγίες για τον τρόπο εκτέλεσης συγκεκριμένων εργασιών. Στην παρούσα ενότητα θα αναλύσουμε τις κυριότερες βάσεις εικόνων και βίντεο που περιλαμβάνουν εκφράσεις προσώπου και μπορούν να χρησιμοποιηθούν για συναισθηματική ανάλυση.

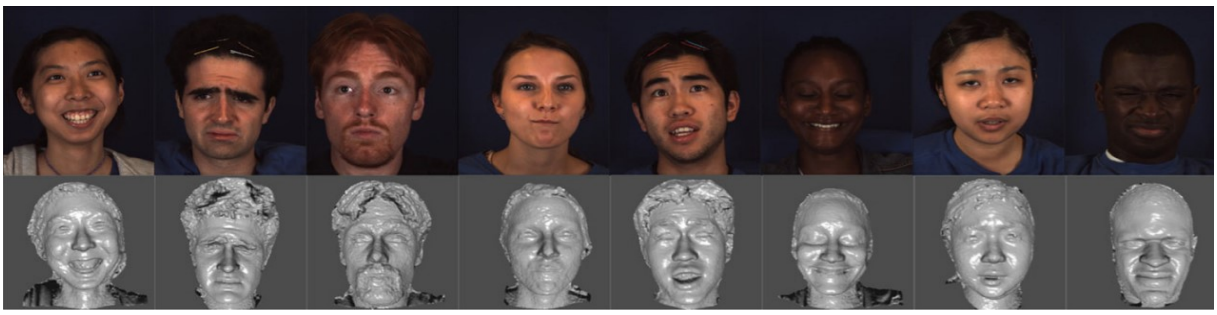
3.2.1 DISFA

Σχήμα 3.1: Δείγμα εικόνων από την βάση DISFA [76]

Η DISFA [76] αποτελεί μια δυναμική βάση δεδομένων που περιλαμβάνει αυθόρμητες ανθρώπινες εκφράσεις. Αναλυτικότερα, περιέχει στερεοφωνικά βίντεο από 27 ενήλικες (12 γυναίκες και 15 άνδρες) διαφορετικών εθνικοτήτων με συνολικά 261.630 χιλιάδες καρέ. Για κάθε εικόνα, έχει καταγραφεί η ένταση ενεργοποίησης 12 Action Units σε κλίμακα από 0 (απούσα) έως 5 (μέγιστη ένταση) από δύο εξειδικευμένους υπομνηματιστές. Η βάση επίσης περιλαμβάνει 66 σημεία αναφοράς προσώπου (landmarks) για κάθε εικόνα.

3.2.2 BP4D-Spontaneous

Η BP4D-Spontaneous [118] αποτελεί μια δυναμική βάση δεδομένων που περιλαμβάνει αυθόρμητες ανθρώπινες εκφράσεις από μια μεγάλη ομάδα ενηλίκων διαφορετικών εθνικοτήτων. Αναλυτικότερα, περιέχει 1.640 βίντεο από 61 άτομα με συνολικά περίπου 223 χιλιάδες καρέ. Διαθέτει επισημειώσεις για την ενεργοποίηση και ένταση 27 Action Units. Υπάρχουν 21 άνθρωποι σε συνολικά 75,6 χιλιάδες εικόνες στο training set, 20 άνθρωποι με 71,2 χιλιάδες εικόνες στο validation set και 20 άνθρωποι με 75,7 χιλιάδες εικόνες στο test set. Η βάση αυτή έχει χρησιμοποιηθεί ως μέρος του FERA 2015 Challenge [105].



Σχήμα 3.2: Δείγμα εικόνων από την βάση BP4D-Spontaneous [118]

3.2.3 BP4D+

Η BP4D+ [119] αποτελεί μια επέκταση της BP4D, η οποία ενσωματώνει μεγαλύτερο αριθμό συμμετεχόντων, 140 στο σύνολο (58 άνδρες και 82 γυναίκες). Περιέχει επισημειώσεις για την εμφάνιση 34 Action Units καθώς και για την ένταση 5 εξ΄ αυτών. Περιλαμβάνει 2.952 βίντεο από 41 άτομα σε 9 διαφορετικές όψεις στο σύνολο εκπαίδευσης, 1431 βίντεο από 20 άτομα σε 9 διαφορετικές όψεις στο σύνολο επικύρωσης και 1080 βίντεο από 30 άτομα στο σύνολο ελέγχου. Η βάση αυτή χρησιμοποιήθηκε ως μέρος του FERA 2017 Challenge [106].

3.2.4 Sayette Gft

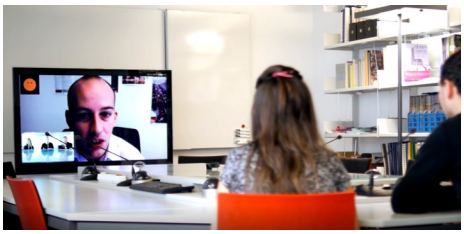
Η Sayette Group Formation Task (GFT) [30] αποτελεί μια οπτικοακουστική βάση με 96 βίντεο που αποτελούνται από 172.800 καρέ.

Σχήμα 3.3: Καρέ από βίντεο της βάσης Sayette Gft [26]

Αναλυτικότερα, περιλαμβάνει 1.800 καρέ για κάθε έναν από τους 96 συμμετέχοντες. Όλες οι εικόνες διαθέτουν επισημειώσεις για την εμφάνιση ή απουσία 10 AUs (1,2,4,6,10,12,14,15,23,24). Για τη δημιουργία της βάσης, συγκροτήθηκαν 32 ομάδες με 3 νεαρά άτομα η καθεμία. Τα μέλη κάθε ομάδας δεν γνωρίζονταν από πριν μεταξύ τους. Στο ένα τρίτο των ομάδων δόθηκε να πιουν ένα αλκοολούχο ποτό, στο ένα τρίτο δόθηκε ένα μη-αλκοολούχο ποτό το οποίο όμως πίστευαν ότι περιέχει αλκοόλ και στο άλλο ένα τρίτο δόθηκε χυμός φρού-των. Τα μέλη κάθε ομάδας κλήθηκαν να πιουν το ποτό/χυμό που τους δόθηκε και να αλληλεπιδράσουν μεταξύ τους με φυσικό τρόπο κάτω από μη-σκηνοθετημένες συνθήκες.

3.2.5 RECOLA

Η REmote COLlaborative and Affective (RECOLA) [92] αποτελεί μια οπτικοακουστική βάση δεδομένων με 46 βίντεο συνολικής διάρκειας άνω των 9,5 ωρών (σύνολο 345.000 καρέ).



Σχήμα 3.4: Καρέ από βίντεο της βάσης RECOLA [92]

Οι άνθρωποι που συμμετείχαν στη δημιουργία της βάσης καταγράφηκαν σε δυάδες κατά τη διάρκεια μιας τηλεδιάσκεψης καθώς εκτελούσαν κάποια εργασία που απαιτούσε συνεργασία. Η βάση περιλαμβάνει ηχητικά βίντεο καθώς και δεδομένα ηλεκτροκαρδιογραφήματος και ηλεκτροδερμικής δραστηριότητας. Συνολικά, 46 Γαλλόφωνοι άνθρωποι έλαβαν μέρος (27 γυναίκες και 19 άνδρες). Για να διευκολυνθεί η διαδικασία προσθήκης ετικετών, επισημειώθηκαν μόνο τα πρώτα 5 λεπτά από κάθε βίντεο, δηλαδή σύνολο 3 ώρες και 50 λεπτά, από 6 υπομνηματιστές (3 άνδρες και 3 γυναίκες). Οι επισημειώσεις αφορούν τις συνεχείς συναισθηματικές συνιστώσες Valence και Arousal, οι οποίες κυμαίνονται στο εύρος [-1,1]. Η βάση χωρίζεται σε τρία μέρη, στο σύνολο εκπαίδευσης (16 άνθρωποι), στο σύνολο επικύρωσης (15 άνθρωποι) και στο σύνολο ελέγχου (15 άνθρωποι), με τρόπο τέτοιο ώστε το φύλο, η ηλικία και η μητρική γλώσσα των ανθρώπων να είναι ισορροπημένα στα τρία σύνολα.

3.3 Βάσεις in-the-wild

Η ραγδαία εξάπλωση της τεχνολογίας, και συγκεκριμένα των Μέσων Μαζικής Ενημέρωσης και του διαδικτύου, ευνόησε τη δημιουργία βάσεων δεδομένων κάτω από μη-ελεγχόμενες συνθήκες, in-the-wild. Στη ενότητα αυτή θα αναλύσουμε τις βασικότερες συλλογές εικόνων και βίντεο που περιλαμβάνουν εκφράσεις προσώπου σε ποικίλες συνθήκες φωτισμού, εναλλακτικές πόζες καθώς και σε διαφορετικό υπόβαθρο.

3.3.1 IMFDB

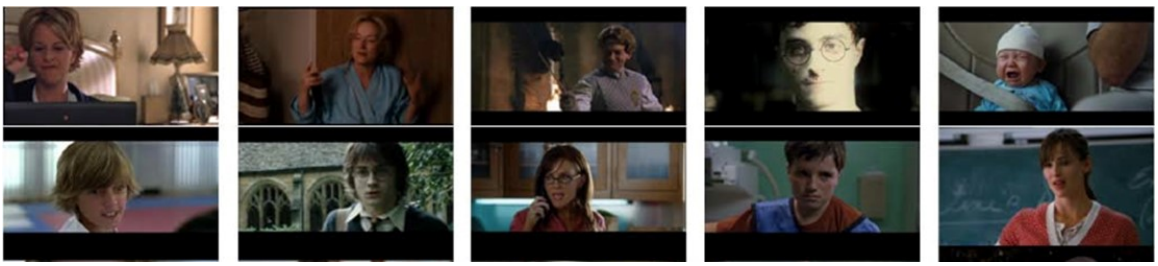
Η Indian Movie Face Database (IMFDB) [96] αποτελεί μια στατική βάση δεδομένων με 34.512 πρόσωπα 100 γνωστών ηθοποιών, συγκεντρωμένα από περίπου 103 Ινδικές ταινίες.

Σχήμα 3.5: Δείγμα εικόνων από την βάση IMFDB [96]

Συγκεκριμένα, περιλαμβάνει 67 άνδρες και 33 γυναίκες ηθοποιούς με μακροχρόνια καριέρα στο χώρο της υποκριτικής. Οι περισσότερες ταινίες συγκεντρώθηκαν από το Youtube με ορισμένες να εμφανίζουν χειρότερη ανάλυση από άλλες. Αξίζει να σημειωθεί πως η βάση περιλαμβάνει τουλάχιστον 200 εικόνες ανά ηθοποιό. Όλες οι εικόνες έχουν επισημειωθεί χειροκίνητα ως προς τα 6 βασικά συναισθήματα, το φωτισμό, την πόζα, την ηλικία, το φύλο, το makeup καθώς και ως προς χαρακτηριστικά που καθιστούν το πρόσωπο όχι εντελώς ορατό (π.χ. γυαλιά, μούσι κτλ). Όλες αυτές οι διαφορετικές ετικέτες που υπάρχουν ανά εικόνα μπορούν να αξιοποιηθούν σε διάφορα προβλήματα ανίχνευσης προσώπου και αναγνώρισης συναισθήματος.

3.3.2 AFEW

Η Acted Facial Expression In The Wild (AFEW) [20] αποτελεί μια δυναμική βάση δεδομένων με 1.809 βίντεο από σκηνές ταινιών και τηλεοπτικών ριάλιτι. Οι άνθρωποι που εμφανίζονται στα βίντεο καλύπτουν ένα ευρύ φάσμα ηλικιών (1-77 ετών) ενώ ξεπερνούν σε αριθμό τους 330. Οι ετικέτες της βάσης έχουν προστεθεί από 3 υπομνηματιστές και αφορούν τα 7 βασικά συναισθήματα (κατηγορικό μοντέλο). Το σύνολο εκπαίδευσης περιλαμβάνει 773 βίντεο, το σύνολο επικύρωσης 383 και το σύνολο ελέγχου 653. Αξίζει να σημειωθεί ότι τόσο το training όσο και το validation set αποτελούνται κυρίως από αποσπάσματα πραγματικών ταινιών, ωστόσο 114 από τα 653 βίντεο του test set συνιστούν πραγματικά τηλεοπτικά κλιπ, επομένως καθιστούν ιδιαίτερα απαιτητική την επίτευξη υψηλής ακρίβειας κατηγοριοποίησης.



Σχήμα 3.6: Δείγμα από βίντεο της βάσης AFEW [20]

3.3.3 FER-2013

Η FER-2013 [33] αποτελεί μια στατική βάση δεδομένων με συνολικά 35.887 εικόνες εκφράσεων προσώπου. Κάθε εικόνα βρίσκεται σε κλίμακα του γκρι και έχει διάσταση 48×48.

Σχήμα 3.7: Δείγμα εικόνων από την βάση FER-2013 [33]

Η βάση δημιουργήθηκε με χρήση της Google, αναζητώντας εικόνες προσώπου που ταιριάζουν με ένα σύνολο από 184 λέξεις-κλειδιά, όπως "blissful" (ευτυχισμένος), "enraged" (εξορκισμένος) κτλ. Οι λέξεις αυτές συνδυάστηκαν με άλλες σχετικές με το φύλο, την ηλικία και την εθνικότητα για την δημιουργία περίπου 600 συμβολοσειρών (strings), οι οποίες χρησιμοποιήθηκαν ως queries κατά την αναζήτηση. Η βάση περιλαμβάνει ετικέτες για τα 7 βασικά συναισθήματα (κατηγορικό μοντέλο). Το σύνολο εκπαίδευσης αποτελείται από 28.709 εικόνες ενώ τα σύνολα επικύρωσης και ελέγχου από 3.589 έκαστο.

3.3.4 EmotioNet

Η EmotioNet [8] αποτελεί μια στατική βάση δεδομένων με περίπου 1 εκατομμύριο εικόνες εκφράσεων προσώπου. Όλες οι εικόνες έχουν συγκεντρωθεί από το διαδίκτυο με χρήση κατάλληλων λέξεων-κλειδιά. 950 χιλιάδες από αυτές έχουν επισημειωθεί με αυτόματο τρόπο ενώ οι υπόλοιπες 50 χιλιάδες έχουν επισημειωθεί χειροκίνητα ως προς την εμφάνιση 11 AUs (1,2,4,5,6,9,12,17,20,25,26). Από τις τελευταίες, οι μισές συνιστούν το σύνολο επικύρωσης ενώ οι άλλες μισές το σύνολο ελέγχου. Επιπλέον, ένα υποσύνολο περίπου 2,5 χιλιάδων εικόνων έχει επισημειωθεί ως προς τα 6 βασικά συναισθήματα καθώς και ως προς 10 σύνθετες συναισθηματικές εκφράσεις.

Σχήμα 3.8: Δείγμα εικόνων από την βάση EmotioNet [8]

3.3.5 OMG-Emotion

Σχήμα 3.9: Καρέ από βίντεο από την βάση OMG-Emotion [7]

Η One-Minute-Gradual Emotion (OMG-Emotion) [7] αποτελεί μια οπτικοακουστική βάση δεδομένων με 567 βίντεο συνολικής διάρκειας 15 ωρών. Αναλυτικότερα, κάθε βίντεο προέρχεται από το Youtube, έχει μέση διάρκεια 1 λεπτό και χωρίζεται σε πολλαπλά διαδοχικά ηχητικά κλιπ (μέσης διάρκειας 8 δευτερολέπτων). Κάθε ένα από τα κλιπ έχει επισημειωθεί από 5 διαφορετικούς υπομνηματιστές τόσο ως προς τις συνεχείς συνιστώσες Valence-Arousal (διανυσματικό μοντέλο) όσο και ως προς τα 7 βασικά συναισθήματα (κατηγορικό μοντέλο). Ο συνολικός αριθμός ετικετών της βάσης είναι 39.803.

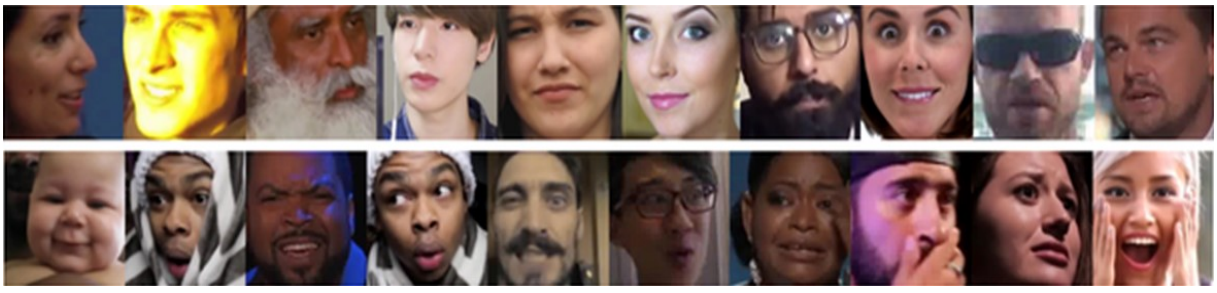
3.3.6 Aff-Wild

Η Aff-Wild [48, 115] αποτελεί μια οπτικοακουστική βάση δεδομένων με 298 βίντεο συνολικής διάρκειας μεγαλύτερης από 30 ώρες (σύνολο 1.224.100 καρέ). Περιλαμβάνει 200 ανθρώπους διαφορετικών εθνικοτήτων (130 άνδρες και 70 γυναίκες) οι οποίοι αντιδρούν αυθόρμητα σε ποικίλα ερεθίσματα (όπως π.χ. σε ανατροπή της πλοκής κάποιας ταινίας, σε δοκιμή μιας καυτερής ή δυσάρεστης γεύσης, σε κάποιο ανέκδοτο κτλ). Τα συναισθήματα που εκδηλώνονται είναι τόσο θετικά όσο και αρνητικά. Η συγκέντρωση των βίντεο πραγματοποιήθηκε από το Youtube χρησιμοποιώντας σαν βασικό κλειδί αναζήτησης τη λέξη 'reaction' (αντίδραση). Κάθε καρέ της βάσης είναι επισημειωμένο από 8 ειδικούς υπομνηματιστές ως προς τις συνεχείς συνιστώσες Valence και Arousal, οι οποίες λαμβάνουν τιμές στο εύρος [-1,1]. Το σύνολο εκπαίδευσης αποτελείται από 252 βίντεο (με περίπου 1 εκατομμύριο καρέ) ενώ το σύνολο ελέγχου από 46 (με περίπου 216 χιλιάδες καρέ).

Σχήμα 3.10: Καρέ από βίντεο της βάσης Aff-Wild [48, 115]

3.3.7 Aff-Wild2

Η Aff-Wild2 [49, 55, 56, 58, 60] συνιστά μια επέκταση της οπτικοακουστικής βάσης Aff-Wild. Συγκεκριμένα, αποτελεί την ένωση της Aff-Wild και άλλων 260 βίντεο, συνολικής διάρκειας 13 ωρών και 5 λεπτών (1.413.000 επιπλέον καρέ). Συνολικά, η βάση Aff-Wild2 αποτελείται από 558 βίντεο με 2.786.201 καρέ. Περιλαμβάνει 458 ανθρώπους (279 άνδρες και 179 γυναίκες) διαφορετικής ηλικίας, εθνικότητας και επαγγέλματος, οι οποίοι εκδηλώνουν ήπια αλλά και έντονα συναισθήματα σε πραγματικές συνθήκες. Όλες οι εικόνες της βάσης διαθέτουν επισημειώσεις ως προς τις συνιστώσες Valence και Arousal. Επίσης, 63 βίντεο (με 398.835 καρέ) έχουν επισημειωθεί από 3 πολύ έμπειρους υπομνηματιστές ως προς τα AUs 1,2,4,6,12,15,20,25. Τέλος, 539 βίντεο (με 2.595.572 καρέ) έχουν επισημειωθεί από 7 ειδικούς ως προς τα 7 βασικά συναισθήματα. Η Aff-Wild2 αποτελεί την πρώτη και μοναδική βάση δεδομένων που περιλαμβάνει ετικέτες και για τα 3 βασικά μοντέλα συναισθηματικής αναπαράστασης (Κατηγορικό μοντέλο 7 βασικών συναισθημάτων, ανίχνευση Action Units, διανυσματικό μοντέλο Valence-Arousal). Η βάση χρησιμοποιήθηκε στους διαγωνισμούς ABAW [49] και ABAW2 [61].



Σχήμα 3.11: Καρέ από βίντεο της βάσης Aff-Wild2 [49, 55, 56, 58]

3.4 Βάσεις που αξιοποιήθηκαν στα πειράματα

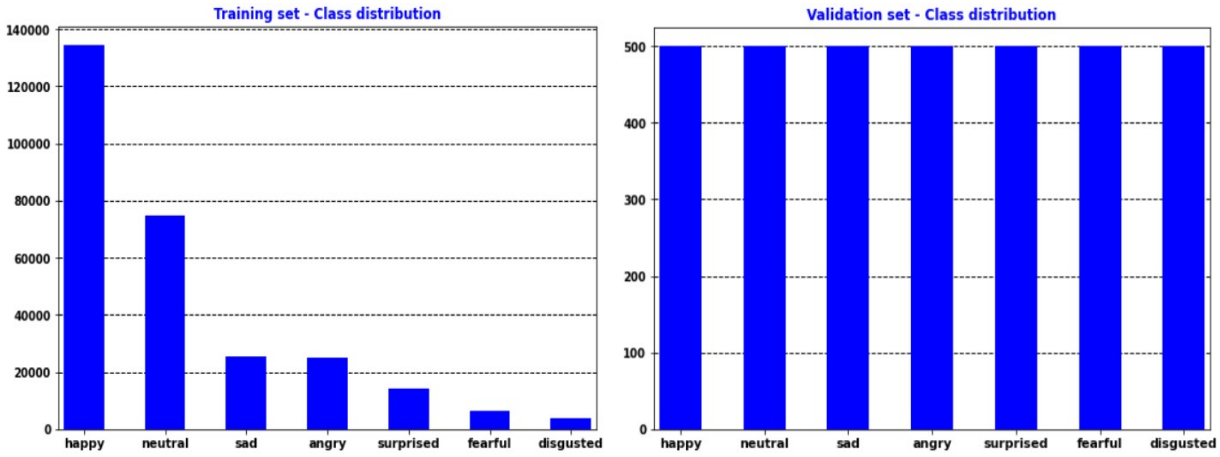
Στα πλαίσια της παρούσας διπλωματικής αξιοποιούμε δύο στατικές in-the-wild βάσεις δεδομένων, οι οποίες διαθέτουν επισημειώσεις ως προς τα 7 βασικά συναισθήματα, τις AffectNet και RAF-DB. Συγκεκριμένα, κάνουμε χρήση δύο διαφορετικών εκδόσεων της κάθε βάσης για μεγαλύτερη συνέπεια στα αποτελέσματά μας.

3.4.1 AffectNet

Η AffectNet [83] αποτελεί μια στατική βάση δεδομένων η οποία περιλαμβάνει πάνω από 1 εκατομμύριο εικόνες εκφράσεων προσώπου κάτω από μη-ελεγχόμενες συνθήκες, in-the-wild.

Σχήμα 3.12: Δείγμα εικόνων από την βάση AffectNet [76]

Οι εικόνες αυτές συγκεντρώθηκαν από το ίντερνετ αξιοποιώντας τρεις βασικές μηχανές αναζήτησης και 1250 λέξεις-κλειδιά σχετικές με το συναίσθημα σε 6 διαφορετικές γλώσσες. Οι μισές περίπου από τις εικόνες (∼440 χιλιάδες) περιέχουν ετικέτες ως προς ένα από τα 7 βασικά συναισθήματα (συν το συναίσθημα της περιφρόνησης, το οποίο στα πλαίσια της δικής μας μελέτης δεν λήφθηκε υπόψη) καθώς και ως προς τις συνιστώσες Valence-Arousal (διανυσματικό μοντέλο). Το σύνολο εκπαίδευσης της βάσης αποτελείται από περίπου 321 χιλ. εικόνες ενώ το σύνολο επικύρωσης από 5 χιλ. Στα πλαίσια της παρούσας διπλωματικής αξιοποιούμε 2 μικρότερα versions, καθένα από τα οποία περιέχει περίπου 290 χιλ. εικόνες. Η κατανομή αυτών σε κλάσεις φαίνεται παρακάτω.



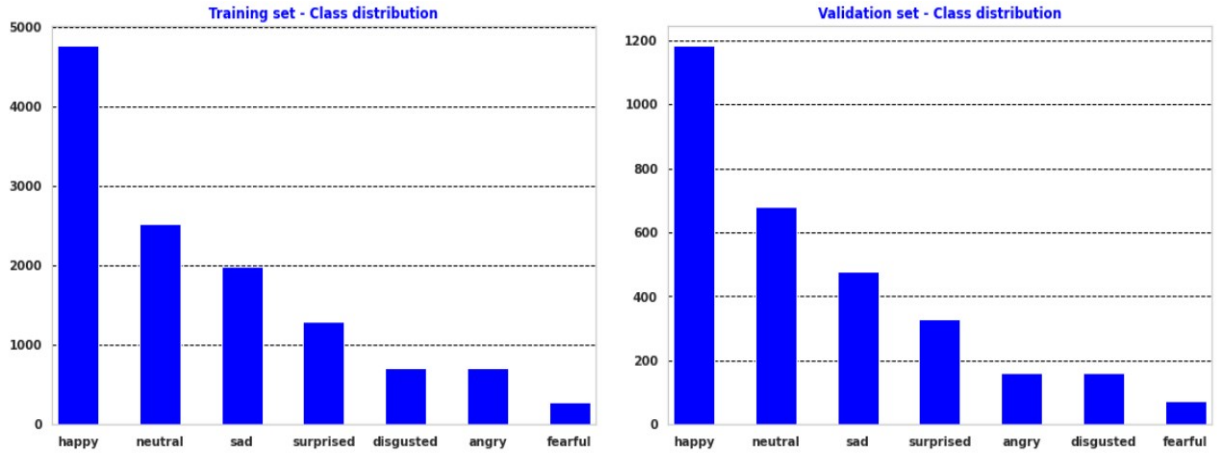
Σχήμα 3.13: Κατανομή δειγμάτων σε κλάσεις για το σύνολο εκπαίδευσης (training set) και για το σύνολο επικύρωσης (validation set) της βάσης AffectNet που χρησιμοποιούμε.

Όπως βλέπουμε στο Σχήμα 3.13, το σύνολο εκπαίδευσης είναι μη-ισορροπημένο, με την κλάση "happy" να έχει με διαφορά τα περισσότερα δείγματα. Δεύτερη βρίσκεται η "neutral" με επίσης μεγάλο αριθμό εικόνων ενώ ακολουθούν όλες οι υπόλοιπες. Το σύνολο επικύρωσης είναι πλήρως ισορροπημένο, με όλες τις κλάσεις να έχουν τον ίδιο αριθμό δειγμάτων (500).

3.4.2 RAF-DB

Η Real-world Affective Faces Database (RAF-DB) [70] αποτελεί μια στατική βάση δεδομένων με περίπου 30 χιλιάδες εικόνες εκφράσεων προσώπου από το διαδίκτυο. Η διαδικασία δημιουργίας ετικετών της βάσης πραγματοποιήθηκε με τη μέθοδο του πληθοπορισμού (crowdsourcing), με την κάθε εικόνα να επισημειώνεται ανεξάρτητα από περίπου 40 υπομνηματιστές. Το σύνολο των εικόνων της βάσης χωρίζεται σε 2 υποσύνολα. Στο πρώτο, υπάρχουν 15.339 εικόνες που έχουν επισημειωθεί ως προς ένα από τα 7 βασικά συναισθήματα (κατηγορικό μοντέλο) ενώ στο δεύτερο, οι εικόνες έχουν χαρακτηριστεί ως προς 12 σύνθετες συναισθηματικές καταστάσεις. Εμείς, στα πλαίσια της παρούσας διπλωματικής, θα ασχοληθούμε μόνο με το πρώτο υποσύνολο.

Σχήμα 3.14: Παραδείγματα εικόνων της RAF-DB από τις 6 βασικές κατηγορίες και από άλλες 12 σύνθετες συναισθηματικές καταστάσεις [118].



Σχήμα 3.15: Κατανομή δειγμάτων σε κλάσεις για το σύνολο εκπαίδευσης (training set) και για το σύνολο επικύρωσης (validation set) της βάσης RAF-DB.

Στο Σχήμα 3.15 απεικονίζεται η κατανομή των εικόνων σε κλάσεις. Όπως διαπιστώνεται, τόσο το σύνολο εκπαίδευσης όσο και το σύνολο επικύρωσης είναι μη-ισορροπημένα, με την κλάση "happy" να έχει με διαφορά τα περισσότερα δείγματα. Δεύτερη βρίσκεται η "neutral" ενώ ακολουθεί η "sad". Και στα δύο σύνολα η κατηγορία με τα λιγότερα δείγματα είναι η "fearful".

Κεφάλαιο 4

Μεθοδολογία

4.1 Ελαχιστοποίηση Εμπειρικού Ρίσκου και mixup

Στην ενότητα αυτή αναλύουμε την τεχνική επαύξησης δεδομένων mixup, την οποία και εφαρμόζουμε πάνω στις βάσεις δεδομένων AffectNet και RAF-DB για το πρόβλημα κατηγοριοποίησης στα 7 βασικά συναισθήματα. Πρόκειται για μια απλή αλλά πολύ αποτελεσματική μέθοδο, η οποία μπορεί να οδηγήσει στην εκπαίδευση Δικτύων Βαθιάς Μάθησης με μεγαλύτερη ευστάθεια και καλύτερη ικανότητα γενίκευσης.

Ας υποθέσουμε πως εξετάζουμε ένα πρόβλημα κατηγοριοποίησης με επιβλεπόμενη μάθηση. Στην περίπτωση αυτή, υπάρχει ο χώρος των διανυσμάτων χαρακτηριστικών X και ο χώρος των ετικετών Y. Στόχος είναι η εκπαίδευση ενός Νευρωνικού Δικτύου, ή ισοδύναμα η εκμάθηση μιας συνάρτησης $f : X \to Y$, που θα δέχεται ως είσοδο ένα δείγμα $x \in X$ και θα δίνει ως έξοδο ένα δείγμα $y \in Y$. Για το σκοπό αυτό έχουμε στη διάθεση μας ένα σύνολο δεδομένων από n παραδείγματα, $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\} = \{(x_i, y_i)\}_{i=1}^n$, όπου $x_i \in X$ είναι μια είσοδος και $y_i \in Y$ είναι η έξοδος που επιθυμούμε να λάβουμε από την $f(x_i)$. Με άλλα λόγια, υποθέτουμε πως υπάρχει μια από κοινού κατανομή πιθανότητας $P(x, y)$ πάνω στα X και Y και το σύνολο εκπαίδευσης αποτελείται από n παραδείγματα, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, με $(x_i, y_i) \sim P(x, y)$. Ορίζουμε τώρα και μια συνάρτηση απώλειας (loss function), έστω l, η οποία επιβάλλει ποινή στις διαφορές που υπάρχουν ανάμεσα στις προβλέψεις $f(x)$ και στις πραγματικές ετικέτες y, για δείγματα $(x, y) \sim P(x, y)$. Το ρίσκο που σχετίζεται με τη συνάρτηση $f(x)$ ορίζεται ως η αναμενόμενη τιμή της συνάρτησης απώλειας, δηλαδή ισχύει:

$$R(f) = \mathbb{E}\left[l(f(x), y)\right] = \int l(f(x), y) dP(x, y) \tag{1}$$

Στις περισσότερες πρακτικές περιπτώσεις, το ρίσκο που παρουσιάζεται στην Εξίσωση 1 δεν μπορεί να υπολογιστεί καθώς η κατανομή $P(x, y)$ είναι άγνωστη. Ωστόσο, αξιοποιώντας το σύνολο εκπαίδευσης $\mathcal{D}$, είναι δυνατόν να υπολογιστεί μια εκτίμησή της P, η οποία ονομάζεται εμπειρική κατανομή:

$$P_\delta(x, y) = \frac{1}{n} \sum_{i=1}^{n} \delta(x = x_i, y = y_i) \tag{2}$$

όπου $\delta(x = x_i, y = y_i)$ είναι μια συνάρτηση Dirac με κέντρο το (x_i, y_i).

Αξιοποιώντας τώρα την εμπειρική κατανομή P_δ, από την Εξίσωση 2, μπορεί να υπολογιστεί μια εκτίμηση και για το αναμενόμενο ρίσκο, την οποία ονομάζουμε εμπειρικό ρίσκο:

$$R_\delta(f) = \int l(f(x), y) dP_\delta(x, y) = \frac{1}{n} \sum_{i=1}^{n} l(f(x_i), y_i) \tag{3}$$

Η διαδικασία εκμάθησης της συνάρτησης f (ή ισοδύναμα η διαδικασία εκπαίδευσης του Νευρωνικού Δικτύου) με ελαχιστοποίηση της Εξίσωσης 3 είναι γνωστή ως Ελαχιστοποίηση Εμπειρικού Ρίσκου (Empirical Risk Minimization) [107]. Αν και αποτελεσματικό στον υπολογισμό του, το εμπειρικό ρίσκο παρακολουθεί την συμπεριφορά της f μόνο σε έναν πεπερασμένο αριθμό n δειγμάτων. Αν αναλογιστούμε συναρτήσεις f με αριθμό παραμέτρων συγκρίσιμο με n, όπως Δίκτυα Βαθιάς Μάθησης, τότε ένας τετριμμένος τρόπος ελαχιστοποίησης του εμπειρικού ρίσκου είναι η απομνημόνευση όλων των δειγμάτων εκπαίδευσης [116]. Ωστόσο, η απομνημόνευση αυτή, με τη σειρά της, οδηγεί σε ανεπιθύμητη συμπεριφορά της f σε δείγματα εκτός του training set. Πιο συγκεκριμένα, Νευρωνικά Δίκτυα που έχουν εκπαιδευτεί με τη μέθοδο Ελαχιστοποίησης Εμπειρικού Ρίσκου, αλλάζουν δραματικά τις προβλέψεις τους όταν εξετάζουν δείγματα που βρίσκονται λίγο έξω από την κατανομή των δεδομένων εκπαίδευσης [101]. Τέτοια δείγματα είναι γνωστά με τον όρο ανταγωνιστικά παραδείγματα (adversarial examples).

Για την αντιμετώπιση του προβλήματος και την βελτίωση της γενίκευσης του δικτύου πάνω σε νέα δεδομένα, εφαρμόζεται συχνά η μέθοδος της επαύξησης δεδομένων [97], η οποία επισημοποιήθηκε με την αρχή της Ελαχιστοποίησης Γειτονικού Ρίσκου (Vicinal Risk Minimization) [14]. Σύμφωνα με την τελευταία, για τον ορισμό της έννοιας της γειτονιάς, γύρω από κάθε δείγμα του συνόλου εκπαίδευσης, απαιτείται ανθρώπινη γνώση. Με βάση τη γνώση αυτή, μπορεί να κατασκευαστούν εικονικά παραδείγματα στην περιοχή γύρω από τα ήδη υπάρχοντα, για την επέκταση της κατανομής των δεδομένων εισόδου. Για παράδειγμα, όπως αναφέραμε και στο εισαγωγικό κεφάλαιο, σε βάσεις που περιέχουν εικόνες, είναι σύνηθες να ορίζεται ως γειτονιά μιας εικόνας, το σύνολο των οριζόντιων αντανακλάσεων, των μικρών περιστροφών καθώς και των ήπιων κλιμακώσεών της.

Έτσι λοιπόν, η αφελής εκτίμηση P_δ δεν αποτελεί την μοναδική επιλογή για την εκτίμηση της πραγματικής κατανομής P. Στο παράδειγμα της αρχής Ελαχιστοποίησης Γειτονικού Ρίσκου, η εκτίμηση της P υπολογίζεται ως εξής:

$$P_\nu(\tilde{x}, \tilde{y}) = \frac{1}{n} \sum_{i=1}^{n} \nu(\tilde{x}, \tilde{y} | x_i, y_i) \tag{4}$$

όπου ν η γειτονική κατανομή που μετράει την πιθανότητας εύρεσης του εικονικού ζεύγους $(\tilde{x}, \tilde{y})$ στην γειτονιά του ζεύγους (x_i, y_i) του συνόλου εκπαίδευσης. Συγκεκριμένα, στο [14] χρησιμοποιούνται Γκαουσιανές κατανομές $\nu(\tilde{x}, \tilde{y} | x_i, y_i) = \mathcal{N}(\tilde{x} - x_i, \sigma^2)\delta(\tilde{y} = y_i)$, το οποίο ισοδυναμεί με επαύξηση των δεδομένων εκπαίδευσης με προσθετικό Γκαουσιανό θόρυβο. Για την εκπαίδευση με Ελαχιστοποίηση Γειτονικού Ρίσκου, δειγματοληπτούμε τη γειτονική κατανομή για την κατασκευή ενός συνόλου δεδομένων $\mathcal{D}_\nu := \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{m}$, και ελαχιστοποιούμε το εμπειρικό γειτονικό ρίσκο:

$$R_\nu(f) = \frac{1}{m} \sum_{i=1}^{m} l(f(\tilde{x}_i), \tilde{y}_i) \tag{5}$$

Παρόλο που η κλασική μέθοδος επαύξησης δεδομένων οδηγεί συνήθως σε καλύτερη ικανότητα γενίκευσης του δικτύου, η διαδικασία είναι εξαρτώμενη από το εκάστοτε σύνολο δεδομένων και συνεπώς απαιτεί την αξιοποίηση εξειδικευμένης γνώσης που σχετίζεται με το πρόβλημα. Επιπλέον, πραγματοποιεί την υπόθεση πως τα παραγόμενα γειτονικά συνθετικά παραδείγματα μοιράζονται όλα την ίδια κλάση ενώ δεν μοντελοποιεί και τη γειτονική σχέση δειγμάτων που ανήκουν σε διαφορετικές κλάσεις.

Παρακινούμενοι από τα ζητήματα αυτά, οι συγγραφείς του [117] προτείνουν μια γενική γειτονική κατανομή που ονομάζεται mixup:

$$\mu(\tilde{x}, \tilde{y}|x_i, y_i) = \frac{1}{n}\sum_{j}^{n} \mathop{\mathbb{E}}_{\lambda}[\delta(\tilde{x} = \lambda \cdot x_i + (1-\lambda)\cdot x_j, \tilde{y} = \lambda \cdot y_i + (1-\lambda)\cdot y_j)] \tag{6}$$

όπου $\lambda \sim \mathcal{B}(\alpha, \alpha)$, για $a \in (0, +\infty)$.

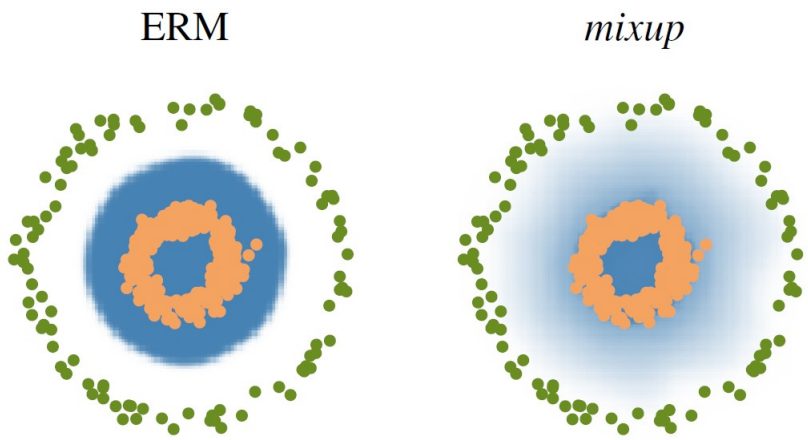
Με λίγα λόγια, η δειγματοληψία από την mixup γειτονική κατανομή παράγει εικονικά ζεύγη δεδομένων και αντίστοιχων ετικετών $(\tilde{x}, \tilde{y})$ ως εξής:

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda)x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda)y_j \end{aligned} \tag{7}$$

όπου x_i,x_j είναι τυχαία επιλεγμένα δείγματα των δεδομένων εκπαίδευσης και y_i,y_j οι αντίστοιχες ετικέτες τους σε one-hot κωδικοποίηση. Το $\lambda \in [0, 1]$ εκφράζει την αναλογία ανάμειξης των δειγμάτων και δειγματοληπτείται από μια Κατανομή Βήτα. Η υπερπαράμετρος a, όπως θα δούμε καλύτερα και στη συνέχεια, καθορίζει την ένταση της παρεμβολής μεταξύ των δεδομένων. Να σημειωθεί ότι για $a \to 0$ η τεχνική προσεγγίζει την αρχή Ελαχιστοποίησης Εμπειρικού Ρίσκου.

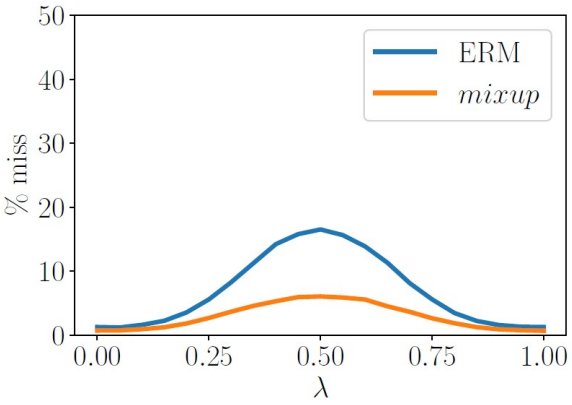
Παρατηρούμε λοιπόν πως η τεχνική mixup επεκτείνει την κατανομή των δεδομένων εκπαίδευσης, ενσωματώνοντας την προγενέστερη γνώση ότι οι γραμμικές παρεμβολές διανυσμάτων χαρακτηριστικών οδηγούν σε ισόποσες γραμμικές παρεμβολές των αντίστοιχων ετικετών. Πρόκειται για μια απλή τεχνική επαύξησης δεδομένων, η οποία είναι ανεξάρτητη του dataset πάνω στο οποίο εφαρμόζεται. Μπορεί να υλοποιηθεί εύκολα, σε λίγες μόλις γραμμές κώδικα, εισάγοντας ελάχιστη υπολογιστική επιβάρυνση (overhead).

Στο Σχήμα 4.1 μπορούμε να διακρίνουμε ένα ενδεικτικό παράδειγμα εφαρμογής της μεθόδου mixup σε ένα πρόβλημα δυαδικής κατηγοριοποίησης με σημεία στο διασδιάστατο χώρο. Τα πράσινα σημεία αφορούν τη μία κλάση (έστω κλάση 0) ενώ τα πορτοκαλί τη δεύτερη (έστω κλάση 1). Η σκιαγραφημένη με μπλε χρώμα περιοχή υποδεικνύει την πιθανότητα ένα δείγμα να ανήκει στην κλάση 1, των πορτοκαλί σημείων, δηλαδή $P(y = 1|x)$. Σε αντίθεση με την αριστερά εικόνα, όπου εφαρμόζεται η κλασική Ελαχιστοποίηση Εμπειρικού Ρίσκου (ERM), στη δεξιά, όπου εφαρμόζεται η τεχνική mixup, βλέπουμε πως τα όρια απόφασης μεταβαίνουν γραμμικά από τη μία κλάση στην άλλη, παρέχοντας μια πιο ομαλή εκτίμηση της αβεβαιότητας.

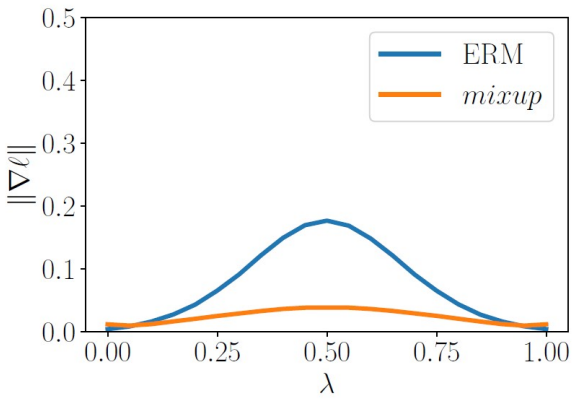


Σχήμα 4.1: Επίδραση της τεχνικής mixup ($a = 1$) σε ένα απλό πρόβλημα δυαδικής κατηγοριοποίησης με σημεία στον δισδιάστατο χώρο [117].

Η συνεισφορά της mixup μπορεί να γίνει καλύτερα αντιληπτή από το Σχήμα 4.2, όπου απεικονίζονται οι μέσες συμπεριφορές δύο Νευρωνικών Δικτύων που εκπαιδεύτηκαν πάνω στο σύνολο δεδομένων CIFAR-10 [63] με χρήση ERM και με χρήση mixup. Και τα δύο μοντέλα έχουν την ίδια αρχιτεκτονική, εκπαιδεύονται με την ίδια διαδικασία και αξιολογούνται σε ένα τυχαία δειγματοληπτημένο σύνολο από σημεία που βρίσκονται ενδιάμεσα των δεδομένων εκπαίδευσης. Στο Σχήμα 4.2α′, όπου απεικονίζεται το σφάλμα πρόβλεψης, βλέπουμε πως το μοντέλο που έχει εκπαιδευτεί με mixup εμφανίζει λιγότερες άστοχες προβλέψεις. Στο σημείο αυτό να σημειωθεί ότι η πρόβλεψη για ένα δείγμα $x = \lambda x_i + (1-\lambda)x_j$ θεωρείται άστοχη στην περίπτωση που αυτή δεν ανήκει στο $\{y_i, y_j\}$. Αντίστοιχα, στο Σχήμα 4.2β′, όπου απεικονίζεται το μέτρο του gradient της συνάρτησης απώλειας, παρατηρούμε πως το εκπαιδευμένο με mixup δίκτυο λαμβάνει σημαντικά μικρότερες τιμές. Συμπερασματικά, η εκπαίδευση του δικτύου με mixup οδηγεί σε καλύτερη προβλεπτική ικανότητα στο χώρο μεταξύ των δειγμάτων εκπαίδευσης, σε σχέση με την παραδοσιακή μέθοδο Ελαχιστοποίησης Εμπειρικού Ρίσκου.



(α′) Σφάλμα πρόβλεψης σε σημεία που βρίσκονται ενδιάμεσα από τα δεδομένα εκπαίδευσης. Η πρόβλεψη για το δείγμα $x = \lambda x_i + (1-\lambda)x_j$ θεωρείται αποτυχημένη ("miss"), αν δεν ανήκει στο $\{y_i, y_j\}$.



(β′) Μέτρο του gradient της συνάρτησης απώλειας για σημεία που βρίσκονται ενδιάμεσα από τα δεδομένα εκπαίδευσης, $x = \lambda x_i + (1-\lambda)x_j$.

Σχήμα 4.2: Επίδραση mixup σε σημεία που βρίσκονται ενδιάμεσα από τα δεδομένα εκπαίδευσης.

4.1.1 Κατανομή Βήτα

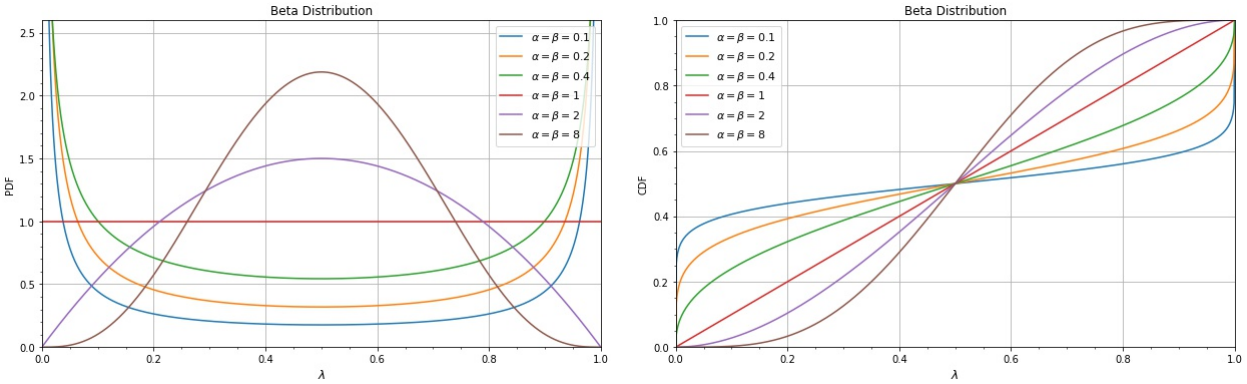
Όπως είδαμε στην Εξίσωση 7 της προηγούμενης ενότητας, η μεταβλητή λ, η οποία καθορίζει την αναλογία ανάμειξης των δειγμάτων εκπαίδευσης, λαμβάνει τιμές από μια κατανομή Βήτα. Στη ενότητα αυτή αναλύουμε ορισμένα σημαντικά χαρακτηριστικά της κατανομής αυτής για την καλύτερη κατανόηση της τεχνικής mixup.

Η κατανομή Βήτα ανήκει στην οικογένεια των συνεχών κατανομών πιθανότητας, έχει δύο θετικές παραμέτρους α,β και έχει συνάρτηση πυκνότητας πιθανότητας η οποία ορίζεται ως εξής:

$$\mathcal{B}(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

όπου με Γ συμβολίζεται η συνάρτηση Γάμμα.

Η συνάρτησης πυκνότητας πιθανότητας (PDF) καθώς και η αθροιστική συνάρτηση κατανομής (CDF) της Βήτα κατανομής μπορεί να πάρουν μία ποικιλία μορφών ανάλογα με την επιλογή τιμών των παραμέτρων της. Εμείς, στα πλαίσια εφαρμογής του mixup, θα εξετάσουμε μόνο περιπτώσεις όπου οι δύο αυτές παράμετροι ταυτίζονται ($\alpha = \beta$), όπως φαίνεται στο Σχήμα 4.3.



(α′) Συνάρτηση πυκνότητας πιθανότητας (PDF) της κατανομής Βήτα

(β′) Αθροιστική συνάρτηση κατανομής (CDF) της κατανομής Βήτα

Σχήμα 4.3: PDF και CDF κατανομής Βήτα για παραμέτρους $\alpha = \beta = \{0.1, 0.2, 0.4, 1, 2, 8\}$

Παρατηρούμε πως όσο μικρότερη είναι η κοινή τιμή των παραμέτρων α, β, τόσο πιθανότερο είναι, κατά την δειγματοληψία, να επιλεγούν τιμές λ κοντά στα άκρα του διαστήματος, δηλαδή τιμές κοντά στο 0 ή το 1. Ενδεικτικά να αναφέρουμε πως, με βάση το Σχήμα 4.3β′, για $\alpha = \beta = 0.1$ είναι $P(\lambda \leq 0.1) > 40\%$ ενώ για $\alpha = \beta = 2$ είναι $P(\lambda \leq 0.1) < 5\%$. Για $\alpha = \beta = 1$ η κατανομή Βήτα ταυτίζεται με Ομοιόμορφη κατανομή στο διάστημα [0,1]. Τέλος, φαίνεται πως για μεγάλες τιμές παραμέτρων, όπως $\alpha = \beta = 8$, η συνάρτηση πυκνότητας πιθανότητας της κατανομής προσεγγίζει σαν μορφή την αντίστοιχη συνάρτηση της Γκαουσιανής κατανομής, με κύρια διαφορά ωστόσο το γεγονός ότι η τελευταία εκτείνεται σε όλο τον άξονα των πραγματικών αριθμών ενώ η κατανομή Βήτα περιορίζεται στο εύρος [0,1].

Με βάση τα παραπάνω, γίνεται κατανοητό ότι στη mixup, όπου $\lambda \sim \mathcal{B}(\alpha, \alpha)$, για $a \in (0, +\infty)$, η τιμή της υπερπαραμέτρου α , καθορίζει σε μεγάλο βαθμό την ισχύ της γραμμικής παρεμβολής των δειγμάτων. Όπως θα δούμε στο Κεφάλαιο 5, όπου πραγματοποιείται η πειραματική μελέτη, στα datasets που εξετάζουμε είναι προτιμότερο να επιλέγεται μικρή τιμή

α, όπως π.χ. 0.1, 0.2 ή 0.4. Εδώ αξίζει να αναφέρουμε πως στην οριακή περίπτωση που το α λάβει μηδενική τιμή ($\alpha = 0$), η μέθοδος mixup ταυτίζεται με την Ελαχιστοποίηση Εμπειρικού Ρίσκου.

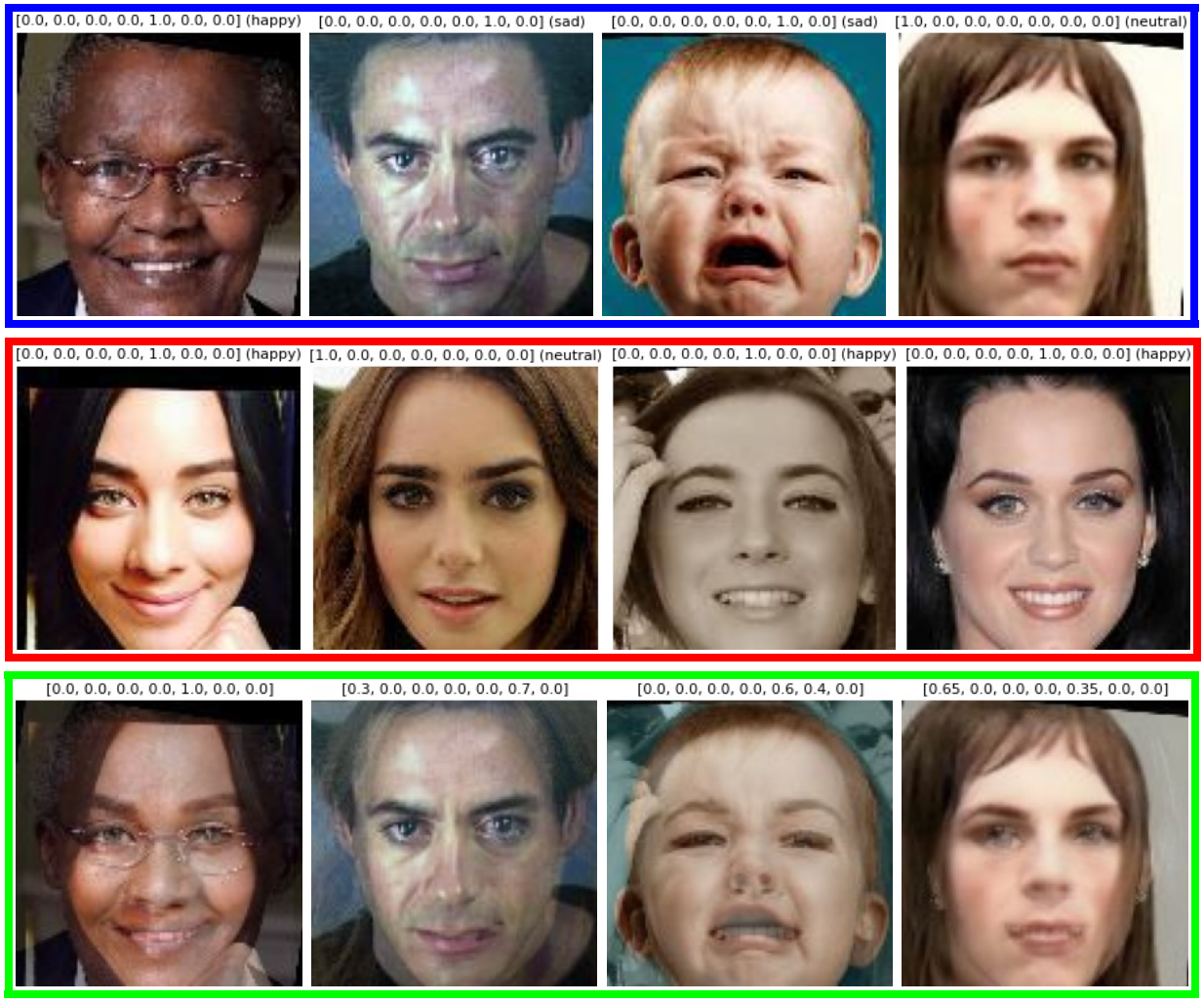
4.2 Mixup στην αναγνώριση συναισθήματος

Η mixup, αν και αποτελεί μια αρκετά απλή και αποτελεσματική μέθοδο επαύξησης δεδομένων, δεν έχει αξιοποιηθεί σε πολλές εφαρμογές της Όρασης Υπολογιστών. Ειδικά στον τομέα της αναγνώρισης συναισθήματος, μόνο στο [91], οι συγγραφείς φαίνεται πως χρησιμοποιούν την εν λόγω τεχνική για βελτίωση της γενίκευσης των μοντέλων που εκπαιδεύουν. Είναι λοιπόν ιδιαίτερα ενδιαφέρον να εξετάσουμε κατά πόσον μπορεί να συνεισφέρει στο απαιτητικό πρόβλημα κατηγοριοποίησης ανθρώπινων συναισθημάτων σε πραγματικές, μη-ελεγχόμενες συνθήκες. Στο Κεφάλαιο 5 πραγματοποιείται μια διεξοδική μελέτη της επίδρασης του mixup γύρω από το συγκεκριμένο πρόβλημα ενώ εφαρμόζεται και μια παραλλαγή της τεχνικής αυτής, η AddMixup, την οποία και αναλύουμε στη συνέχεια.

4.2.1 AddMixup: Μια αποτελεσματική παραλλαγή της mixup

Στην παραδοσιακή μέθοδο mixup, τυχαία επιλεγμένα ζεύγη δεδομένων παρεμβάλλονται γραμμικά και τροφοδοτούνται στο Νευρωνικό Δίκτυο για εκπαίδευση. Με άλλα λόγια, το training πραγματοποιείται πάνω στα παραγόμενα δείγματα που προκύπτουν από την διαδικασία ανάμειξης. Προτείνουμε μια παραλλαγή της τεχνικής αυτής, την οποία ονομάζουμε AddMixup. Σύμφωνα με αυτή, το δίκτυο, πέρα από τα εικονικά παραδείγματα, βλέπει και αντίστοιχα πραγματικά από το σύνολο εκπαίδευσης. Αναλυτικότερα, για κάθε mixed εικόνα $\tilde{x} = \lambda x_i + (1-\lambda)x_j$ που παράγεται, όταν αυτή πρόκειται να δοθεί στο δίκτυο για εκπαίδευση, τροφοδοτείται παράλληλα και η x_i. Με αυτό τον τρόπο, σε κάθε εποχή, το δίκτυο πραγματοποιεί ένα πέρασμα πάνω σε όλα τα εικονικά παραδείγματα αλλά και σε όλα τα δεδομένα εκπαίδευσης. Ουσιαστικά, αν N είναι ο συνολικός αριθμός των δεδομένων εκπαίδευσης, στην AddMixup το μοντέλο δέχεται $2N$ δείγματα ανά εποχή. Η τεχνική αυτή δεν εισάγει σημαντική υπολογιστική επιβάρυνση και επιτρέπει στο δίκτυο να έχει μια καλύτερη εποπτεία του γενικότερου προβλήματος κατηγοριοποίησης.

Η τεχνική AddMixup μπορεί να γίνει καλύτερα κατανοητή μέσω ενός παραδείγματος. Έστω ότι εκπαιδεύουμε ένα Νευρωνικό Δίκτυο με batches μεγέθους 4. Στην περίπτωση του κλασικού mixup, όπως φαίνεται και στο Σχήμα 4.4, 4 δείγματα του dataset (μπλε πλαίσιο) παρεμβάλλονται γραμμικά με 4 άλλα τυχαία δείγματα του dataset (κόκκινο πλαίσιο), οπότε και παράγονται 4 νέα συνθετικά δεδομένα (πράσινο πλαίσιο). Στη συνέχεια, τα παραγόμενα δείγματα τροφοδοτούνται στο δίκτυο για εκπαίδευση. Στην περίπτωση της AddMixup, μαζί με τα συνθετικά αυτά παραδείγματα, τροφοδοτούνται και άλλα τόσα αντίστοιχα πραγματικά. Στον εν λόγω απλοϊκό παράδειγμα, επομένως, εκτός από τις εικόνες στο πράσινο πλαίσιο, δίνονται ταυτόχρονα ως είσοδος στο μοντέλο και οι εικόνες που βρίσκονται στο μπλε πλαίσιο. Όπως θα δούμε στο Κεφάλαιο 5, η διαδικασία αυτή οδηγεί σε σημαντικά καλύτερη γενίκευση κατηγοριοποίησης.



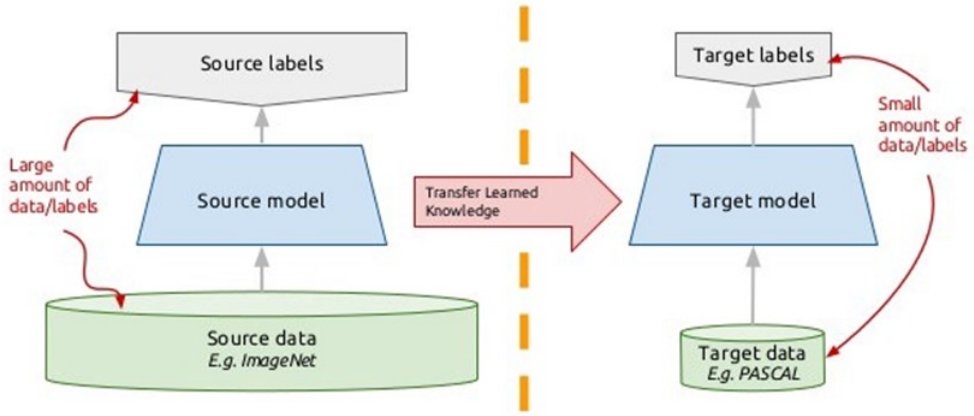
Σχήμα 4.4: Εφαρμογή mixup σε εικόνες εκφράσεων προσώπου

4.3 Μεταφορά Μάθησης

Όπως έχουμε ήδη αναφέρει, τα Δίκτυα Βαθιάς Μάθησης διαθέτουν έναν πολύ μεγάλο αριθμό παραμέτρων που πρέπει να προσδιοριστεί. Στην περίπτωση που το σύνολο εκπαίδευσης δεν έχει ικανοποιητικό μέγεθος, όπως συμβαίνει στα περισσότερα προβλήματα επιβλεπόμενης μάθησης, υπάρχει ο κίνδυνος της υπερεκπαίδευσης, κατά την οποία το δίκτυο απομνημονεύει τα δεδομένα εισόδου, αδυνατώντας να αποδώσει καλά στο γενικότερο πρόβλημα κατηγοριοποίησης. Μια ευρέως διαδεδομένη τεχνική που χρησιμοποιείται για την αντιμετώπιση του συγκεκριμένου προβλήματος είναι η Μεταφορά Μάθησης (Transfer Learning).

Σύμφωνα με αυτή, η γνώση που αποκομίζεται από την εκπαίδευση κάποιου DNN για ένα πρόβλημα κατηγοριοποίησης, μπορεί να διατηρηθεί και να "μεταφερθεί" στην εκπαίδευση ενός άλλου DNN, για την επίλυση κάποιου άλλου classification task. Στο Σχήμα 4.5 βλέπουμε μια διαγραμματική απεικόνιση της τεχνικής αυτής. Παρατηρούμε πως η γνώση που έχει αποκτηθεί κατά την εκπαίδευση ενός μοντέλου πάνω στην βάση δεδομένων ImageNet [93] αξιοποιείται κατά την εκπαίδευση ενός νέου μοντέλου, για κατηγοριοποίηση εικόνων της βάσης PASCAL [27]. Αξίζει να αναφέρουμε πως συνήθως η εκμετάλλευση της γνώσης πραγματοποιείται από κάποιο πρόβλημα κατηγοριοποίησης για το οποίο υπάρχει διαθέσιμος μεγάλος όγκος επισημειωμένων δεδομένων, σε αντίθεση με το πρόβλημα που αντιμετωπίζου-

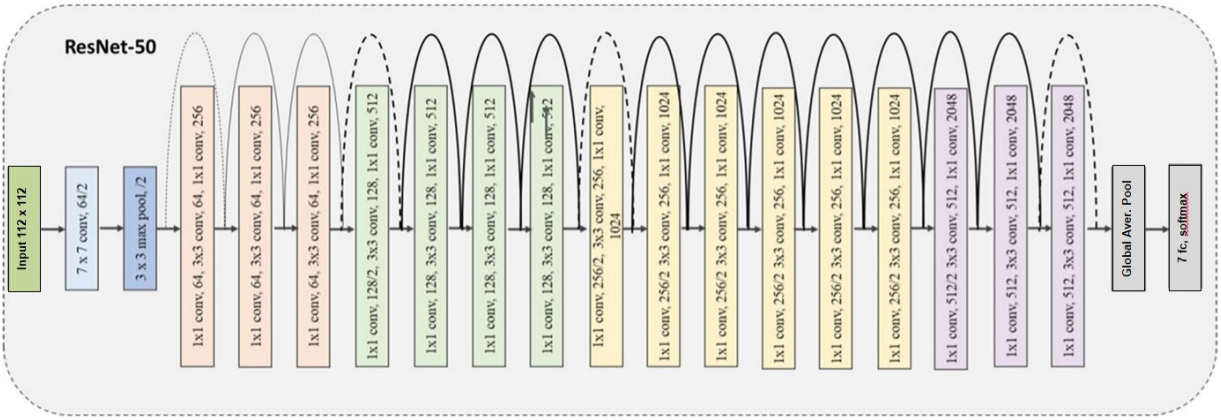
με, όπου τα δεδομένα είναι αρκετά λιγότερα. Στο παράδειγμα μας, η ImageNet αποτελεί μια τεράστια βάση με περισσότερες από 14 εκατομμύρια εικόνες, σε αντίθεση με την PASCAL που περιέχει μόλις μερικές χιλιάδες.



Σχήμα 4.5: Διάγραμμα απεικόνισης Μεταφοράς Μάθησης

4.4 Αρχιτεκτονική δικτύου

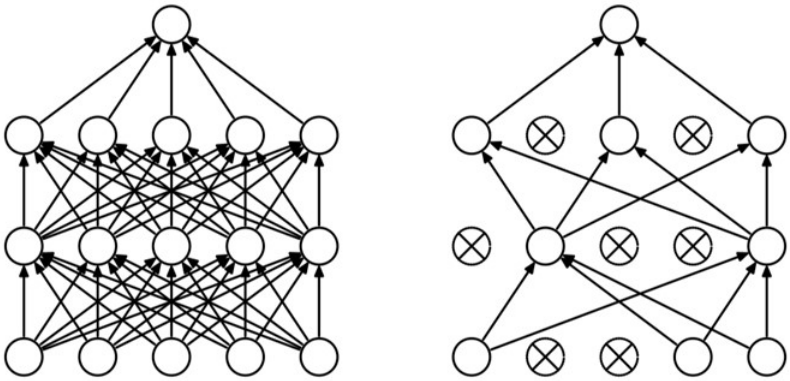
Βασιζόμαστε στην παραπάνω μέθοδο για την εκπαίδευση πάνω στις βάσεις δεδομένων AffectNet και RAF-DB. Συγκεκριμένα, χρησιμοποιούμε ένα ResNet-50 [38] με βάρη προεκπαιδευμένα πάνω στην ImageNet [93]. Όπως φαίνεται στο Σχήμα 4.6, από το pretrained δίκτυο, αφαιρείται η παλιά κεφαλή ταξινόμησης και στη θέση της προστίθεται ένα επίπεδο Global Average Pooling καθώς και ένα πλήρως συνδεδεμένο επίπεδο 7 νευρώνων (ένας νευρώνας για κάθε μία από τις 7 βασικές συναισθηματικές κατηγορίες). Τέλος, το Dense layer ακολουθείται από μια συνάρτηση ενεργοποίησης softmax για τον υπολογισμό της κατανομής πιθανότητας πάνω στις κλάσεις. Για την εκπαίδευση του δικτύου γίνονται unfreeze όλα τα συνελικτικά επίπεδα, καθώς σε αντίθετη περίπτωση έχουμε σημαντική μείωση στην ακρίβεια κατηγοριοποίησης.



Σχήμα 4.6: Αρχιτεκτονική ResNet-50 [38] με προσαρμοσμένη κεφαλή ταξινόμησης.

4.4.1 Dropout

Μια ευρέως διαδεδομένη τεχνική που χρησιμοποιείται για αντιμετώπιση του προβλήματος της υπερεκπαίδευσης είναι το dropout [98]. Αποτελεί μια μορφή κανονικοποίησης που συμβάλει στον περιορισμό των αλληλεξαρτήσεων μεταξύ διαφορετικών νευρώνων. Σύμφωνα με αυτό, κατά τη διάρκεια της εκπαίδευσης ενός Νευρωνικού Δικτύου, αγνοείται με τυχαίο τρόπο ένα ποσοστό των κόμβων. Μια οπτικοποίηση της τεχνικής απεικονίζεται στο Σχήμα 4.7. Όπως φαίνεται, ένα μέρος των συνολικών νευρώνων του δικτύου διαγράφεται και δεν συνεισφέρει κατά την εκπαιδευτική διαδικασία. Να σημειωθεί ότι στην φάση της αξιολόγησης, οι νευρώνες που προηγουμένως είχαν αγνοηθεί ενεργοποιούνται, ωστόσο η επιρροή τους είναι εξασθενημένη. Όπως θα δούμε στο Κεφάλαιο 5, σε ορισμένα πειράματα, αξιοποιούμε την εν λόγω τεχνική για να εξετάσουμε κατά πόσο βελτιώνει την απόδοση του δικτύου μας.



(a) Standard Neural Net (b) After applying dropout.

Σχήμα 4.7: Αρχιτεκτονική Δικτύου Βαθιάς Μάθησης χωρίς και με εφαρμογή dropout [98].

4.4.2 Υπερπαράμετροι βελτιστοποίησης

Κατά την εκπαίδευση του δικτύου, ως συνάρτηση απώλειας (loss function) χρησιμοποιείται η κατηγορική διασταυρούμενη εντροπία (categorical cross-entropy) ενώ ως βελτιστοποιητής (optimizer) ο Adam, με ρυθμό μάθησης (learning rate) 10^{-4}. Οι εικόνες εισόδου που δίνονται στο δίκτυο είναι κανονικοποιημένες RGB στο εύρος [0,1]. Η εκπαίδευση πραγματοποιείται με batch size μεγέθους 150 στην AffectNet και μεγέθους 32 στην RAF-DB.

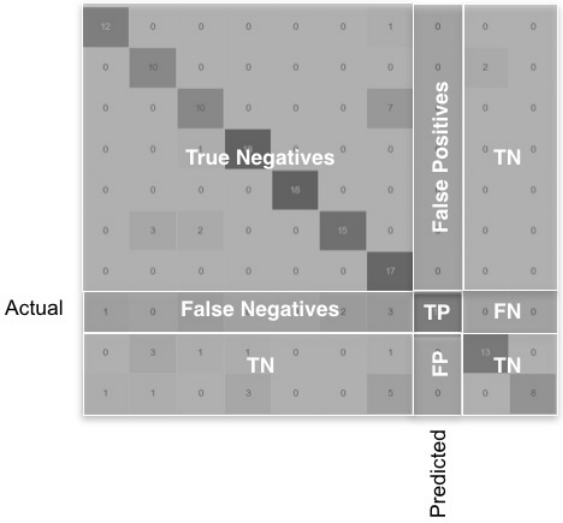
Κεφάλαιο 5

Πειραματική μελέτη και σχολιασμός

5.1 Μετρικές αξιολόγησης

Στη ενότητα αυτή θα αναλύσουμε τις μετρικές αξιολόγησης που χρησιμοποιούμε για την παρακολούθηση της απόδοσης των μοντέλων μας. Προτού ωστόσο τις ορίσουμε, χρειάζεται να προσδιορίσουμε τις βασικές έννοιες: **True positives** ($\boldsymbol{T_p}$), **False positives** ($\boldsymbol{F_p}$), **True negatives** ($\boldsymbol{T_n}$) και **False negatives** ($\boldsymbol{F_n}$) για μια δεδομένη κλάση $\boldsymbol{C_k}$.

- **True positives** ($\boldsymbol{T_p}$): Το σύνολο των δειγμάτων που ταξινομήθηκαν στην κλάση $\boldsymbol{C_k}$ και πράγματι άνηκαν στην κλάση $\boldsymbol{C_k}$.
- **False positives** ($\boldsymbol{F_p}$): Το σύνολο των δειγμάτων που ταξινομήθηκαν στην κλάση $\boldsymbol{C_k}$ ενώ δεν άνηκαν σε αυτή.
- **True negatives** ($\boldsymbol{T_p}$): Το σύνολο των δειγμάτων που δεν ταξινομήθηκαν στην κλάση $\boldsymbol{C_k}$ και πράγματι δεν άνηκαν στην κλάση $\boldsymbol{C_k}$.
- **False negatives** ($\boldsymbol{F_n}$): Το σύνολο των δειγμάτων που δεν ταξινομήθηκαν στην κλάση $\boldsymbol{C_k}$ ενώ άνηκαν σε αυτή.



Σχήμα 5.1: Πίνακας Σύγχυσης που απεικονίζει τα $\boldsymbol{T_p}$, $\boldsymbol{F_p}$, $\boldsymbol{T_n}$ και $\boldsymbol{F_n}$ για μια δεδομένη κλάση $\boldsymbol{C_k}$

Τώρα, με τη βοήθεια των παραπάνω εννοιών, μπορούμε να ορίσουμε ορισμένες μετρικές αξιολόγησης:

- **Πιστότητα - Accuracy**

  Αποτελεί την πιο συνηθισμένη μετρική αξιολόγησης και εκφράζει το σύνολο των ορθών προβλέψεων στο σύνολο όλων των δειγμάτων του dataset. Για ισορροπημένα σύνολα δεδομένων αποτελεί μια αρκετά αξιόπιστη ένδειξη της απόδοσης του μοντέλου. Ωστόσο, για imbalanced datasets μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα, καθώς στην περίπτωση αυτή οι υψηλές τιμές του accuracy δεν αποτελούν ένδειξη καλής προβλεπτικής ικανότητας για τις κλάσεις με μικρό αριθμό δειγμάτων. Αυτό αποτελεί πρόβλημα, καθώς σε αρκετές περιπτώσεις οι κατηγορίες που συγκεντρώνουν το μεγαλύτερο ενδιαφέρον είναι αυτές που περιλαμβάνουν και τη μειονότητα των δεδομένων (π.χ. ανίχνευση ανωμαλιών).

- **Ακρίβεια - Precision (P)**

  Ορίζεται ως ο λόγος των True positives (T_p) προς τον αριθμό των True positives (T_p) συν τον αριθμό των False positives (F_p). Εκφράζει δηλαδή το λόγο των δειγμάτων που ορθώς ταξινομήθηκαν σε μια κλάση προς το σύνολο όλων των δειγμάτων που προβλέφθηκε ότι ανήκουν στην κλάση αυτή. Είναι ιδιαίτερα χρήσιμη μετρική όταν το κόστος των False positives είναι υψηλό (π.χ. email spam detection).

$$P = \frac{T_p}{T_p + F_p}$$

- **Ανάκληση - Recall (R)**

  Ορίζεται ως ο λόγος των True positives (T_p) ως προς τον αριθμό των True positives (T_p) συν τον αριθμό των False negatives (F_n). Εκφράζει δηλαδή το λόγο των δειγμάτων που ορθώς ταξινομήθηκαν σε μια κλάση προς το σύνολο όλων των δειγμάτων που ανήκουν στην κλάση αυτή. Είναι ιδιαίτερα σημαντική μετρική όταν το κόστος των False negatives είναι υψηλό (π.χ. διάγνωση καρκίνου).

$$R = \frac{T_p}{T_p + F_n}$$

  Ιδανικά θέλουμε και υψηλή ακρίβεια και υψηλή ανάκληση, ωστόσο μεταξύ των δύο αυτών μετρικών υπάρχει γενικά ένα trade-off. Για παράδειγμα, στην οριακή περίπτωση ενός δυαδικού ταξινομητή που επιστρέφει σταθερά μόνο τη θετική κλάση, το Recall είναι 1 αλλά το Precision λαμβάνει τη μικρότερη δυνατή τιμή του. Γενικά, κατεβάζοντας το κατώφλι της απόφασης του ταξινομητή, αυξάνουμε την ανάκληση και μειώνουμε την ακρίβεια και αντιστρόφως.

- **F1-score ($F1$)**

  Ορίζεται ως ο αρμονικός μέσος της ακρίβειας (Precision) και της ανάκλησης (Recall). Χρησιμοποιείται όταν υπάρχει μια άνιση κατανομή των κλάσεων και θέλουμε να πετύχουμε μια ισορροπία μεταξύ (Precision) και (Recall). Είναι ιδιαίτερα χρήσιμη μετρική όταν το κόστος των False Negatives και False Positives είναι υψηλό. Σε αρκετά προβλήματα, το Precision ενδιαφέρει περισσότερο από ότι το Recall και αντίστροφα. Στην περίπτωση αυτή, είναι προφανές πως η χρήση του F1-score δεν εξυπηρετεί, καθώς σταθμίζει ισόποσα την ακρίβεια και την ανάκληση.

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} = \frac{2 \cdot T_p}{2 \cdot T_p + F_p + F_n}$$

Η ταξινόμηση εικόνων εκφράσεων προσώπου στα 7 βασικά συναισθήματα αποτελεί ένα **multi-classification** πρόβλημα. Θέλουμε λοιπόν με κάποιο τρόπο να συνδυάσουμε τις τιμές μια μετρικής για κάθε κλάση (π.χ. F1-score), έτσι ώστε να λάβουμε μια τελική τιμή που αντιπροσωπεύει τη συνολική ικανότητα κατηγοριοποίησης του μοντέλου μας. Υπάρχουν διάφοροι τρόποι να πραγματοποιηθεί αυτό. Ο απλούστερος από αυτούς είναι να λάβουμε τον

αριθμητικό μέσο των τιμών της μετρικής κάθε κλάσης. Η τεχνική αυτή ονομάζεται **macro-average**. Αντίστοιχα, μια άλλη μέθοδος είναι η **micro-average**, κατά την οποία όλες οι κλάσεις εξετάζονται μαζί. Συγκεκριμένα, για οποιαδήποτε εκ των μετρικών Precision, Recall και F1-score, η τεχνική **micro-average** δίνει την ίδια τιμή, η οποία συμπίπτει με αυτή του accuracy. Είναι δηλαδή:

$$\textbf{Precision\_micro-avg} = \textbf{Recall\_micro-avg} = \textbf{F1\_micro-avg} = \textbf{Accuracy}$$

Σε μη-ισορροπημένα datasets, όταν εξετάζουμε τη μετρική F1-score, προτιμάται η χρήση της τεχνικής macro-average, καθώς δίνει ίση βαρύτητα σε όλες τις κλάσεις, σε αντίθεση με την micro-average η οποία δίνει ίση βαρύτητα σε όλα τα δείγματα (το οποίο σημαίνει ότι όσο μεγαλύτερος είναι ο αριθμός των δειγμάτων μιας κλάσης τόσο μεγαλύτερη η επιρροή της στην τιμή της συνολικής μετρικής, όπως ακριβώς συμβαίνει στην μετρική Accuracy) [85].

5.2 Τεχνικές προεπεξεργασίας εικόνων

Σε ένα τυπικό πρόβλημα κατηγοριοποίησης συναισθήματος από ανθρώπινες εκφράσεις, προτού δοθούν οι εικόνες στο Νευρωνικό Δίκτυο για εκπαίδευση, πραγματοποιείται αυτόματος εντοπισμός και απομόνωση της περιοχής του προσώπου, μέσω κάποιου κατάλληλου face detector. Έπειτα, εντοπίζονται ορισμένα χαρακτηριστικά σημεία που σκιαγραφούν το περίγραμμα του κεφαλιού, τα μάτια, τα φρύδια, τη μύτη και το στόμα τα οποία ονομάζονται landmarks. Αυτά αξιοποιούνται στη συνέχεια, για την ευθυγράμμιση του προσώπου με χρήση κατάλληλου γεωμετρικού μετασχηματισμού.

Τόσο η διαδικασία της ανίχνευσης όσο και αυτή της ευθυγράμμισης προσώπου παίζουν ιδιαίτερα σημαντικό ρόλο στην βελτίωση της απόδοσης ενός Νευρωνικού Δικτύου. Η πρώτη επιτρέπει στο δίκτυο να εστιάσει μόνο στην περιοχή ενδιαφέροντος (πρόσωπο), αγνοώντας τα υπόλοιπα τμήματα της εικόνας. Η δεύτερη ενισχύει την ομοιομορφία ανάμεσα στις διαφορετικές εικόνες, θέτοντας μια κοινή βάση για όλα τα δείγματα εισόδου (πρόσωπα ευθυγραμμισμένα και κεντραρισμένα στις εικόνες). Αυτό ευνοεί τη δημιουργία αντιστοιχιών μεταξύ των διαφορετικών εκφράσεων προσώπου, ενισχύοντας την εξαγωγή χρήσιμων χαρακτηριστικών.

Τέλος, μια ακόμα τυπική διαδικασία προεπεξεραγασίας που εφαρμόζεται στις εικόνες εισόδου, προτού αυτές τροφοδοτηθούν στο δίκτυο, είναι η κανονικοποίηση (normalization). Κάθε εικόνα μετατρέπεται σε RGB κλίμακα (αν δεν βρίσκεται ήδη) και κανονικοποιείται στο εύρος [0,1], γεγονός που επιταχύνει την εκπαίδευση και σύγκλιση του μοντέλου.

5.2.1 AffectNet

Για την εκτέλεση των πειραμάτων αξιοποιούμε δύο διαφορετικές εκδόσεις της βάσης AffectNet, οι οποίες διαθέτουν επισημειώσεις ως προς τα 7 βασικά συναισθήματα. Στο ένα από τα δύο διαφορετικά versions έχει πραγματοποιηθεί τόσο ανίχνευση όσο και ευθυγράμμισή των προσώπων. Το face detection έχει γίνει με χρήση της βιβλιοθήκης OpenCV, ενώ το alignment έχει υλοποιηθεί αξιοποιώντας έναν αλγόριθμο παλινδρόμησης τοπικών δυαδικών χαρακτηριστικών [90]. Στην άλλη έκδοση της βάσης που χρησιμοποιούμε, έχει υλοποιηθεί μόνο η διαδικασία της ανίχνευση προσώπου και όχι αυτή της ευθυγράμμισης, την οποία και εφαρμόζουμε όπως περιγράφεται στη συνέχεια.

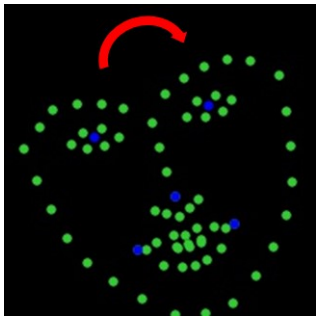
Για κάθε εικόνα προσώπου της βάσης, υπάρχει διαθέσιμη η ετικέτα της καθώς και ένας πίνακας με 68 facial landmarks. Αξιοποιούμε ορισμένα από τα 68 αυτά landmarks για τον υπολογισμό 5 σημείων ενδιαφέροντος σε κάθε εικόνα [100]. Αναλυτικότερα, βρίσκουμε πρώτα

το μέσον των σημείων που σκιαγραφούν την κόρη κάθε ματιού. Με βάση το Σχήμα 5.2, τα σημεία που οριοθετούν την κόρη του αριστερού ματιού είναι τα 38, 39, 41 και 42 ενώ του δεξιού είναι τα 44, 45, 47 και 48. Λαμβάνουμε επίσης την άκρη της μύτης (σημείο 31) καθώς και την αριστερή και δεξιά γωνία του στόματος (σημεία 49 και 55 αντίστοιχα). Τα 5 αυτά σημεία ενδιαφέροντος απεικονίζονται στο ίδιο σχήμα με μαύρες κουκκίδες.

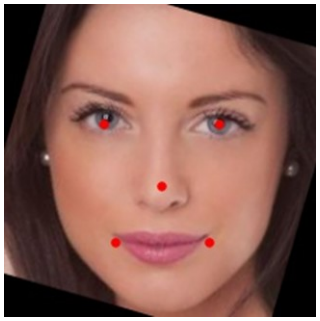
Σχήμα 5.2: Με κόκκινο σημειώνονται τα 68 facial landmarks που εξάγονται από μια εικόνα προσώπου [3], ενώ με μαύρο τα 5 σημεία ενδιαφέροντος που αξιοποιούνται για την ευθυγράμμιση.

Ορίζουμε τώρα κάποια αντίστοιχα σημεία αναφοράς (για το κέντρο των ματιών, την άκρη της μύτης και τις γωνίες στου στόματος) για ένα ιδανικά ευθυγραμμισμένο και κεντραρισμένο πρόσωπο (λαμβάνοντας υπόψη διαστάσεις ίδιες με αυτές των εικόνων). Υπολογίζουμε το μετασχηματισμό ομοιότητας (similarity transformation) μεταξύ των 5 σημείων ενδιαφέροντος που υπολογίσαμε και των αντίστοιχων 5 σημείων αναφοράς του πρότυπου προσώπου. Στη συνέχεια, εφαρμόζουμε το μετασχηματισμό αυτό σε ολόκληρη την εικόνα προσώπου για ευθυγράμμιση. Ένα παράδειγμα απεικονίζεται στο Σχήμα 5.3.

(α′) Αρχική εικόνα προσώπου πριν την ευθυγράμμιση



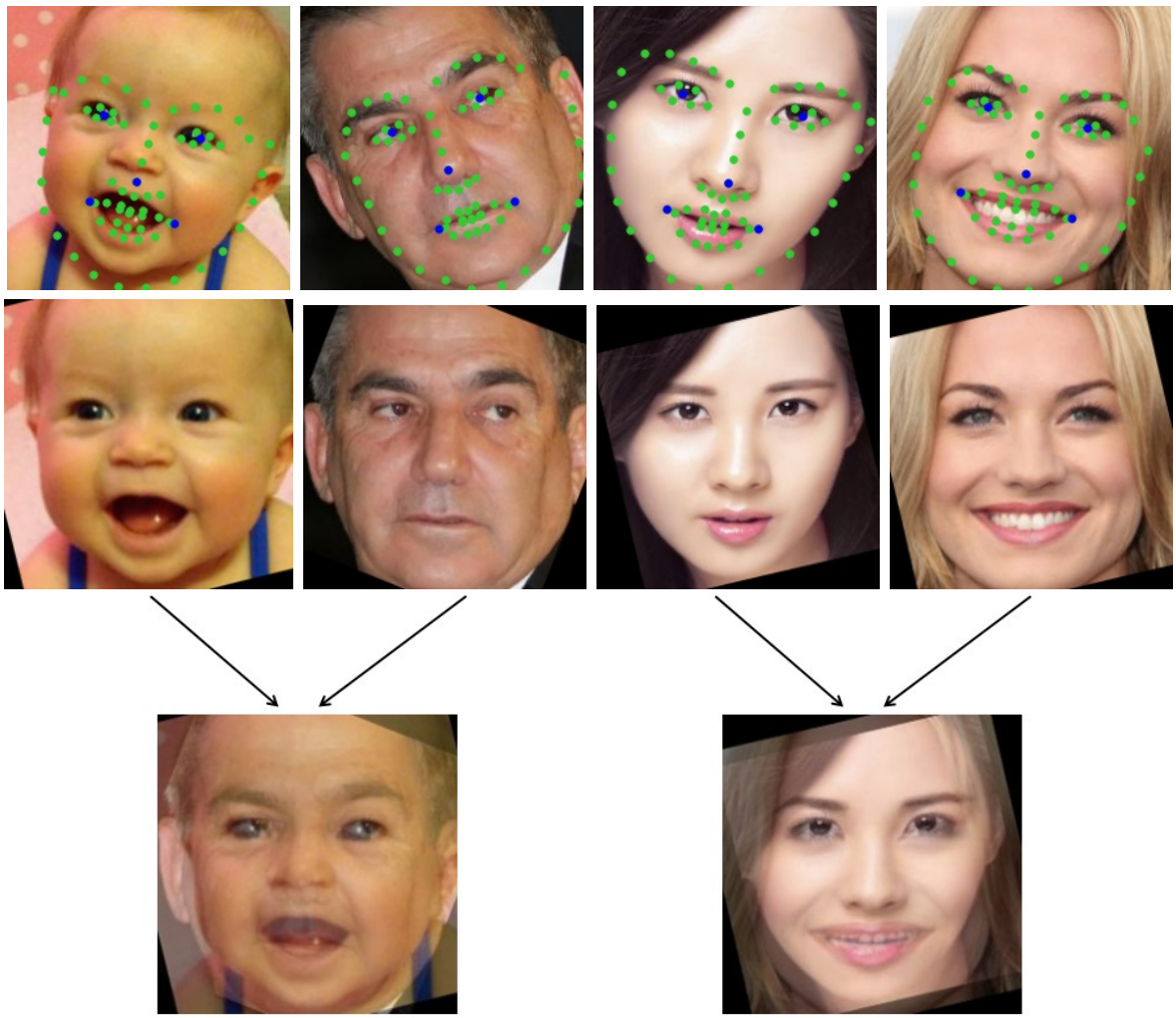
(β′) Τα 68 facial landmarks της εικόνας απεικονίζονται με πράσινο. Με μπλε απεικονίζονται τα 5 σημεία ενδιαφέροντος.



(γ′) Εικόνα προσώπου μετά την ευθυγράμμιση. Με κόκκινο απεικονίζονται τα 5 σημεία του πρότυπου προσώπου αναφοράς.

Σχήμα 5.3: Διαδικασία ευθυγράμμισης προσώπου με χρήση μετασχηματισμού ομοιότητας.

Στο Σχήμα 5.4 απεικονίζονται μερικά ακόμα ενδεικτικά παραδείγματα ευθυγράμμισης με βάση την παραπάνω μέθοδο. Στην πρώτη σειρά διακρίνουμε τις αρχικές, μη-ευθυγραμμισμένες, εικόνες ενώ στη δεύτερη βρίσκονται οι αντίστοιχες που προκύπτουν μετά τη διαδικασία του alignment. Σε κάθε εικόνα σημειώνονται με πράσινο χρώμα τα 68 διαθέσιμα facial landmarks ενώ με μπλε τα 5 σημεία ενδιαφέροντος που αξιοποιούνται κατά την ευθυγράμμιση.



Σχήμα 5.4: Παράδειγμα ευθυγράμμισης και ανάμειξης προσώπων της βάσης AffectNet.

Παρατηρούμε πως σε όλες τις ευθυγραμμισμένες εικόνες, το πρόσωπο είναι κεντραρισμένο στο μέσον της εικόνας ενώ η ευθεία των ματιών βρίσκεται στο ίδιο ύψος. Αυτό είναι ιδιαίτερα σημαντικό για την βελτίωση της απόδοσης κατηγοριοποίησης του μοντέλου που εκπαιδεύουμε. Ειδικά κατά την εφαρμογή της τεχνικής mixup, θέλουμε να πραγματοποιηθεί ανάμειξη των αντίστοιχων σημείων του προσώπου κάθε εικόνας (π.χ. το αριστερό μάτι της μίας εικόνας να συνδυαστεί με το αριστερό μάτι της άλλης, η μύτη της μίας με τη μύτη της άλλης κτλ). Στο κάτω μέρος του σχήματος φαίνονται δύο παραδείγματα γραμμικής παρεμβολής των ευθυγραμμισμένων προσώπων. Παρατηρούμε πως οι νέες εικόνες που προκύπτουν προσεγγίζουν φυσικά πρόσωπα τα οποία διαθέτουν μια μείξη από τα χαρακτηριστικά των επιμέρους προσώπων που συνδυάστηκαν. Αν η διαδικασία ευθυγράμμισης δεν ήταν επιτυχημένη θα είχαμε πολλά παραδείγματα συγκεχυμένων εικόνων κατά την διαδικασία της ανάμειξης, γεγονός που θα επιδρούσε αρνητικά στην απόδοση του μοντέλου. Μετά το πέρας και της διαδικασίας ευθυγράμμισης, οι εικόνες κανονικοποιούνται στο εύρος [0,1] και δίνονται ως είσοδος στο Δίκτυο Βαθιάς Μάθησης για εκπαίδευση.

5.2.2 RAF-DB

Για την εκτέλεση των πειραμάτων αξιοποιούμε δύο διαφορετικές εκδόσεις της βάσης RAF-DB, οι οποίες διαθέτουν επισημειώσεις ως προς τα 7 βασικά συναισθήματα. Και στις δύο έχει πραγματοποιηθεί τόσο ανίχνευση όσο και ευθυγράμμισή των προσώπων. Η διαφοροποίησή τους έγκειται στον τρόπο με τον οποίο εντοπίζεται και απομονώνεται η περιοχή γύρω από το πρόσωπο καθώς και στην εξαγωγή των landmarks. Στη μία έκδοση χρησιμοποιείται ο ανιχνευτής της βιβλιοθήκης dlib ενώ στην άλλη ο RetinaFace [17]. Σε κάθε περίπτωση οι εικόνες κανονικοποιούνται στο εύρος [0,1] προτού δοθούν ως είσοδος στο Νευρωνικό Δίκτυο για εκπαίδευση.

5.3 Πειράματα

5.3.1 AffectNet: Έκδοση 1η

Ξεκινάμε με τη βάση AffectNet και συγκεκριμένα με το version στο οποίο οι εικόνες προσώπων είναι ευθυγραμμισμένες. Εκπαιδεύουμε ένα ResNet-50 για 10 εποχές με την κλασική μέθοδο Ελαχιστοποίησης Επειρικού Ρίσκου (ERM). Έπειτα, πραγματοποιούμε εκπαίδευση με την τεχνική mixup για διάφορες τιμές της υπερπαραμέτρου α. Σε κάθε περίπτωση η ανάμειξη των δειγμάτων εκπαίδευσης πραγματοποιείται με τυχαίο τρόπο. Όλα τα μοντέλα αξιολογούνται πάνω στο σύνολο ελέγχου και τα αποτελέσματα συνοψίζονται στον Πίνακα 5.1.

Mixup	Accuracy (%)	F1-score (%)
ERM (no mixup)	**55,50**	55,09
$\alpha = 0.1$	54,47	54,75
$\alpha = 0.2$	54,50	54,37
$\alpha = 0.3$	54,16	54,21
$\boldsymbol{\alpha = 0.4}$	55,22	**55,20**
$\alpha = 1$	53,70	53,80

Πίνακας 5.1: Εκπαίδευση ResNet-50 χωρίς και με mixup στην 1η έκδοση της βάσης AffectNet

Παρατηρούμε πως για $\alpha = 0.4$ το εκπαιδευμένο δίκτυο εμφανίζει μια μικρή βελτίωση στη μετρική F1-score σε σχέση με την τυπική μέθοδο ERM. Για τις υπόλοιπες τιμές του α δεν διαπιστώνεται κάποια καλύτερη απόδοση. Στο σημείο αυτό δοκιμάζουμε να εισάγουμε dropout στο δίκτυο. Συγκεκριμένα, ορίζουμε ένα dropout layer μετά το τελευταίο συνελικτικό επίπεδο, πριν το Global Average Pooling και εκπαιδεύουμε εκ νέου με $\alpha = 0.4$. Παρατηρούμε πως η επίδοση του δικτύου βελτιώνεται ακόμα περισσότερο με χρήση του dropout, όπως φαίνεται στον Πίνακα 5.2.

Mixup	Dropout	Accuracy	F1-score
$\alpha = 0.4$	✗	55,22	54,21
	✓	**55,90**	**54,55**

Πίνακας 5.2: Εισαγωγή dropout μετά το τελευταίο συνελικτικό επίπεδο του ResNet-50.

Παρακινούμενοι από το παραπάνω αποτέλεσμα, εξετάζουμε γενικότερα την επίδραση του dropout στην εφαρμογή της mixup. Πιο συγκεκριμένα, ενσωματώνουμε τη μέθοδο αυτή και εκπαιδεύουμε εκ νέου με mixup για διάφορες παραμετροποιήσεις, όπως φαίνεται στον Πίνακα 5.3. Αυτή τη φορά μάλιστα, δοκιμάζουμε και τιμές α αρκετά μεγαλύτερες της μονάδας (π.χ. $\alpha = 8$, $\alpha = 32$).

Mixup	Accuracy (%)	F1-score (%)
ERM (no mixup)	56,33	55,20
$\boldsymbol{\alpha = 0.1}$	**56,76**	**56,07**
$\boldsymbol{\alpha = 0.2}$	**56,62**	**55,82**
$\alpha = 0.3$	56,13	55,27
$\alpha = 0.4$	55,90	54,55
$\alpha = 0.6$	55,82	55,61
$\alpha = 0.8$	54,79	52,72
$\boldsymbol{\alpha = 1}$	**56,99**	**55,92**
$\alpha = 2$	53,56	52,11
$\alpha = 3$	54,19	53,06
$\alpha = 4$	51,90	49,04
$\alpha = 8$	49,56	47,39
$\alpha = 32$	50,21	46,83

Πίνακας 5.3: Εκπαίδευση ResNet-50 με χρήση και dropout στην 1η έκδοση της βάσης AffectNet

Παρατηρούμε πως για τιμές $\alpha = 0.1, 0.2, 1$ το δίκτυο γενικεύει καλύτερα πάνω στα δεδομένα ελέγχου, σε σχέση με την εκπαίδευση με Ελαχιστοποίηση Εμπειρικού Ρίσκου. Φαίνεται επομένως, πως η κανονικοποίηση που εισάγει το dropout στο δίκτυο συνεισφέρει κατά την εφαρμογή της mixup.

Ένα άλλο αξιοσημείωτο συμπέρασμα είναι πως για τιμές $\alpha > 1$ υπάρχει μια αισθητή πτώση στην απόδοση. Στο σημείο αυτό να υπενθυμίσουμε πως, με βάση την Κατανομή Βήτα, όσο μεγαλύτερη η τιμή της υπερπαραμέτρου α τόσο πιο πιθανό να δειγματοληπτούνται τιμές για το λ οι οποίες βρίσκονται κοντά στο μέσον του διαστήματος [0,1]. Μάλιστα στην οριακή περίπτωση όπου $\alpha \to +\infty$, η κατανομή Βήτα ταυτίζεται με μια συνάρτηση Dirac στο 0.5, γεγονός που ισοδυναμεί με ισόποση ανάμειξη των δεδομένων εισόδου.

5.3.2 AffectNet: Έκδοση 2η

Προχωράμε τώρα στην 2η έκδοση της AffectNet που θα εξετάσουμε. Πρόκειται για το version στο οποίο έχει πραγματοποιηθεί περικοπή των εικόνων γύρω από την περιοχή του προσώπου αλλά όχι ευθυγράμμισή τους. Προτού εκτελέσουμε οποιοδήποτε πείραμα, εφαρμόζουμε τη διαδικασία του alignment, όπως αυτή περιγράφεται αναλυτικά στην Ενότητα 5.2.1. Εκπαιδεύουμε ξανά ένα ResNet-50 για 10 εποχές με την κλασική μέθοδο Ελαχιστοποίησης Εμπειρικού Ρίσκου χωρίς και με χρήση dropout. Έπειτα, εκπαιδεύουμε με mixup για διάφορες τιμές της υπερπαραμέτρου α τόσο χωρίς όσο και με dropout. Οι μετρικές αξιολόγησης των μοντέλων μας συνοψίζονται στον Πίνακα 5.4.

Mixup	Dropout	Accuracy	F1-score
ERM (no mixup)	✗	**57,69**	**56,79**
	✓	55,54	54,59
$\alpha = 0.1$	✗	53,57	51,92
	✓	53,06	50,45
$\alpha = 0.2$	✗	53,43	51,61
	✓	54,14	53,28
$\alpha = 0.4$	✗	54,54	53,17
	✓	54,91	54,02
$\alpha = 0.6$	✗	54,29	52,31
	✓	52,77	51,59
$\alpha = 1$	✗	54,71	53,40
	✓	54,06	52,93

Πίνακας 5.4: Εκπαίδευση ResNet-50 χωρίς και με dropout στην 2η έκδοση της βάσης AffectNet

Στη συγκεκριμένη έκδοση της AffectNet φαίνεται πως η mixup δεν συνεισφέρει στη βελτίωση της γενίκευσης του μοντέλου μας. Η προσθήκη dropout φαίνεται να λειτουργεί θετικά σε ορισμένες παραμετροποιήσεις ($\alpha = 0.2, 0.4$) ενώ σε άλλες επιδρά αρνητικά.

Στο [117], όπου οι συγγραφείς εφαρμόζουν τη mixup σε διάφορες στατικές βάσεις δεδομένων, διαπιστώνεται πως η συνεισφορά της μεθόδου είναι μεγαλύτερη στην περίπτωση που εφαρμόζεται σε πιο βαθιές αρχιτεκτονικές Νευρωνικών Δικτύων, όπως π.χ. σε ResNet-101 αντί για ResNet-50. Δοκιμάζουμε λοιπόν να εκτελέσουμε τα ίδια πειράματα με πριν αλλά σαν δίκτυο να χρησιμοποιήσουμε ένα ResNet-101. Τα αποτελέσματα συνοψίζονται στον Πίνακα 5.5.

Mixup	Dropout	Accuracy	F1-score
ERM (no mixup)	✗	**56,77**	**55,71**
	✓	56,77	55,27
$\alpha = 0.1$	✗	55,26	54,06
	✓	54,23	52,37
$\alpha = 0.2$	✗	54,66	53,04
	✓	55,57	54,02
$\alpha = 0.4$	✗	55,06	53,74
	✓	54,11	52,60
$\alpha = 0.6$	✗	52,97	50,61
	✓	53,74	52,14

Πίνακας 5.5: Εκπαίδευση ResNet-101 χωρίς και με dropout στην 2η έκδοση της βάσης AffectNet

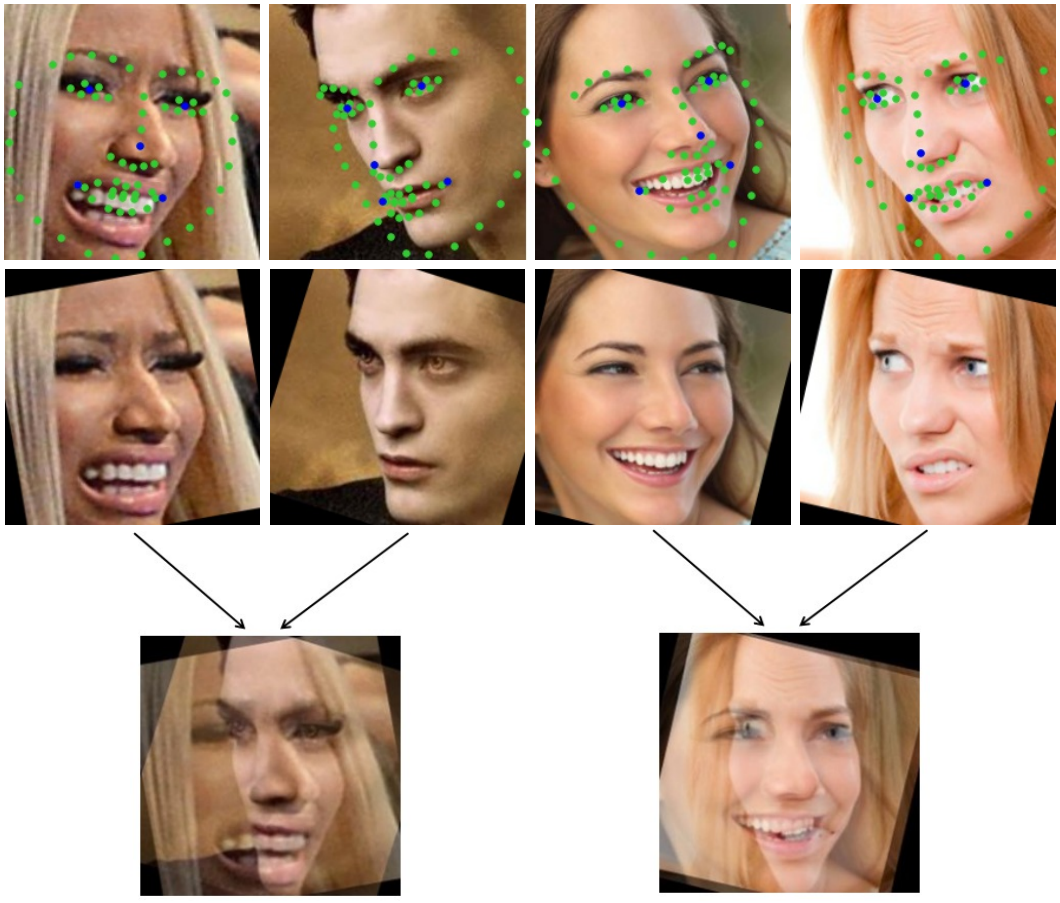
Τα συμπεράσματα που προκύπτουν από τα πειράματα είναι αντίστοιχα αυτών που λάβαμε με εκπαίδευση του ResNet-50. Φαίνεται δηλαδή πως η mixup δεν οδηγεί σε υψηλότερη ακρίβεια ενώ το dropout δίνει καλύτερα αποτελέσματα μόνο σε κάποιες παραμετροποιήσεις.

5.3.3 Σχολιασμός και παρατηρήσεις

Τα αποτελέσματα των πειραμάτων στην 1η έκδοση της βάσης AffectNet καταδεικνύουν πως η τεχνική mixup μπορεί να συνεισφέρει στην βελτίωση της γενίκευσης του δικτύου μας στο πρόβλημα της κατηγοριοποίησης εικόνων στα 7 βασικά συναισθήματα. Επίσης, φαίνεται πως αν στο δίκτυο προστεθεί επιπλέον κανονικοποίηση, μέσω dropout, τα αποτελέσματα μπορεί να βελτιωθούν ακόμα περισσότερο.

Στην 2η έκδοση της βάσης, φαίνεται πως η τεχνική δεν επιφέρει κάποια θετική επίδραση, ακόμα και όταν αυτή συνδυάζεται με dropout. Το γεγονός αυτό μπορεί να οφείλεται σε διάφορους παράγοντες:

- **Μη αποτελεσματική ευθυγράμμιση μερικών προσώπων της βάσης**: Αυτός ενδεχομένως να είναι και ένας από τους βασικότερους λόγος για τους οποίους η τεχνική mixup δείχνει να μην συνεισφέρει στη συγκεκριμένη έκδοση της βάσης. Παρατηρώντας λίγο πιο αναλυτικά τις εικόνες, διαπιστώνουμε πως ένα μέρος αυτών απεικονίζουν πρόσωπα τα οποία δεν έχουν το βλέμμα τους στραμμένο προς το φακό, αλλά κοιτάζουν προς κάποια κατεύθυνση. Τα πρόσωπα αυτά εξακολουθούν να έχουν γυρισμένο το βλέμμα τους προς την κατεύθυνση αυτή και μετά την διαδικασία της ευθυγράμμισης. Αυτό έχει σαν αποτέλεσμα στη mixup, η ανάμειξή τους με άλλες εικόνες να μην είναι ιδιαίτερα αποτελεσματική, καθώς παρεμβάλλονται διαφορετικά σημεία των προσώπων που συνδυάζονται, γεγονός που δημιουργεί ένα συγκεχυμένο αποτέλεσμα και δεν βοηθάει το δίκτυο να εξάγει χρήσιμα χαρακτηριστικά. Στο Σχήμα 5.5 απεικονίζονται ορισμένα χαρακτηριστικά παραδείγματα.



Σχήμα 5.5: Ευθυγράμμιση και ανάμειξη προσώπων που είναι στραμμένα αριστερά ή δεξιά.

- **Τυχαιότητα κατά την ανάμειξη των δειγμάτων**: Η επιλογή των δειγμάτων εκπαίδευσης που παρεμβάλλονται καθώς και η ένταση της παρεμβολής (καθορίζεται από το λ) ορίζονται με τυχαίο τρόπο. Αυτό σημαίνει πως πρακτικά υπάρχουν άπειροι δυνατοί συνδυασμοί δεδομένων και κατά συνέπεια άπειρα διαφορετικά συνθετικά παραδείγματα πάνω στα οποία μπορεί να εκπαιδευτεί το δίκτυο μας. Όπως είδαμε και προηγουμένως, υπάρχουν περιπτώσεις όπου η παρεμβολή εικόνων οδηγεί σε ένα καλό αποτέλεσμα αναμειγμένου προσώπου, ωστόσο υπάρχουν και άλλες στις οποίες η παραγόμενη εικόνα δεν συμβάλει στην βελτίωση της απόδοσης του δικτύου. Επομένως, όσο πιο "καλή" τυχαίνει να είναι η ανάμειξη των δεδομένων εκπαίδευσης, τόσο πιο υψηλή και η ικανότητα γενίκευσης του δικτύου πάνω στα δεδομένα ελέγχου.
- **Τυχαιότητα κατά την διαγραφή νευρώνων**: Η τεχνική κανονικοποίησης dropout, κατά την εκπαίδευση, αγνοεί με τυχαίο τρόπο ένα ποσοστό των κόμβων του δικτύου. Ανάλογα με το ποιοι νευρώνες θα επιλεγούν να μην ληφθούν υπόψη ενδέχεται τα αποτελέσματα λίγο έως πολύ να διαφέρουν.
- **Μικρός αριθμός εποχών**: Στα πλαίσια των πειραμάτων που εξετάζουμε, πραγματοποιούμε εκπαίδευση του δικτύου για 10 εποχές. Λόγω του μεγέθους της AffectNet, ενδέχεται η σύγκλιση να επιτυγχάνεται μετά από μεγαλύτερο αριθμό εποχών.

5.3.4 RAF-DB: Έκδοση 1η

Έχοντας ολοκληρώσει έναν ικανοποιητικό αριθμό πειραμάτων σε δύο διαφορετικές εκδόσεις της βάσης AffectNet, προχωράμε σε μια αρκετά μικρότερη in-the-wild στατική βάση, την RAF-DB. Στόχος είναι να εξετάσουμε αν και κατά πόσον μπορεί να συνεισφέρει η mixup σε πιο μικρά facial databases. Και σε αυτή την περίπτωση εξετάζουμε δύο διαφορετικά versions, οι εικόνες των οποίων έχουν ευθυγραμμιστεί αξιοποιώντας διαφορετικούς ανιχνευτές. Ξεκινάμε με μία εξ' αυτών η οποία περιέχει RGB εικόνες διάστασης $100 \times 100 \times 3$. Εκπαιδεύουμε πάνω σε αυτή ένα ResNet-50 για 50 εποχές με την κλασική μέθοδο Ελαχιστοποίησης Εμπειρικού Ρίσκου (ERM). Έπειτα, πραγματοποιούμε εκπαίδευση με την τεχνική mixup για διάφορες τιμές της υπερπαραμέτρου α. Τέλος, αξιολογούμε όλα τα μοντέλα πάνω στο σύνολο ελέγχου, όπως φαίνεται στον Πίνακα 5.6.

Mixup	Dropout	Accuracy	F1-score
ERM (no mixup)	✗	79,92	71,66
	✓	80,59	**71,67**
$\alpha = 0.1$	✗	80,40	71,20
	✓	80,18	70,16
$\boldsymbol{\alpha = 0.2}$	✗	80,57	70,53
	✓	**81,09**	70,93
$\alpha = 0.4$	✗	78,84	68,70
	✓	79,36	69,42
$\alpha = 0.6$	✗	79,39	70,04
	✓	79,95	70,67
$\alpha = 1$	✗	76,80	66,55
	✓	75,51	66,16

Πίνακας 5.6: Εκπαίδευση ResNet-50 χωρίς και με dropout στην 1η έκδοση της RAF-DB

Παρατηρούμε πως για $\alpha = 0.2$ και ύπαρξη dropout στο δίκτυο, λαμβάνουμε την καλύτερη δυνατή ακρίβεια κατηγοριοποίησης. Ωστόσο, κανένα από τα μοντέλα που εκπαιδεύτηκε με την τεχνική mixup δεν εμφανίζει καλύτερη γενίκευση ως προς την μετρική F1-score, συγκριτικά με την κλασική μέθοδο ERM.

Στο σημείο αυτό εφαρμόζουμε μια εναλλακτική της μεθόδου mixup, την Addmixup, η οποία επεξηγείται λεπτομερώς στην Ενότητα 4.2.1. Εν ολίγοις, σύμφωνα με την τεχνική αυτή, σε αντίθεση με την παραδοσιακή προσέγγιση mixup, η εκπαίδευση του δικτύου πραγματοποιείται τόσο πάνω στα εικονικά δείγματα που παράγονται όσο και στα ίδια τα δεδομένα εκπαίδευσης. Η ιδέα είναι ότι με αυτό τον τρόπο το δίκτυο αποκτάει μια γρηγορότερη και καλύτερη εποπτεία του γενικότερου προβλήματος κατηγοριοποίησης. Παράλληλα μετριάζεται και το πρόβλημα της μη-αποτελεσματικής ευθυγράμμισης ορισμένων προσώπων, αφού κάθε φορά που το δίκτυο τροφοδοτείται με μια mixed εικόνα, δέχεται ταυτόχρονα ως είσοδο και την αντίστοιχη αρχική εικόνα του dataset.

Δοκιμάζουμε λοιπόν να εκπαιδεύουμε ένα ResNet-50 για 50 εποχές με την τεχνική Addmixup για τις ίδιες τιμές της υπερπαραμέτρου α. Έπειτα αξιολογούμε όλα τα μοντέλα πάνω στο σύνολο ελέγχου, όπως φαίνεται στον Πίνακα 5.7.

Addmixup	Dropout	Accuracy	F1-score
ERM (no mixup)	✗	79,92	71,66
	✓	80,59	71,67
$\boldsymbol{\alpha = 0.1}$	✗	**82,33**	73,07
	✓	82,13	**73,41**
$\boldsymbol{\alpha = 0.2}$	✗	**82,33**	73,21
	✓	81,58	72,41
$\alpha = 0.4$	✗	81,38	72,87
	✓	80,01	70,54
$\boldsymbol{\alpha = 0.6}$	✗	81,77	71,86
	✓	**82,30**	73,25

Πίνακας 5.7: Εκπαίδευση ResNet-50 με την τεχνική Addmixup στην 1η έκδοση της RAF-DB

Παρατηρούμε πως για όλες σχεδόν τις εκτελέσεις, η τεχνική Addmixup οδηγεί σε καλύτερα αποτελέσματα συγκριτικά με την κλασική μέθοδο Ελαχιστοποίησης Εμπειρικού Ρίσκου. Μάλιστα, η βέλτιστη συνεισφορά στην ικανότητα γενίκευσης του μοντέλου προκύπτει για μικρές τιμές του α (0,1 και 0,2), όπου τόσο η μετρική Accuracy όσο και η F1-score εμφανίζουν μια άνοδο της τάξεως του 2%.

Για να γίνει καλύτερα αντιληπτή η συνεισφορά της Addmixup, στον Πίνακα 5.8 παραθέτουμε συγκεντρωτικά τα αποτελέσματα για την εκπαίδευση του ResNet-50 για 50 εποχές με τις μεθόδους mixup και Addmixup. Όπως μπορούμε εύκολα να διακρίνουμε, για οποιαδήποτε παραμετροποίηση α και ανεξαρτήτως εισαγωγής ή όχι του dropout, η τεχνική Addmixup οδηγεί πάντοτε σε αισθητά καλύτερα αποτελέσματα.

Hyperparameter	Dropout	Method	Accuracy		F1-score	
$\alpha = 0.1$	✗	Mixup	80,40	**+1,93**	71,20	**+1,87**
		Addmixup	82,33		73,07	
	✓	Mixup	80,18	**+1,95**	70,16	**+3,25**
		Addmixup	82,13		73,41	
$\alpha = 0.2$	✗	Mixup	80,57	**+1,76**	70,53	**+2,68**
		Addmixup	82,33		73,21	
	✓	Mixup	81,09	**+0,49**	70,93	**+1,48**
		Addmixup	81,58		72,41	
$\alpha = 0.4$	✗	Mixup	78,84	**+2,54**	68,70	**+4,17**
		Addmixup	81,38		72,87	
	✓	Mixup	79,36	**+0,68**	69,42	**+1,12**
		Addmixup	80,01		70,54	
$\alpha = 0.6$	✗	Mixup	79,39	**+2,38**	70,04	**+1,82**
		Addmixup	81,77		71,86	
	✓	Mixup	79,95	**+2,35**	70,67	**+2,58**
		Addmixup	82,30		73,25	

Πίνακας 5.8: Σύγκριση των τεχνικών mixup και Addmixup στην 1η έκδοση της RAF-DB

Στον Πίνακα 5.9 παραθέτουμε την ελάχιστη, τη μέγιστη και τη μέση μεταβολή στην απόδοση του ResNet-50 όταν αυτό εκπαιδεύεται για 50 εποχές τη μέθοδο Addmixup, έναντι της παραδοσιακής mixup, για τις παραμετροποιήσεις που εξετάσαμε. Παρατηρούμε πως σχετικά με το Accuracy, υπάρχει μια άνοδος από 0,49-2,54%, με μέση αύξηση 1,76%. Αντίστοιχα, η θετική επίδραση στη μετρική F1-score είναι ακόμα πιο αισθητή, καθώς αυτή κυμαίνεται μεταξύ 1,12% και 4,17%, με μέση αύξηση 2,37%. Μάλιστα η F1-score αποτελεί μια ακόμα πιο καλή ένδειξη της ικανότητας γενίκευσης του μοντέλου μας, δεδομένου ότι το σύνολο επαλήθευσης είναι μη-ισορροπημένο.

Evaluation Metric	Min	Max	Mean
Accuracy	0,49	2,54	1,76
F1-score	1,12	4,17	2,37

Πίνακας 5.9: Στατιστικά για την βελτιωμένη απόδοση της Addmixup έναντι της mixup

Στο σημείο αυτό λαμβάνουμε τις παραμετροποιήσεις που έδωσαν την καλύτερη απόδοση ($\alpha = 0.1, 0.2, 0.6$) και εκπαιδεύουμε ένα μοντέλο ResNet-50 για 100 εποχές με τη μέθοδο Addmixup και με τη μέθοδο ERM. Επειδή όπως αναφέραμε το dataset είναι μη-ισορροπημένο, για την καλύτερη μελέτη, εισάγουμε και μια επιπλέον μετρική αξιολόγησης, την Average Accuracy. Η μετρική αυτή εκφράζει τη μέση ακρίβεια όλων των κλάσεων και υπολογίζεται ως η μέση τιμή της διαγωνίου του κανονικοποιημένου Πίνακα Σύγχυσης. Στην ουσία υπολογίζεται η Ανάκληση (Recall) κάθε κλάσης και έπειτα λαμβάνεται ο μέσος όρος αυτών, δηλαδή η τιμή της μετρικής ταυτίζεται με αυτή της Macro Recall.

Addmixup	Dropout	Accuracy	F1-score	Average Accuracy
ERM (no mixup)	✗	82,00	73,55	71,16
	✓	82,33	74,02	71,86
$\boldsymbol{\alpha = 0.1}$	✗	82,20	73,14	71,42
	✓	82,56	**74,09**	71,80
$\boldsymbol{\alpha = 0.2}$	✗	**82,78**	73,92	71,85
	✓	82,62	73,72	71,54
$\boldsymbol{\alpha = 0.6}$	✗	82,39	73,88	**71,93**
	✓	82,00	73,55	71,16

Πίνακας 5.10: Εκπαίδευση δικτύου για 100 εποχές με Addmixup στην 1η έκδοση της RAF-DB

Όπως φαίνεται στον Πίνακα 5.10, και για τις τρεις μετρικές αξιολόγησης, η βέλτιστη τιμή λαμβάνεται για κάποια παραμετροποίηση της μεθόδου Addmixup. Συνεπώς, συμπεραίνουμε πως αυτή οδηγεί σε καλύτερη ικανότητα γενίκευσης του μοντέλου συγκριτικά με την κλασική εκπαίδευση με ERM.

5.3.5 RAF-DB: Έκδοση 2η

Προχωράμε τώρα στην 2η έκδοση της RAF-DB που θα εξετάσουμε. Και σε αυτό το version οι εικόνες προσώπου έχουν ευθυγραμμιστεί. Ακολουθούμε μια αντίστοιχη πειραματική διαδικασία με αυτή που εφαρμόστηκε στην άλλη έκδοση της βάσης. Ξεκινάμε δηλαδή εκπαιδεύοντας ένα ResNet-50 για 50 εποχές με την κλασική μέθοδο Ελαχιστοποίησης Εμπειρικού Ρίσκου, χωρίς και με χρήση dropout. Έπειτα, εκπαιδεύουμε με mixup για διάφορες τιμές της υπερπαραμέτρου α τόσο χωρίς όσο και με dropout. Η αξιολόγηση των μοντέλων μας συνοψίζεται στον Πίνακα 5.11.

Mixup	Dropout	Accuracy	F1-score	Average Accuracy
ERM (no mixup)	✗	82,63	74,12	72,61
	✓	**83,02**	**75,47**	**74,07**
$\alpha = 0.1$	✗	82,69	73,88	71,83
	✓	82,56	73,54	71,53
$\alpha = 0.2$	✗	82,37	73,59	71,69
	✓	82,66	73,96	72,02
$\alpha = 0.4$	✗	81,39	71,64	69,57
	✓	81,10	71,82	70,17
$\alpha = 0.6$	✗	80,96	72,09	69,49
	✓	80,90	71,86	69,51
$\alpha = 1$	✗	79,99	70,68	68,71
	✓	80,67	71,69	69,05

Πίνακας 5.11: Εκπαίδευση ResNet-50 χωρίς και με dropout στην 2η έκδοση της RAF-DB

Σε αυτή την έκδοση της βάση φαίνεται πως η κλασική τεχνική mixup, αν και προσεγγίζει για κάποιες παραμετροποιήσεις την μέθοδο ERM, αδυνατεί να συμβάλει σε βελτιωμένη γενίκευση. Το γεγονός αυτό είναι πιθανό να οφείλεται σε ένα συνδυασμό παραγόντων που αναφέραμε και κατά την πειραματική μελέτη της βάσης AffectNet (παραδείγματα μη-αποτελεσματικής ευθυγράμμισης προσώπων, τυχαιότητα κατά την εφαρμογή mixup και dropout).

Στο σημείο αυτό εξετάζουμε την τεχνική Addmixup, η οποία έδωσε πολύ ικανοποιητικά αποτελέσματα στην προηγούμενη έκδοση της βάσης. Εκπαιδεύουμε λοιπόν εκ νέου ένα ResNet-50 για 50 εποχές με τη μέθοδο αυτή, για τιμές της υπερπαραμέτρου α ίδιες με πριν. Στη συνέχεια, αξιολογούμε όλα τα εκπαιδευμένα μοντέλα πάνω στο σύνολο ελέγχου, ως προς τις μετρικές Accuracy, F1-score και Average Accuracy. Τα αποτελέσματα φαίνονται στον Πίνακα 5.12.

Addmixup	**Dropout**	**Accuracy**	**F1-score**	**Average Accuracy**
ERM (no mixup)	✗	82,63	74,12	72,61
	✓	83,02	75,47	74,07
$\boldsymbol{\alpha = 0.1}$	✗	**84,45**	76,37	74,72
	✓	83,67	76,03	73,77
$\boldsymbol{\alpha = 0.2}$	✗	84,19	76,57	74,74
	✓	84,39	**76,64**	**75,26**
$\alpha = 0.4$	✗	84,13	75,04	73,38
	✓	84,26	76,46	74,38
$\alpha = 0.6$	✗	83,74	75,43	73,87
	✓	83,51	75,58	74,36

Πίνακας 5.12: Εκπαίδευση ResNet-50 με την τεχνική Addmixup στην 2η έκδοση της RAF-DB

Παρατηρούμε πως για οποιαδήποτε τιμή της υπερπαραμέτρου α, είτε με είτε χωρίς την επιπλέον κανονικοποίηση που εισάγει το dropout, η Addmixup οδηγεί σε αρκετά καλύτερα αποτελέσματα συγκριτικά με την κλασική μέθοδο Ελαχιστοποίησης Εμπειρικού Ρίσκου. Μάλιστα οι καλύτερες επιδόσεις του μοντέλου προκύπτουν για αρκετά μικρές τιμές της υπερπαραμέτρου α. Συγκεκριμένα, η βέλτιστη τιμή για την μετρική Accuracy (84.45%) εμφανίζεται όταν $\alpha = 0.1$ ενώ οι υψηλότερες τιμές για τις F1-score και Average Accuracy (76.64% και 75.26% αντίστοιχα) προκύπτουν όταν $\alpha = 0.2$. Φαίνεται επομένως για άλλη μια φορά πως, όταν το δίκτυο εκπαιδεύεται ταυτόχρονα πάνω στα εικονικά παραδείγματα αλλά και στα αντίστοιχα πραγματικά, εμφανίζει αυξημένη ικανότητα γενίκευσης.

Για να μπορέσουμε να έχουμε μια καλύτερη εικόνα του κατά πόσον συνεισφέρει η τεχνική Addmixup έναντι της παραδοσιακής μεθόδου, σχηματίζουμε τον πίνακα Πίνακα 5.13. Εκεί παραθέτουμε συγκεντρωτικά τα αποτελέσματα για την εκπαίδευση του ResNet-50 για 50 εποχές με τις μεθόδους mixup και Addmixup. Όπως μπορούμε εύκολα να διακρίνουμε, για οποιαδήποτε παραμετροποίηση α και ανεξαρτήτως προσθήκης ή μη του dropout, η τεχνική Addmixup οδηγεί πάντοτε σε αισθητά καλύτερα αποτελέσματα.

Hyperparameter	Dropout	Method	Accuracy		F1-score		Avrg. Acc.	
$\alpha = 0.1$	✗	Mixup	82,69	**+1,76**	73,88	**+2,49**	71,83	**+2,89**
		Addmixup	84,45		76,37		74,72	
	✓	Mixup	82,56	**+1,11**	73,54	**+2,49**	71,53	**+2,24**
		Addmixup	83,67		76,03		73,77	
$\alpha = 0.2$	✗	Mixup	82,37	**+1,82**	73,59	**+2,97**	71,69	**+3,05**
		Addmixup	84,19		76,56		74,74	
	✓	Mixup	82,66	**+1,73**	73,96	**+2,68**	72,02	**+3,24**
		Addmixup	84,39		76,64		75,26	
$\alpha = 0.4$	✗	Mixup	81,39	**+2,74**	71,64	**+3,4**	69,57	**+3,81**
		Addmixup	84,13		75,04		73,38	
	✓	Mixup	81,10	**+3,16**	71,82	**+4,82**	70,17	**+4,21**
		Addmixup	84,26		76,46		74,38	
$\alpha = 0.6$	✗	Mixup	80,96	**+2,78**	72,09	**+3,34**	69,49	**+4,38**
		Addmixup	83,74		75,43		73,87	
	✓	Mixup	80,90	**+2,61**	71,86	**+3,72**	69,51	**+4,85**
		Addmixup	83,51		75,58		74,36	

Πίνακας 5.13: Σύγκριση των τεχνικών mixup και Addmixup στην 2η έκδοση της RAF-DB

Στον Πίνακα 5.14 παραθέτουμε την ελάχιστη, τη μέγιστη και τη μέση μεταβολή στην απόδοση του ResNet-50 όταν αυτό εκπαιδεύεται για 50 εποχές με τη μέθοδο Addmixup, έναντι της παραδοσιακής τεχνικής mixup, για τις παραμετροποιήσεις που εξετάσαμε. Παρατηρούμε πως στο Accuracy υπάρχει μια άνοδος 1,11% - 3,16%, με μέση αύξηση 2,21%. Στις άλλες δύο μετρικές, οι οποίες αποτελούν και καλύτερη ένδειξη της ικανότητας γενίκευσης του μοντέλου μας (μη-ισορροπημένο dataset), η βελτίωση είναι ακόμα πιο αισθητή. Συγκεκριμένα, η F1-score παρουσιάζει μια άνοδο που κυμαίνεται μεταξύ 2,49% και 4,82%, με μέση άνοδο 3,24%. Αντίστοιχα, το Average Accuracy αυξάνεται 2,24% - 4,85%, έχοντας τη μεγαλύτερη μέση αύξηση με 3,58%. Παρατηρούμε επομένως, πως στη συγκεκριμένη έκδοση της RAF-DB, η τεχνική Addmixup συνεισφέρει σημαντικά στη βελτίωση της απόδοσης του δικτύου μας, σε αντίθεση με τη μέθοδο mixup, η οποία δεν επιδρά θετικά.

Evaluation Metric	Min	Max	Mean
Accuracy	1,11	3,16	2,21
F1-score	2,49	4,82	3,24
Average Accuracy	2,24	4,85	3,58

Πίνακας 5.14: Στατιστικά για την βελτιωμένη απόδοση της Addmixup έναντι της mixup

Στο σημείο αυτό λαμβάνουμε τις παραμετροποιήσεις που έδωσαν την καλύτερη απόδοση ($\alpha = 0.1, 0.2$) και εκπαιδεύουμε ένα μοντέλο ResNet-50 για 100 εποχές με τη μέθοδο Addmixup και με τη μέθοδο ERM. Τα αποτελέσματα φαίνονται στον Πίνακα 5.15.

Addmixup	Dropout	Accuracy	F1-score	Average Accuracy
ERM (no mixup)	✗	84,00	75,59	73,70
	✓	82,66	74,35	73,77
$\alpha = 0.1$	✗	**85,04**	**77,30**	**75,13**
	✓	83,96	75,50	73,74
$\alpha = 0.2$	✗	84,32	76,45	75,06
	✓	84,75	75,75	74,09

Πίνακας 5.15: Εκπαίδευση δικτύου για 100 εποχές με Addmixup στην 2η έκδοση της RAF-DB

Παρατηρούμε πως για $\alpha = 0.1$ λαμβάνουμε τις υψηλότερες τιμές και για τις τρεις μετρικές αξιολόγησης. Γενικότερα όλες οι εκπαιδεύσεις με την Addmixup οδηγούν σε καλύτερα αποτελέσματα από ότι η κλασική μέθοδος Ελαχιστοποίησης Εμπειρικού Ρίσκου.

Για να γίνει ακόμα καλύτερα αντιληπτή η συνεισφορά της εν λόγω τεχνικής απομονώνουμε το μοντέλο που έδωσε τα βέλτιστα αποτελέσματα με Addmixup ($\alpha = 0.1$) καθώς και αυτό που έδωσε τα βέλτιστα αποτελέσματα με την τυπική μέθοδος Ελαχιστοποίησης Εμπειρικού Ρίσκου. Με βάση αυτά κατασκευάζουμε τον Πίνακα 5.16, όπου παραθέτουμε διάφορες μετρικές αξιολόγησης για κάθε κλάση του συνόλου ελέγχου για τα δύο αυτά μοντέλα. Με bold απεικονίζονται οι μετρικές που προκύπτουν με την Addmixup, οι οποίες είναι μεγαλύτερες από τις αντίστοιχες της τεχνικής ERM.

Class	Precision		Recall		F1-score		Support
	ERM	Addmixup	ERM	Addmixup	ERM	Addmixup	
surprised	83,49	↑ **85,14**	79.94	↑ **83,59**	81,68	↑ **84,36**	329
fearful	69,09	↑ **78,85**	51,35	↑ **55,41**	58,91	↑ **65,08**	74
disgusted	62,60	↑ **65,89**	51,25	↑ **53,12**	56,36	↑ **58,82**	160
happy	92,33	↑ **92,73**	93,50	↓ 92, 57	92,91	↓ 92, 65	1185
sad	81,01	↑ **83,37**	80,33	↑ **82,85**	80,67	↑ **83,11**	478
angry	79,47	↓ 75, 08	74,07	↓ 71, 60	76,68	↓ 75, 08	162
neutral	78,30	↓ 77, 73	85,44	↑ **86,76**	81,72	↑ **82,00**	680
Accuracy					83,96	↑ **84,75**	3068
Macro average	78,04	↑ **80,37**	73,70	↑ **75,13**	75,56	↑ **77,30**	3068
Weighted average	83,72	↑ **84,67**	83,96	↑ **84,75**	83,74	↑ **84,56**	3068

Πίνακας 5.16: Σύγκριση Addmixup και ERM σε όλες τις μετρικές για κάθε κλάση του dataset.

Όπως μπορούμε να παρατηρήσουμε, όταν εφαρμόζεται η Addmixup, με εξαίρεση ελάχιστες περιπτώσεις, σημειώνεται άνοδος σε όλες τις μετρικές αξιολόγησης τόσο ξεχωριστά για κάθε κλάση όσο και συνολικά. Μάλιστα αξίζει να αναφέρουμε πως διαπιστώνεται μια ιδιαίτερα σημαντική αύξηση στην κλάση με τα λιγότερα δείγματα, την "fearful". Συγκεκριμένα, για την κατηγορία αυτή, η μετρική Precision παρουσιάζει άνοδο σχεδόν 10%, η Recall περίπου

4% ενώ η F1-score πάνω από 6%. Το γεγονός αυτό είναι ιδιαίτερα ενθαρρυντικό για την αποτελεσματικότητα της τεχνικής σε προβλήματα κατηγοριοποίησης που έχουν μικρό αριθμό δειγμάτων σε ορισμένες κατηγορίες.

Είδαμε λοιπόν πως η Addmixup οδηγεί σε σημαντική βελτίωση των μετρικών αξιολόγησης του μοντέλου μας, συγκριτικά με τη μέθοδο Ελαχιστοποίησης Εμπειρικού Ρίσκου. Στο σημείο αυτό έχει ενδιαφέρον να εξάγουμε και ορισμένα επιπλέον στατιστικά για την βεβαιότητα με την οποία τα δίκτυα κατατάσσουν τις test εικόνες σε κλάσεις. Αξιοποιούμε πάλι τόσο το μοντέλο που εκπαιδεύτηκε με την κλασική ERM μέθοδο καθώς και το βέλτιστο που εκπαιδεύτηκε με την Addmixup. Στους Πίνακες 5.17 και 5.18 παραθέτουμε τη μέση και ενδιάμεση τιμή για τη βεβαιότητα με την οποία τα δύο αυτά δίκτυα κατηγοριοποιούν δείγματα ορθά αλλά και λανθασμένα αντίστοιχα. Παρατηρούμε πως με χρήση της τεχνικής Addmixup, το δίκτυο λαμβάνει σωστές προβλέψεις με μεγαλύτερη βεβαιότητα ενώ αντίστοιχα λαμβάνει λανθασμένες προβλέψεις με μικρότερη βεβαιότητα, γεγονός επιθυμητό.

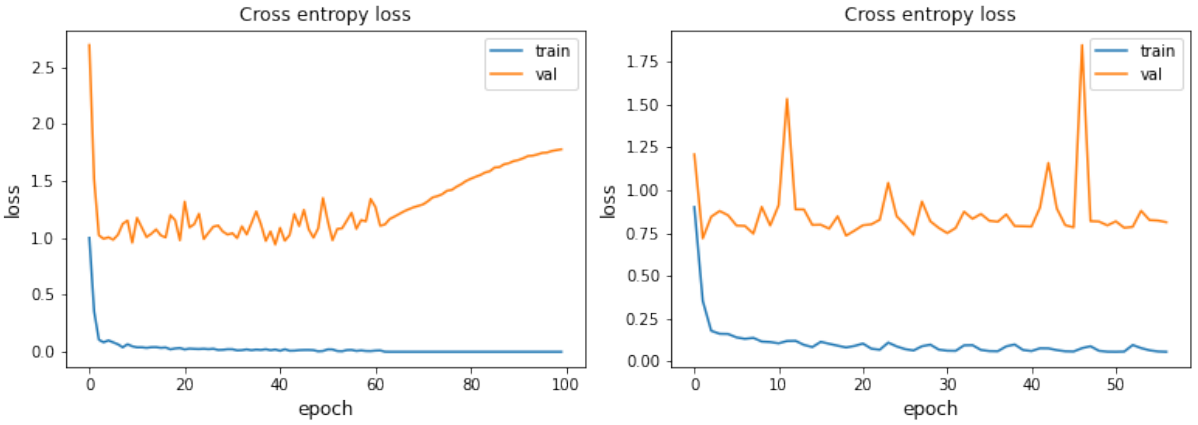
Confidence for correct predictions		
Models	mean	median
ERM	96,37	98,82
Addmixup	**98,77**	**100,0**

Πίνακας 5.17: Βεβαιότητα κατηγοριοποίησης δικτύων για τις σωστές προβλέψεις τους

Confidence for wrong predictions		
Models	mean	median
ERM	92,66	99,69
Addmixup	**84,24**	**90,75**

Πίνακας 5.18: Βεβαιότητα κατηγοριοποίησης δικτύων για τις λάθος προβλέψεις τους

Τέλος, εξετάζουμε κατά πόσον η εκπαίδευση με τη μέθοδο Addmixup μειώνει το σφάλμα της συνάρτησης απώλειας για τα δεδομένα ελέγχου. Στο Σχήμα 5.6 απεικονίζεται η συνάρτηση κατηγορικής διασταυρούμενης εντροπίας για τα δύο μοντέλα που μελετήσαμε και προηγουμένως (εκπαίδευση με ERM και με Addmixup). Παρατηρούμε πως η επαύξηση δεδομένων με Addmixup οδηγεί σε μικρότερες τιμές σφάλματος ενώ παράλληλα συμβάλλει και σε ταχύτερη σύγκληση του δικτύου, αφού, όπως φαίνεται στην περίπτωση αυτή, η εκπαίδευση διακόπτεται λίγο μετά τις 50 εποχές (πραγματοποιείται αυτόματα Early Stopping σε περίπτωση που δεν υπάρχει βελτίωση στην κατηγοριοποίηση του δικτύου για πάνω από 20 εποχές.)



(α′) Συνάρτηση απώλειας με ERM

(β′) Συνάρτηση απώλειας με Addmixup

Σχήμα 5.6: Απεικόνιση της συνάρτησης απώλειας για εκπαίδευσης με ERM και με Addmixup

Κεφάλαιο 6

Συμπεράσματα και μελλοντικές επεκτάσεις

6.1 Ανακεφαλαίωση και συμπεράσματα

Στα πλαίσια της παρούσας διπλωματικής εργασίας, εξετάζουμε την αποτελεσματικότητα της τεχνικής mixup στο πρόβλημα της συναισθηματικής αναγνώρισης σε πραγματικές, μη-ελεγχόμενες συνθήκες (in-the-wild). Συγκεκριμένα, εκπαιδεύουμε δίκτυα Βαθιάς Μάθησης με την εν λόγω τεχνική για κατηγοριοποίηση εικόνων εκφράσεων προσώπου στα 7 βασικά συναισθήματα. Παράλληλα, προτείνουμε και μια παραλλαγή της mixup, την Addmixup, με βάση την οποία το δίκτυο εκπαιδεύεται ταυτόχρονα πάνω σε εικονικά και πραγματικά παραδείγματα. Συγκρίνουμε τις δύο αυτές μεθόδους με την κλασική Ελαχιστοποίηση Εμπειρικού Ρίσκου ενώ παράλληλα εξετάζουμε και την επίδραση του dropout, μιας μορφής κανονικοποίησης του δικτύου, σε όλες τις προαναφερθείσες τεχνικές. Το σύνολο των πειραμάτων πραγματοποιείται πάνω σε δύο διαφορετικές εκδόσεις των στατικών in-the-wild βάσεων AffectNet και RAF-DB.

Όπως προκύπτει, από τα αποτελέσματα της πειραματικής μας μελέτης, για συγκεκριμένες παραμετροποιήσεις και σε συνδυασμό με την ύπαρξη dropout, η mixup συνεισφέρει θετικά σε ορισμένες περιπτώσεις, σε σύγκριση με την κλασική μέθοδο ERM. Ωστόσο, αισθητά μεγαλύτερη βελτίωση διαπιστώνεται όταν εφαρμόζεται η τεχνική Addmixup. Στην περίπτωση αυτή το εκπαιδευμένο μοντέλο εμφανίζει βελτιωμένη ικανότητα γενίκευσης, μεγαλύτερη βεβαιότητα ως προς τις προβλέψεις του και μικρότερο σφάλμα πάνω στα δεδομένα ελέγχου.

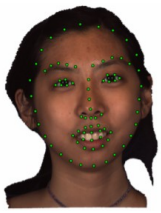
6.2 Μελλοντικές επεκτάσεις

Τα αποτελέσματα που λαμβάνουμε κατά τη διεξαγωγή της πειραματικής μας μελέτης είναι ιδιαίτερα θετικά. Υπάρχουν ωστόσο αρκετές ακόμα επεκτάσεις που μπορούν να υλοποιηθούν για την εξαγωγή περισσότερων χρήσιμων συμπερασμάτων γύρω από το πρόβλημα της συναισθηματικής αναγνώρισης και την εφαρμογή της τεχνικής mixup. Αυτές είναι οι εξής:

- **3D ευθυγράμμιση προσώπων**: Για την ευθυγράμμιση κάθε εικόνας της AffectNet πραγματοποιείται ένας αφινικός μετασχηματισμός μεταξύ συγκεκριμένων σημείων ενδιαφέροντος του προσώπου και αντίστοιχων σημείων αναφοράς. Πρόκειται για μια γρήγορη και αρκετά αποτελεσματική μέθοδο, η οποία λειτουργεί αρκετά ικανοποιητικά σε όλες τις controlled βάσεις δεδομένων. Ωστόσο, όπως αναφέραμε στην Ενότητα 5.3.3, επειδή η AffectNet είναι μια in-the-wild βάση, περιέχει έναν μεγάλο αριθμό

εικόνων που απεικονίζουν πρόσωπα που δεν είναι στραμμένα ευθεία προς το φακό. Σε τέτοιες περιπτώσεις, η μέθοδος ευθυγράμμισης που υλοποιούμε δεν είναι ιδιαίτερα αποτελεσματική καθώς, και μετά το πέρας της διαδικασίας, τα πρόσωπα διατηρούν στραμμένο το βλέμμα τους προς την ίδια αρχική κατεύθυνση και όχι προς τα εμπρός. Για την αντιμετώπιση του συγκεκριμένου προβλήματος μπορεί να αξιοποιηθεί κάποια πιο σύνθετη τεχνική 3D ευθυγράμμισης. Για παράδειγμα, στο [42] παρουσιάζεται μια μέθοδος, κατά την οποία από κάθε δισδιάστατη εικόνα προσώπου εξάγεται ένα 3D πλέγμα. Στη συνέχεια, πάνω στο πλέγμα αυτό πραγματοποιείται μια ανίχνευση 3D landmarks, τα οποία αξιοποιούνται για την κατάλληλη περιστροφή της πόζας του κεφαλιού, όπως φαίνεται στο Σχήμα 6.1. Αν το σύνολο των εικόνων των βάσεων που εξετάζουμε ευθυγραμμιστεί μέσω μιας τέτοιας διαδικασίας, εκτιμάται πως τόσο η τεχνική mixup όσο και η Addmixup θα λειτουργήσουν πολύ πιο αποτελεσματικά.

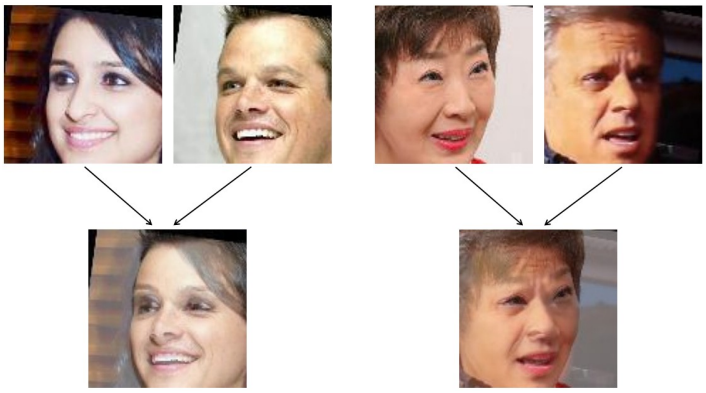
(α′) Εικόνα προσώπου με βλέμμα στραμμένο προς κάποια κατεύθυνση



(β′) Ευθυγραμμισμένη εικόνα προσώπου με βλέμμα στραμμένο προς τα μπροστά

Σχήμα 6.1: 3D ευθυγράμμιση προσώπου αξιοποιώντας 2D facial landmarks [42]

- **Χωρισμός του dataset σε υποκατηγορίες ανάλογα με την πόζα του κεφαλιού**: Σε περίπτωση που δεν εφαρμοστεί κάποια πιο εξεζητημένη μέθοδος ευθυγράμμισης μπορεί, σαν τεχνική προεπεξεργασίας, να πραγματοποιηθεί μία διάκριση του συνόλου εκπαίδευσης σε υποκατηγορίες, ανάλογα με την πόζα του κεφαλιού. Για το σκοπό αυτό μπορεί να αξιοποιηθεί ένα προεκπαιδευμένο DNN το οποίο θα ταξινομεί όλες τις εικόνες της βάσης σε subsets, με βάση τη γωνία στροφής του προσώπου. Στη συνέχεια, η διαδικασία ανάμειξης των δειγμάτων θα πραγματοποιείται εντός κάθε υποσυνόλου, οπότε και ο τυχαίος συνδυασμός εικόνων που θα δίνεται στο δίκτυο για εκπαίδευση θα είναι πολύ πιο αποτελεσματικός. Δύο χαρακτηριστικά παραδείγματα ανάμειξης προσώπων που είναι στραμμένα κατά την ίδια κατεύθυνση και γωνία φαίνονται στην Εικόνα 6.2. Όπως παρατηρούμε η παρεμβολή των εικόνων είναι αρκετά επιτυχημένη καθώς συνδυάζονται μεταξύ τους τα αντίστοιχα χαρακτηριστικά κάθε προσώπου (μάτια,μύτη, στόμα κτλ).



Σχήμα 6.2: Παραδείγματα ανάμειξης προσώπων που είναι στραμμένα κατά την ίδια γωνία

- **Εκπαίδευση για μεγαλύτερο αριθμό εποχών**: Λόγω του μεγάλου μεγέθους της βάσης AffectNet, στα πλαίσια της παρούσας διπλωματικής η μελέτη μας περιορίστηκε σε έναν σχετικά μικρό αριθμό εποχών. Μελλοντικός μας στόχος είναι η εκπαίδευση με τη συγκεκριμένη βάση για αρκετά μεγαλύτερο αριθμό εποχών, έτσι ώστε να διαμορφωθεί μια περισσότερο ενδεικτική εικόνα της συνεισφοράς της τεχνικής mixup.

- **Εκπαίδευση βαθύτερων αρχιτεκτονικών με Addmixup**: Σε όλα τα πειράματα που εφαρμόστηκε η τεχνική Addmixup αξιοποιήθηκε ένα προεκπαιδευμένο δίκτυο ResNet-50 με αλλαγμένη κεφαλή ταξινόμησης. Στο [117] οι συγγραφείς εφαρμόζουν την τεχνική mixup και σε πιο μεγάλα Συνελικτικά Δίκτυα, όπως ResNet-101 και ResNeXt-101, και μάλιστα εκεί παρατηρούν σημαντικότερη βελτίωση. Θα ήταν ενδιαφέρον να εξετάσουμε κατά πόσον η τεχνική Addmixup μπορεί να συνδράμει στην βελτίωση της γενίκευσης και άλλων, πιο βαθιών, Νευρωνικών Δικτύων.

- **Εφαρμογή της Addmixup σε περισσότερες βάσεις δεδομένων**: Τα ιδιαίτερα θετικά αποτελέσματα της εφαρμογής της Addmixup στην RAF-DB αποδεικνύουν πως πρόκειται για μια πολύ αποτελεσματική τεχνική που μπορεί να συνεισφέρει σημαντικά στο πρόβλημα της συναισθηματική αναγνώρισης. Σύντομα πρόκειται να εξετάσουμε τη συγκεκριμένη μέθοδο τόσο στη βάση AffectNet όσο και σε άλλα in-the-wild σύνολα εικόνων και βίντεο, με στόχο την εξαγωγή ορισμένων καθολικών συμπερασμάτων.

Βιβλιογραφία

[1] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, pages 1–4. IEEE, 2010. 36

[2] Nancy Alvarado. Arousal and valence in the direct scaling of emotional response to film clips. *Motivation and Emotion*, 21(4):323–348, 1997. 19

[3] Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Claudio Vairo. A comparison of face verification with facial landmarks and deep features. In *10th International Conference on Advances in Multimedia (MMEDIA)*, pages 1–6, 2018. 12, 56

[4] Panagiotis Antoniadis, Ioannis Pikoulis, Panagiotis P Filntisis, and Petros Maragos. An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3645–3651, 2021. 10, 29

[5] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017. 30

[6] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–6. IEEE, 2015. 18

[7] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. The omg-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018. 11, 36, 40

[8] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, and Aleix M Martinez. Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv:1703.01210*, 2017. 11, 18, 40

[9] Simone Bianco, Luigi Celona, Gianluigi Ciocca, Davide Marelli, Paolo Napoletano, Stefano Yu, and Raimondo Schettini. A smart mirror for emotion monitoring in home environments. *Sensors*, 21(22):7453, 2021. 10, 19

[10] Sanjay Bilakhia, Stavros Petridis, Anton Nijholt, and Maja Pantic. The mahnob mimicry database: A database of naturalistic human interactions. *Pattern recognition letters*, 66:52–61, 2015. 36

[11] Sven Buechel and Udo Hahn. Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. *arXiv preprint arXiv:1806.08890*, 2018. 20

[12] Pedro Cano, Markus Koppenberger, and Nicolas Wack. Content-based music audio recommendation. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 211–212, 2005. 34

[13] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 20

[14] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, pages 416–422, 2001. 45

[15] Chen Chen, Zuxuan Wu, and Yu-Gang Jiang. Emotion in context: Deep semantic feature fusion for video emotion recognition. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 127–131, 2016. 28

[16] Jiaao Chen, Zichao Yang, and Diyi Yang. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *arXiv preprint arXiv:2004.12239*, 2020. 11, 33

[17] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 58

[18] Weihong Deng, Jiani Hu, Shuo Zhang, and Jun Guo. Deepemo: Real-world facial expression analysis via deep learning. In *2015 Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2015. 25

[19] Terrance DeVries and Graham W Taylor. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017. 34

[20] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. From individual to group-level emotion recognition: Emotiw 5.0. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 524–528, 2017. 11, 36, 39

[21] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014. 10, 17, 18, 20

[22] Zach Eaton-Rosen, Felix Bragman, Sebastien Ourselin, and M Jorge Cardoso. Improving data augmentation for medical image segmentation. 2018. 10, 30, 31

[23] Paul Ekman. Facial action coding system (facs). *A human face*, 2002. 16, 20

[24] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971. 16

[25] Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997. 20

[26] Itir Onal Ertugrul, Jeffrey F Cohn, László A Jeni, Zheng Zhang, Lijun Yin, and Qiang Ji. Crossing domains for au coding: Perspectives, approaches, and measures. *IEEE transactions on biometrics, behavior, and identity science*, 2(2):158–171, 2020. 11, 38

[27] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 50

[28] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016. 18

[29] Peter Foster, Siddharth Sigtia, Sacha Krstulovic, Jon Barker, and Mark D Plumbley. Chime-home: A dataset for sound source recognition in a domestic environment. In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE, 2015. 34

[30] Jeffrey M Girard, Wen-Sheng Chu, László A Jeni, and Jeffrey F Cohn. Sayette group formation task (gft) spontaneous facial expression database. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 581–588. IEEE, 2017. 38

[31] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 21

[32] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 21

[33] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer, 2013. 11, 21, 30, 36, 40

[34] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010. 36

[35] Hongyu Guo. Nonlinear mixup: Out-of-manifold data augmentation for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4044–4051, 2020. 11, 32

[36] Hongyu Guo, Yongyi Mao, and Richong Zhang. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*, 2019. 10, 31

[37] Stephan Hamann. Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in cognitive sciences*, 16(9):458–466, 2012. 20

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 11, 30, 51

[39] K Heilman, RD Lane, and L Nadel. Cognitive neuroscience of emotion, 2000. 19

[40] Junlin Hou, Jilan Xu, Rui Feng, Yuejie Zhang, Fei Shan, and Weiya Shi. Cmc-cov19d: Contrastive mixup classification for covid-19 diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 454–461, 2021. 30

[41] Hiroshi Inoue. Data augmentation by pairing samples for images classification. *arXiv preprint arXiv:1801.02929*, 2018. 34

[42] László A Jeni, Jeffrey F Cohn, and Takeo Kanade. Dense 3d face alignment from 2d videos in real-time. In *2015 11th IEEE international conference and workshops*

on automatic face and gesture recognition (FG), volume 1, pages 1–8. IEEE, 2015. 12, 71

[43] Yue Jin, Tianqing Zheng, Chao Gao, and Guoqiang Xu. A multi-modal and multi-task learning method for action unit and expression recognition. *arXiv preprint arXiv:2107.04187*, 2021. 10, 27, 28

[44] Pooya Khorrami, Thomas Paine, and Thomas Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE international conference on computer vision workshops*, pages 19–27, 2015. 20

[45] Dimitrios Kollias, Anastasios Arsenos, Levon Soukissian, and Stefanos Kollias. Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis. *arXiv preprint arXiv:2106.07524*, 2021. 30

[46] Dimitrios Kollias, Shiyang Cheng, Maja Pantic, and Stefanos Zafeiriou. Photorealistic facial synthesis in the dimensional affect space. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 21

[47] Dimitrios Kollias, Shiyang Cheng, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, 128(5):1455–1484, 2020. 21, 30

[48] Dimitrios Kollias, Mihalis A Nicolaou, Irene Kotsia, Guoying Zhao, and Stefanos Zafeiriou. Recognition of affect in the wild using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 26–33, 2017. 11, 41

[49] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 637–643. IEEE, 2020. 11, 18, 26, 36, 41

[50] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 18

[51] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 20

[52] Dimitrios Kollias, Athanasios Tagaris, Andreas Stafylopatis, Stefanos Kollias, and Georgios Tagaris. Deep neural architectures for prediction in healthcare. *Complex & Intelligent Systems*, 4(2):119–131, 2018. 30

[53] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6):907–929, 2019. 10, 19

[54] Dimitris Kollias, Y Vlaxos, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate, and S Kollias. Transparent adaptation in deep medical image diagnosis. In *International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning*, pages 251–267. Springer, 2020. 30

[55] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018. 11, 18, 26, 36, 41

[56] Dimitrios Kollias and Stefanos Zafeiriou. A multi-task learning & generation framework: Valence-arousal, action units & primary expressions. *arXiv preprint arXiv:1811.07771*, 2018. 11, 18, 26, 36, 41

[57] Dimitrios Kollias and Stefanos Zafeiriou. Training deep neural networks with different datasets in-the-wild: The emotion recognition paradigm. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018. 26

[58] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 10, 11, 18, 20, 21, 26, 36, 41

[59] Dimitrios Kollias and Stefanos Zafeiriou. Va-stargan: Continuous affect generation. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 227–238. Springer, 2020. 21

[60] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 26, 41

[61] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 27, 29, 41

[62] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1667–1675, 2017. 28

[63] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 30, 47

[64] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research). 30

[65] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 26

[66] Felix Kuhnke, Lars Rumberg, and Jörn Ostermann. Two-stream aural-visual affect analysis in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 600–605. IEEE, 2020. 10, 26, 27

[67] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, Skyler T Hawk, and AD Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010. 10, 17

[68] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10143–10152, 2019. 10, 28

[69] Penelope A Lewis, Hugo D Critchley, Pia Rotshtein, and Raymond J Dolan. Neural correlates of processing valence and arousal in affective words. *Cerebral cortex*, 17(3):742–748, 2007. 19

[70] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 2584–2593. IEEE, 2017. 23, 26, 30, 36, 43

[71] Chengjun Liu and Harry Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image processing*, 11(4):467–476, 2002. 26

[72] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010. 30, 36

[73] Michael Lyons, Miyuki Kamachi, and Jiro Gyoba. Japanese female facial expression (jaffe) database. 2017. 36

[74] Leandro Y Mano, Bruno S Faiçal, Vinícius P Gonçalves, Gustavo Pessin, Pedro H Gomes, André CPLF de Carvalho, and Jó Ueyama. An intelligent and generic approach for detecting human emotions: a case study with facial expressions. *Soft Computing*, 24(11):8467–8479, 2020. 10, 16

[75] David Matsumoto. More evidence for the universality of a contempt expression. *Motivation and Emotion*, 16(4):363–368, 1992. 19

[76] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 11, 26, 36, 37, 42

[77] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17, 2011. 36

[78] Marc Mehu and Klaus R Scherer. Emotion categories and dimensions in the facial communication of affect: An integrated approach. *Emotion*, 15(6):798, 2015. 20

[79] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 30

[80] Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. Dcase 2017 challenge setup: Tasks, datasets and baseline system. In *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017. 35

[81] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Tut database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1128–1132. IEEE, 2016. 34

[82] Ali Mollahosseini, David Chan, and Mohammad H Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016. 26

[83] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 10, 22, 23, 36, 42

[84] Ali Mollahosseini, Behzad Hasani, Michelle J Salvador, Hojjat Abdollahi, David Chan, and Mohammad H Mahoor. Facial expression recognition from world wild web. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 58–65, 2016. 26

[85] Harikrishna Narasimhan, Weiwei Pan, Purushottam Kar, Pavlos Protopapas, and Harish G Ramaswamy. Optimizing the multiclass f-measure via biconcave programming. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 1101–1106. IEEE, 2016. 55

[86] Armando M Oliveira, Marta P Teixeira, Isabel B Fonseca, and Miguel Oliveira. Joint model-parameter validation of self-estimates of valence and arousal: Probing a differential-weighting model of affective intensity. *Proceedings of Fechner Day*, 22:245–250, 2006. 19

[87] Tianyu Pang, Kun Xu, and Jun Zhu. Mixup inference: Better exploiting mixup to defend adversarial attacks. *arXiv preprint arXiv:1909.11515*, 2019. 22

[88] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pages 5–pp. IEEE, 2005. 36

[89] Karol J Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE, 2015. 34

[90] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014. 55

[91] Muhammad Naveed Riaz, Yao Shen, Muhammad Sohail, and Minyi Guo. Exnet: An efficient approach for emotion recognition in the wild. *Sensors*, 20(4):1087, 2020. 30, 49

[92] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013. 11, 36, 38

[93] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 30, 50, 51

[94] James A Russell. Evidence of convergent validity on the dimensions of affect. *Journal of personality and social psychology*, 36(10):1152, 1978. 18

[95] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 18

[96] Shankar Setty, Moula Husain, Parisa Beham, Jyothi Gudavalli, Menaka Kandasamy, Radhesyam Vaddi, Vidyagouri Hemadri, JC Karure, Raja Raju, B Rajan, et al. Indian movie face database: a benchmark for face recognition under wide variations. In *2013 fourth national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG)*, pages 1–5. IEEE, 2013. 11, 39

[97] Patrice Y Simard, Yann A LeCun, John S Denker, and Bernard Victorri. Transformation invariance in pattern recognition—tangent distance and tangent propagation. In *Neural networks: tricks of the trade*, pages 239–274. Springer, 1998. 45

[98] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 11, 52

[99] Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip S Yu, and Lifang He. Mixup-transformer: Dynamic data augmentation for nlp tasks. *arXiv preprint arXiv:2010.02394*, 2020. 11, 33

[100] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013. 55

[101] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 45

[102] Luke Taylor and Geoff Nitschke. Improving deep learning with generic data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1542–1547. IEEE, 2018. 10, 21

[103] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001. 36

[104] Michel Valstar, Maja Pantic, et al. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65. Paris, France., 2010. 36

[105] Michel F. Valstar, Timur Almaev, Jeffrey M. Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F. Cohn. Fera 2015 - second facial expression recognition and analysis challenge. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 06, pages 1–8, 2015. 37

[106] Michel F Valstar, Enrique Sánchez-Lozano, Jeffrey F Cohn, László A Jeni, Jeffrey M Girard, Zheng Zhang, Lijun Yin, and Maja Pantic. Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 839–847. IEEE, 2017. 38

[107] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999. 45

[108] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018. 33

[109] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In *International MICCAI brainlesion workshop*, pages 178–190. Springer, 2017. 30

[110] Shengyun Wei, Kele Xu, Dezhi Wang, Feifan Liao, Huaimin Wang, and Qiuqiang Kong. Sample mixed-based data augmentation for domestic audio tagging. *arXiv preprint arXiv:1808.03883*, 2018. 11, 34

[111] Cynthia M Whissel. The dictionary of affect in language r. plutchnik and h. kellerman (eds) emotion: Theory, research and experience: vol 4, the measurement of emotions, 1989. 18

[112] Kele Xu, Dawei Feng, Haibo Mi, Boqing Zhu, Dezhi Wang, Lilun Zhang, Hengxing Cai, and Shuwen Liu. Mixup-based acoustic scene classification using multi-channel convolutional neural network. In *Pacific Rim conference on multimedia*, pages 14–23. Springer, 2018. 11, 35

[113] L Yin, X Chen, Y Sun, T Worm, and M Reale. 3d dynamic facial expression database, ieee inter. In *Conf. on Automatic Face and Gesture Recognition, Amsterdam, the Netherlands*, 2008. 36

[114] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006. 36

[115] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–41, 2017. 11, 41

[116] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 45

[117] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 11, 22, 29, 30, 32, 46, 47, 60, 72

[118] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 11, 36, 37, 43

[119] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3438–3446, 2016. 38