



Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και
Υπολογιστών

**Πολυτροπική Προσέγγιση Αναγνώρισης
Συναισθήματος Βασισμένη σε Δίκτυα Βαθιάς
Μάθησης**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΣΤΥΛΙΑΝΗ ΤΕΡΖΑΚΗ ΠΑΠΑΔΟΠΟΥΛΟΥ

Επιβλέπων:

Στέφανος Κόλλιας

Καθηγητής Ε.Μ.Π.

Συμμετοχή στην Επίβλεψη:

Τζούβελη Παρασκευή

μέλος ΕΔΙΠ

Αθήνα, Ιούλιος 2021



Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και
Υπολογιστών

**Πολυτροπική Προσέγγιση Αναγνώρισης
Συναισθήματος Βασισμένη σε Δίκτυα Βαθιάς
Μάθησης**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΣΤΥΛΙΑΝΗ ΤΕΡΖΑΚΗ ΠΑΠΑΔΟΠΟΥΛΟΥ

Επιβλέπων:

Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Συμμετοχή στην Επίβλεψη:

Τζούβελη Παρασκευή
μέλος ΕΔΙΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 14η Ιουλίου 2021.

Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Γιώργος Στάμου
Αν. Καθηγητής Ε.Μ.Π.

Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2021

Στυλιανή Τερζάκη Παπαδοπούλου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © Στυλιανή Τερζάκη Παπαδοπούλου, 2021.

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

Η παρούσα διπλωματική εργασία πραγματεύεται το θέμα της αυτόματης Αναγνώρισης Συναισθήματος χρησιμοποιώντας μεθόδους Βαθιάς Μάθησης που εξάγουν πληροφορία από οπτικά και ακουστικά δεδομένα. Το αντικείμενο της ανάλυσης της συναισθηματικής συμπεριφοράς έχει εφαρμογή στην αλληλεπίδραση ανθρώπου-υπολογιστή, στην αυτόματη παρακολούθηση ασθενών, στην ασφάλεια, κ.α.

Η προσέγγιση του προβλήματος αναγνώρισης συναισθημάτων από οπτικοακουστικά δεδομένα γίνεται με την χρήση σύγχρονων Συνελικτικών Νευρωνικών Δικτύων και την εκμετάλλευση της μεθόδου της μεταφοράς μάθησης από προεκπαιδευμένα μοντέλα. Συγκεκριμένα, στη παρούσα εργασία καλούμαστε να επιλύσουμε δύο είδη προβλήματων Επιβλεπόμενης Μάθησης, ένα πρόβλημα ταξινόμησης και ένα πρόβλημα παλινδρόμησης. Τα δύο αυτά προβλήματα προκύπτουν από τα είδη ετικετών που περιέχει το σύνολο δεδομένων, το οποίο χρησιμοποιείται για την εκπαίδευση των μοντέλων.

Το σύνολο αυτό ονομάζεται Aff-Wild2 και απαρτίζεται από μία συλλογή βίντεο που έχουν δημιουργηθεί σε πραγματικές συνθήκες ή όπως αναφέρεται χαρακτηριστικά "in-the-wild". Το συγκεκριμένο σύνολο έχει χαρακτηρισθεί με ετικέτες που βασίζονται σε δύο διαφορετικά μοντέλα αναπαράστασης συναισθημάτων. Το πρώτο μοντέλο είναι κατηγορικό και χωρίζει το πλήθος των συναισθημάτων σε 7 βασικές κατηγορίες, οι οποίες είναι ο Φόβος, ο Θυμός, η Αηδία, η Χαρά, η Λύπη, η Έκπληξη και η Ουδετερότητα. Συνεπώς, οι ετικέτες αυτές χρησιμοποιούνται για την ανάπτυξη μοντέλων ταξινόμησης πολλαπλών κλάσεων. Το δεύτερο μοντέλο είναι διαστατικό και εντάσσει τα συναισθήματα σε ένα δισδιάστατο χώρο με άξονες το σθένος και τη διέγερση. Οι ετικέτες του σθένους και της διέγερσης παίρνουν συνεχόμενες τιμές στο διάστημα $[-1, 1]$ και η πρόβλεψή τους από τα μοντέλα μάθησης αποτελεί πρόβλημα παλινδρόμησης.

Για το κάθε πρόβλημα μάθησης αναπτύσσονται διαφορετικά μοντέλα, τα οποία δέχονται οπτικά ή ακουστικά δεδομένα ή και τον συνδυασμό τους. Τα συνελικτικά νευρωνικά δίκτυα που χρησιμοποιούνται ως βάση για τα παραπάνω μοντέλα είναι τα δίκτυα VGG, τα Διαφορικά δίκτυα (Residual Networks) και τα Πυκνά Συνδεδεμένα δίκτυα (Dense Networks). Μετά τη σχεδίαση των μοντέλων γίνεται η εκπαίδευση και η τελική αξιολόγησή τους στο συγκεκριμένο σύνολο δεδομένων.

Λέξεις κλειδιά

Αναγνώριση συναισθήματος, ανάλυση συναισθηματικής συμπεριφοράς, μηχανική μάθηση, νευρωνικά δίκτυα, βαθιά μάθηση, συνελικτικά νευρωνικά δίκτυα, ενσωμάτωση δεδομένων, ταξινόμηση, παλινδρόμηση

Abstract

This diploma thesis deals with the topic of automatic Emotion Recognition using Deep Learning methods that extract information from visual and audio data. The field of affective behavior analysis is applicable to human-computer interaction, automatic patient monitoring, security, etc.

The problem of recognizing emotions from audiovisual data is approached by using modern Convolutional Neural Networks and exploiting the method of transfer learning from pre-trained models. Specifically, in this paper we are called to solve two types of Supervised Learning problems, a classification problem and a regression problem. These two problems arise from the type of labels contained in the data set, which is used to train the models.

This dataset is called Aff-Wild2 and it consists of a collection of videos created in real-life conditions or as it is typically called "in-the-wild". The dataset has been annotated with labels based on two different emotion representation models. The first model is categorical and it divides the range of different emotions into 7 basic categories, which are Fear, Anger, Disgust, Happiness, Sadness, Surprise and Neutral. These labels are therefore used to develop multi-class classification models. The second model is dimensional and incorporates emotions into a two-dimensional space with axis of valence and arousal. Valence and arousal labels take continuous values in the range $[-1, 1]$ and their prediction by a learning model is a regression problem.

Different models are developed for each learning problem, which accept visual or audio data or their fusion. The convolutional neural networks used as the basis for the above models are the VGG networks, the Residual Networks and the Dense Networks. After the design of the models, their training and final evaluation in the specific dataset takes place.

Key words

Emotion recognition, affective behavior analysis, machine learning, neural networks, deep learning, convolutional neural networks, data fusion, classification, regression

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στο εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης της σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου. Πρωτίστως, θα ήθελα να ευχαριστήσω θερμά τον καθηγητή κ. Στέφανο Κόλλια, ο οποίος μου ανέθεσε το παρόν θέμα διπλωματικής εργασίας και μου έδειξε πλήρη εμπιστοσύνη για την επιτυχή ανάπτυξή του.

Επίσης, θα ήθελα να ευχαριστήσω από καφδιάς τη Δρ. Παρασκευή Τζούβελη, η οποία με χαρά και αφοσίωση καθοδήγησε κάθε μου βήμα και με βοήθησε να ξεπεράσω εύκολα οποιαδήποτε δυσκολία αντιμετώπισα στη διάρκεια εκπόνησης της παρούσας εργασίας. Ακόμη, θα ήθελα να ευχαριστήσω και τον Δρ. Δημήτρη Κόλλια για την παροχή των δεδομένων που χρησιμοποιήθηκαν στη παρούσα εργασία, αλλά και τη καθοδήγησή του όσο αναφορά τη προσέγγιση του θέματος.

Τέλος, ευχαριστώ την οικογένειά μου και τους φίλους μου για την στήριξη τους σε όλη τη διάρκεια των σπουδών μου αλλά και το γάτο μου που με τον τρόπο του αποτελούσε πολύτιμη συντροφιά στις πολλές ώρες διαβάσματος.

Στυλιανή Τερζάκη Παπαδοπούλου,

Αθήνα, 14η Ιουλίου 2021

Περιεχόμενα

Περίληψη	5
Abstract	7
Ευχαριστίες	9
Περιεχόμενα	11
Κατάλογος πινάκων	15
Κατάλογος σχημάτων	17
1. Εισαγωγή	19
1.1 Σκοπός της Εργασίας	19
1.2 Παρεμφερείς Εργασίες	19
1.3 Δομή της Εργασίας	20
2. Θεωρητικό Υπόβαθρο	23
2.1 Μηχανική Μάθηση	23
2.1.1 Επιβλεπόμενη Μάθηση	23
2.1.2 Μη-Επιβλεπόμενη Μάθηση	24
2.1.3 Ενισχυτική Μάθηση	25
2.1.4 Μεταφορική Μάθηση	25
2.2 Τεχνητά Νευρωνικά Δίκτυα (ANN)	26
2.2.1 Τρόπος Λειτουργίας	26
2.2.2 Βαθιά Νευρωνικά Δίκτυα	27
2.3 Συνελικτικά Νευρωνικά Δίκτυα (CNN)	28
2.3.1 Συνελικτικό Επίπεδο (Convolutional Layer)	28
2.3.2 Επίπεδο Ενεργοποίησης (Activation Layer)	29
2.3.3 Επίπεδο Υποδειγματοληψίας (Pooling Layer)	30
2.3.4 Πλήρως Συνδεδεμένο Επίπεδο (Fully Connected Layer)	31
2.3.5 Επίπεδο Ενεργοποίησης Softmax	31
2.3.6 Επίπεδο Κανονικοποίησης Παρτίδας (Batch Normalization Layer)	32
2.3.7 Επίπεδο Dropout	32
2.4 Σύγχρονα Συνελικτικά Νευρωνικά Δίκτυα	33
2.4.1 AlexNet	33

2.4.2	Δίκτυο VGG	33
2.4.3	Διαφορικά Δίκτυα (Residual Networks)	34
2.4.4	Πυκνά Συνδεδεμένα Δίκτυα (DenseNet)	36
3.	Δεδομένα	39
3.1	Τρόποι Αναπαράστασης Συναυσθήματος	39
3.1.1	Κατηγορικό Μοντέλο	39
3.1.2	Διαστατικό Μοντέλο	40
3.2	Το σύνολο δεδομένων Aff-Wild2	40
3.2.1	Πρόβλημα Παλινδρόμισης Σθένους/Διέγερσης (VA Task)	42
3.2.2	Πρόβλημα Ταξινόμησης Βασικών Συναυσθημάτων (Expressions Task)	43
3.3	Προεπεξεργασία Δεδομένων	44
3.3.1	Προεπεξεργασία Οπτικών Δεδομένων	45
3.3.2	Προεπεξεργασία Ακουστικών Δεδομένων	45
3.3.2.1	Φασματογράφημα Mel (Mel Spectrogram)	46
3.3.2.2	Εξαγωγή Φασματογραφημάτων Mel	48
4.	Μεθοδολογία	53
4.1	Μετρικές Αξιολόγησης Μοντέλων	53
4.1.1	Μετρική Αξιολόγησης Προβλήματος Παλινδρόμησης	53
4.1.2	Μετρική Αξιολόγησης Προβλήματος Ταξινόμησης	54
4.2	Συναρτήσεις Κόστους (Loss Functions)	55
4.2.1	Συνάρτηση Κόστους Προβλήματος Παλινδρόμησης	55
4.2.2	Συνάρτηση Κόστους Προβλήματος Ταξινόμησης	56
4.3	Προτεινόμενα Μοντέλα Οπτικής Αναγνώρισης	56
4.3.1	VGG-16	58
4.3.2	ResNet-50	58
4.3.3	DenseNet-121	59
4.4	Προτεινόμενα Μοντέλα Ακουστικής Αναγνώρισης	61
4.4.1	VGG-11	61
4.4.2	ResNet-18	62
4.5	Συνδυαστικό Μοντέλο Οπτικοακουστικής Αναγνώρισης	63
4.5.1	Μέθοδοι Ενσωμάτωσης Δεδομένων	63
4.5.2	Συνδυαστικά Μοντέλα Δύο Κλάδων	64
5.	Εκπαίδευση και Αξιολόγηση των Μοντέλων	67
5.1	Εκπαίδευση των Μοντέλων	67
5.1.1	Τυπολογιστικά Συστήματα	67
5.1.2	Λογισμικό	68
5.1.3	Διαδικασία Εκπαίδευσης	68
5.2	Αξιολόγηση Μοντέλων	69
5.2.1	Αξιολόγηση Μοντέλων VA	69
5.2.2	Αξιολόγηση Μοντέλων Basic Expressions	71
5.3	Ανάλυση Αποτελεσμάτων	76

5.3.1	Απόδοση Μοντέλων VA	76
5.3.2	Απόδοση Μοντέλων Basic Expressions	77
6.	Επίλογος και Μελλοντικές Επεκτάσεις	79
6.1	Συμπεράσματα Διπλωματικής Εργασίας	79
6.2	Μελλοντικές Επεκτάσεις	80
	Βιβλιογραφία	83

Κατάλογος πινάκων

3.1	Στατιστικά Στοιχεία Aff-Wild2	43
4.1	Αρχιτεκτονική VGG-16	58
4.2	Αρχιτεκτονική ResNet-50	59
4.3	Αρχιτεκτονική DenseNet-121	60
4.4	Αρχιτεκτονική VGG-11	62
4.5	Αρχιτεκτονική ResNet-18	63
5.1	Αποτελέσματα Μοντέλων VA	71
5.2	Αποτελέσματα Μοντέλων Basic Expressions	73

Κατάλογος σχημάτων

2.1	Μεταφορική Μάθηση	25
2.2	Βαθύ Νευρωνικό Δίκτυο	27
2.3	Συνελικτικό Νευρωνικό Δίκτυο	28
2.4	Συνάρτηση Ενεργοποίησης ReLU	30
2.5	Τποδειγματοληψία Μέγιστης Τιμής (Max Pooling)	31
2.6	Αρχιτεκτονική AlexNet	33
2.7	Αρχιτεκτονική VGG	34
2.8	Διαφορικό Μπλοκ (Residual Block)	35
2.9	Αρχιτεκτονική του ResNet-34	36
2.10	Αρχιτεκτονική DenseNet	37
3.1	Βασικά Συναισθημάτα	40
3.2	Ο τροχός των συναισθημάτων	41
3.3	Δείγμα του Aff-Wild2	42
3.4	Δισδιάστατο Ιστόγραμμα των δεδομένων VA	43
3.5	Κατανομή Δεδομένων Basic Expressions	44
3.6	Παραδείγματα Οπτικών Δεδομένων του συνόλου VA	46
3.7	Παραδείγματα Οπτικών Δεδομένων του συνόλου Basic Expressions	47
3.8	Παράδειγμα Ηχητικού Σήματος και Φασματογραφήματος	47
3.9	Βραχυπρόθεσμος Μετασχηματισμός Fourier (STFT)	48
3.10	Παραδείγματα Φασματογραφημάτων του συνόλου VA	50
3.11	Παραδείγματα Φασματογραφημάτων του συνόλου Basic Expressions	51
4.1	Προτεινόμενο Οπτικό Μοντέλο VA	57
4.2	Προτεινόμενο Οπτικό Μοντέλο Basic Expressions	57
4.3	Μέθοδοι Ενσωμάτωσης Δεδομένων	64
4.4	Συνδυαστικό Μοντέλο VA	65
4.5	Συνδυαστικό Μοντέλο Basic Expressions	65
5.1	Καμπύλες Κόστους και CCC Οπτικού Μοντέλου VA	70
5.2	Καμπύλες Κόστους και CCC Ακουστικού Μοντέλου VA	70
5.3	Καμπύλες Κόστους και CCC Συνδυαστικού Μοντέλου VA	70
5.4	Παράδειγμα Προβλέψεων του Μοντέλου VA	72
5.5	Καμπύλες Κόστους και Κριτηρίου Οπτικού Μοντέλου Basic Expressions	72
5.6	Καμπύλες Κόστους και Κριτηρίου Ακουστικού Μοντέλου Basic Expressions	73

5.7	Καμπύλες Κόστους και Κριτηρίου Συνδυαστικού Μοντέλου Basic Expressions	73
5.8	Παραδείγματα Προβλέψεων του Μοντέλου Basic Expressions	74
5.9	Παραδείγματα Προβλέψεων του Μοντέλου Basic Expressions	75

Κεφάλαιο 1

Εισαγωγή

Τα ανθρώπινα συναισθήματα αποτελούν μυστήριο για το ίδιο το ανθρώπινο είδος, πόσο μάλλον για τους υπολογιστές. Ολόκληρη η επιστήμη της Ψυχολογίας έχει αναπτυχθεί για την ανάλυση και την κατανόηση των ανθρώπινων συναισθημάτων καθώς και τι τα προκαλεί. Συνέπως, η αναγνώρισή τους από ένα υπολογιστικό σύστημα ήταν αδιανόητη μέχρι πριν από κάποια χρόνια. Όμως, η Επιστήμη των Υπολογιστών εξελίσσεται με ραγδαίους ρυθμούς με αποτέλεσμα να αναπτύξει μεθόδους ώστε να καταστήσει την επίλυση του προβλήματος αυτού δυνατή.

1.1 Σκοπός της Εργασίας

Σκοπός της παρούσας Διπλωματικής Εργασίας είναι η προσέγγιση του θέματος της Αναγνώρισης Συναισθήματος από οπτικοακουστικές πηγές χρησιμοποιώντας τεχνικές της Μηχανικής Μάθησης. Η προσέγγιση αυτή περιλαμβάνει ανάπτυξη μοντέλων Βαθιάς Μάθησης, τα οποία βασίζονται σε σύγχρονα Συνελικτικά Νευρωνικά Δίκτυα. Μέρος της εργασίας αποτελεί η μελέτη της θεωρίας των Νευρωνικών Δικτύων αποσκοπώντας στη βαθύτερη κατανόησή της και έπειτα στην εφαρμογή της για το σχεδιασμό, την εκπαίδευση και την αξιολόγηση μοντέλων που επιλύουν το πρόβλημα της αναγνώρισης συναισθήματος.

Επιπρόσθετα, σημαντικό ρόλο για την τελική απόδοση των μοντέλων παίζει και η επιλογή και η προεπεξεργασία του συνόλου δεδομένων, το οποίο θα αποτελέσει είσοδο των μοντέλων μας.

1.2 Παρεμφερείς Εργασίες

Το αντικείμενο της Αναγνώρισης Συναισθήματος (Emotion Recognition) ή αλλιώς Αναγνώριση Συναισθηματικής Συμπεριφοράς (Affective Behavior Recognition) έχει απασχολήσει αρκετά την ερευνητική κοινότητα τα τελευταία 20 χρόνια.

Τα πρώτα βήματα της επιστήμης υπολογιστών που έγιναν προς αυτή την κατεύθυνση ήταν η ανάπτυξη μεθόδων ανίχνευσης προσώπων σε εικόνες και βίντεο [1, 2], ο εκσυγχρονισμός των μεθόδων ανάλυσης εικόνων και βίντεο [3, 4, 5, 6, 7] για εφαρμογές όπως η αναγνώριση ανθρώπινων ενεργειών (Action Recognition) και η ταξινόμηση βίντεο. Οι πρώτες δοκιμές για την αναγνώριση συναισθήματος από το πρόσωπο αναπτύχθηκαν κάνοντας χρήση ασαφών συστημάτων

(fuzzy systems) [8]. Επίσης, εκτός από το πρόβλημα της αναγνώρισης έγιναν και προσπάθειες σύνθεσης ανθρώπινων συναισθημάτων και εκφράσεων από τον υπολογιστή [9, 10].

Όσο αναφορά το πεδίο των νευρωνικών δικτύων, η έρευνα είχε ξεκινήσει από τις αρχές της δεκαετίας του 80 με χρήση της μεθόδου των ελαχίστων τετραγώνων [11]. Η περαιτέρω ανάπτυξη της θεωρίας των νευρωνικών δικτύων έγινε με την έναρξη του 21ου αιώνα αναπτύσσοντας διάφορες εφαρμογές που αναφέρονται στις εργασίες [12, 13, 14].

Μετά το 2014, η ανάπτυξη του τομέα της Μηχανικής Μάθησης ήταν ραγδαία και συγκεκριμένα της μάθησης μέσω νευρωνικών δικτύων. Ειδικότερα, η εξέλιξη των Συνελικτικών Νευρωνικών Δικτύων που κατέστησε ευκολότερη και πιο επιτυχή τη μάθηση μέσω δεδομένων εικόνας αλλά και την εκπαίδευση δικτύων με μεγάλο αριθμό επιπέδων ήταν ο λόγος που αναπτύχθηκε τόσο το αντικείμενο της αναγνώρισης συναισθημάτων. Η Βαθιά Μάθηση αποτέλεσε τη κύρια τεχνική ανάπτυξης μοντέλων αναγνώρισης συναισθημάτων όπως γίνεται αντιληπτό στις εργασίες [15, 16, 17]. Στο πλαίσιο της Βαθιάς Μάθησης προτείνονται και τεχνικές που συνδυάζουν Συνελικτικά (CNN) και Αναδρομικά (RNN) Νευρωνικά Δίκτυα για την αναγνώριση συναισθημάτων "in-the-wild", όπως γίνεται στις εργασίες [18, 19, 20]. Τέλος, οι πιο σύγχρονες έρευνες αφορούν τη σύνθεση προσώπων που εκφράζουν ανθρώπινα συναισθημάτα, τα οποία συμβάλλουν και στο πρόβλημα της αναγνώρισης συναισθημάτων [19, 21, 22].

Στο πλαίσιο της έρευνας της αναγνώρισης συναισθημάτων μέσω της μηχανικής μάθησης, κάθε χρόνο διοργανώνονται διαγωνισμοί που απευθύνονται σε ερευνητικές ομάδες αλλά και μεμονωμένους ερευνητές με σκοπό τη σύγκριση και την εξέλιξη των μοντέλων ανάλυσης συναισθημάτων. Ο πρώτος διεθνής διαγωνισμός αναγνώρισης συναισθημάτων ήταν το AVEC (Audio/Visual Emotion Challenge) το οποίο ξεκίνησε το 2011 [23] και διεξάγεται κάθε χρόνο. Γνωστός ετήσιος διαγωνισμός αποτελεί και το EmotiW (Emotion Recognition in the Wild Challenge), το οποίο ξεκίνησε το 2013 [24]. Το σύνολο δεδομένων Aff-Wild2 [25] που θα χρησιμοποιήσουμε για την εκπαίδευση των μοντέλων μας, αναπτύχθηκε και αυτό στο πλαίσιο ενός σύγχρονου διαγωνισμού αναγνώρισης συναισθηματικής συμπεριφοράς, το ABAW (Affective Behavior Analysis in-the-wild Challenge) [26, 27].

1.3 Δομή της Εργασίας

Η παρούσα εργασία πραγματεύεται την επίλυση του προβλήματος της αναγνώρισης ανθρώπινων συναισθημάτων μέσω οπτικοακουστικών μέσων, κάνοντας χρήση της θεωρίας της Μηχανικής Μάθησης. Συνεπώς, στο Κεφάλαιο 2 παρουσιάζουμε το θεωρητικό υπόβαθρο των πεδίων της Βαθιάς Μάθησης και των Νευρωνικών Δικτύων, το οποίο θα συμβάλει στη χαλύτερη κατανόηση της εργασίας. Επίσης γίνεται εκτενής παρουσίαση των Συνελικτικών Νευρωνικών Δικτύων, τα οποία αποτελούν τον πιο ευρέως γνωστό τρόπο επίλυσης προβλημάτων αναγνώρισης προτύπων από εικόνες.

Στο Κεφάλαιο 3 γίνεται η παρουσίαση των μοντέλων αναπαράστασης του ανθρώπινου συναισθήματος, και στη συνέχεια του συνόλου δεδομένων Aff-Wild2, το οποίο θα χρησιμοποιηθεί για την εκπαίδευση των μοντέλων αναγνώρισης που θα αναπτυχθούν στο πλαίσιο της εργασίας. Ακόμη περιγράφεται η διαδικασία προεπεξεργασίας που διενεργείται στο σύνολο δεδομένων ώστε να αποτελέσει είσοδο των μοντέλων. Έπειτα, στο Κεφάλαιο 4, αναλύεται η μεθοδολογία που ακο-

λουθήθηκε για την ανάπτυξη των μοντέλων αναγνώρισης συναισθήματος και γίνεται περιγραφή των Σύγχρονων Συνελικτικών Δικτύων που χρησιμοποιούνται για την εξαγωγή χαρακτηριστικών της εισόδου.

Στο Κεφάλαιο 5 αναφέρονται τα υπολογιστικά συστήματα και το λογισμικό που χρησιμοποιήθηκε για την εκπόνηση της παρούσας εργασίας. Ακολούθως γίνεται η περιγραφή της διαδικασίας εκπαίδευσης και αξιολόγησης των μοντέλων μας και η παρουσίαση και η ανάλυση των τελικών αποτελεσμάτων. Τέλος, στο Κεφάλαιο 6, παρουσιάζονται τα τελικά συμπεράσματα της διπλωματικής εργασίας και γίνονται προτάσεις για μελλοντικές επεκτάσεις.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Μηχανική Μάθηση

Η Μηχανική Μάθηση (Machine Learning-ML) αποτελεί υποκλάδο της Τεχνητής Νοημοσύνης (Artificial Intelligence-AI) και της επιστήμης υπολογιστών, η οποία έχει ως σκοπό τη μίμηση του τρόπου μάθησης των ανθρώπων κάνοντας χρήση δεδομένων και ειδικών αλγορίθμων. Ουσιαστικά είναι η ικανότητα των υπολογιστών να μαθαίνουν χωρίς να έχουν προγραμματιστεί ρητά για αυτό, όπως ορίζει τη μηχανική μάθηση για πρώτη φορά ο Άρθουρ Σάμουελ το 1959 [28]. Οι αλγόριθμοι της μηχανικής μάθησης λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, έτσι ώστε να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.

Τηράχουν τρεις κατηγορίες μηχανικής μάθησης, η Επιβλεπόμενη μάθηση (Supervised learning), η Μη-Επιβλεπόμενη μάθηση (Unsupervised learning) και η Ενισχυτική μάθηση (Reinforcement learning). Η διαφορά των παραπάνω κατηγοριών είναι στον τρόπο που εξάγει αποτελέσματα ο εκάστοτε αλγόριθμος μάθησης. Στην πρώτη κατηγορία τα μοντέλα εκπαιδεύονται γνωρίζοντας τη σωστή απάντηση, ενώ στη δεύτερη δεν υπάρχει σωστή απάντηση και οι αλγόριθμοι εξάγουν αποτελέσματα αναγνωρίζοντας μοτίβα και συσχετισμούς στα δεδομένα εισόδου, τέλος στη τρίτη κατηγορία ακολουθείται η τεχνική επιβράβευσης και τιμωρίας όπως συμβαίνει συχνά και στα έμβια όντα.

2.1.1 Επιβλεπόμενη Μάθηση

Η Επιβλεπόμενη Μάθηση προϋποθέτει τη δημιουργία ενός μοντέλου αντιστοίχισης ενός συνόλου δεδομένων εισόδου σε μία μεταβλητή εξόδου και στην συνέχεια την εφαρμογή του για την πρόβλεψη των μεταβλητών εξόδου όγνωστων δεδομένων εισόδου. Οι αλγόριθμοι επιβλεπόμενης μάθησης πετυχαίνουν τη παραπάνω λειτουργία προσαρμόζοντας τις παραμέτρους του μοντέλου ή αλλιώς τα βάρη του, έτσι ώστε να συμφωνούν όσο το δυνατόν καλύτερα με το σύνολο εκπαίδευσης (training set). Το σύνολο εκπαίδευσης αποτελείται από ζευγάρια εισόδου-εξόδου ή αλλιώς παραδείγματα, η είσοδος είναι συνήθως ένα διάνυσμα χαρακτηριστικών (feature vector) ενώ η έξοδος ονομάζεται ετικέτα (label). Μετά το πέρας της διαδικασίας εκπαίδευσης του μοντέλου, δηλαδή της προσαρμογής των βαρών του στο σύνολο εκπαίδευσης, το μοντέλο αξιολογείται κάνοντας προβλέψεις σε ένα όγνωστο σύνολο δεδομένων, το σύνολο δοκιμής (test set). Η επιτυχία της εκπαίδευσης του μοντέλου καθορίζεται από το ποσοστό σωστών προβλέψεων των ετικετών των νέων δεδομένων ή αλλιώς η ικανότητα γενίκευσης του μοντέλου από το σύνολο εκπαίδευσης

στο σύνολο δοκιμής.

Στην επιβλεπόμενη μάθηση εντάσσονται δύο κατηγορίες προβλημάτων, τα προβλήματα *Ταξινόμησης* και τα προβλήματα *Παλινδρόμησης*. Ο διαχωρισμός των δύο κατηγοριών έχει άμεση σχέση με το αν η τιμή της ετικέτας είναι ποιοτική ή ποσοτική, συγκεκριμένα:

- *Ταξινόμηση (Classification)*: σε αυτή τη κατηγορία προβλημάτων η τιμή της ετικέτας είναι ποιοτική, δηλαδή αντιπροσωπεύει μία κλάση που περιγράφει το διάνυσμα χαρακτηριστικών. Σκοπός του μοντέλου ταξινόμησης είναι η σωστή κατηγοριοποίηση των δεδομένων εισόδου σε κλάσεις. Ένα πρόβλημα ταξινόμησης μπορεί να πρέπει να χωρίσει τα δεδομένα σε δύο ή και περισσότερες κλάσεις, με τον αριθμό των κλάσεων να είναι προκαθορισμένος. Παραδείγματα προβλημάτων ταξινόμησης, τα οποία μπορούν εύκολα να επιλυθούν με μηχανική μάθηση, είναι η ταξινόμηση της ηλεκτρονικής αλληλογραφίας σε ανεπιθύμητη ή μη, η αναγνώριση αν σε μία εικόνα απεικονίζεται μία γάτα, ένα σκύλος ή ένα άλογο, ή ο προσδιορισμός του είδους ενός όγκου, καλοήθης ή κακοήθης, από ένα σύνολο χαρακτηριστικών του (μέγεθος, σύσταση, κτλ.).
- *Παλινδρόμηση (Regression)*: σε αυτή τη κατηγορία προβλημάτων η τιμή της ετικέτας είναι ποσοτική, το οποίο σημαίνει ότι παίρνει συνεχόμενες τιμές περιγράφοντας ένα μέγεθος. Τα μοντέλα που επιλύουν προβλήματα παλινδρόμησης αποσκοπούν στη πρόβλεψη μίας τιμής ενός μεγέθους για το οποίο έχουν στατιστικά ή ιστορικά δεδομένα. Τα μοντέλα παλινδρόμησης συναντώνται συχνά στον τομέα των πωλήσεων ως εναλλακτική μέθοδος στατιστικής, καθώς μπορούν να προβλέψουν ποσοστά πωλήσεων ενός προϊόντος έχοντας τα μέχρι τώρα δεδομένα. Άλλα παραδείγματα είναι η πρόβλεψη της θερμοκρασίας, της τιμής μιας μετοχής στο χρηματιστήριο ή της αξίας ενός ακινήτου.

2.1.2 Μη-Επιβλεπόμενη Μάθηση

Η Μη-Επιβλεπόμενη Μάθηση αναφέρεται στη περίπτωση όπου το σύνολο δεδομένων δεν έχει κατηγοριοποιηθεί ήδη από ανθρώπινο παράγοντα όπως στην Επιβλεπόμενη Μάθηση. Σκοπός των αλγορίθμων Μη-Επιβλεπόμενης μάθησης είναι η εύρεση μοτίβων (patterns) και δομών που υπάρχουν μέσα σε ένα σύνολο δεδομένων με αυτοματοποιημένο τρόπο. Η εξαγωγή γνώσης από τα δεδομένα γίνεται αυτόματα από τα μοντέλα μέσω του μηχανισμού ανάδρασης, ο οποίος τροποποιεί τις παραμέτρους των μοντέλων έτσι ώστε να γίνει η αναγνώριση των μοτίβων. Πρόκειται για ένα πολύ χρήσιμο εργαλείο από το οποίο μπορούμε να εξάγουμε γνώση από διάφορα είδη δεδομένων, τα οποία υπάρχουν σε πληθώρα στη σύγχρονη κοινωνία.

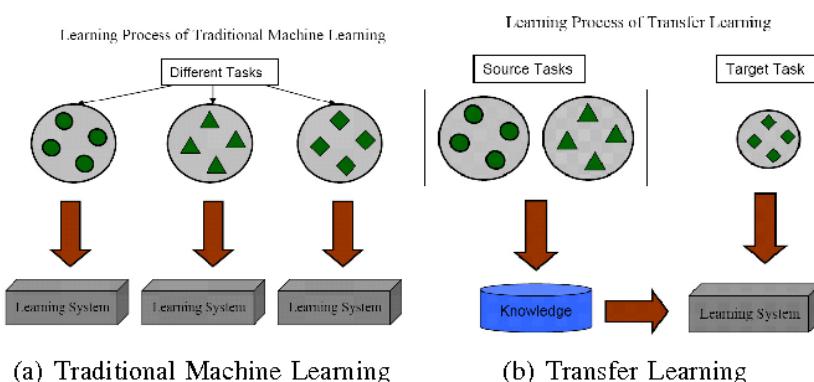
Προβλήματα Μη-Επιβλεπόμενης μάθησης αποτελούν η ομαδοποίηση (clustering) και η μείωση των διαστάσεων (dimensionality reduction). Το πρόβλημα της ομαδοποίησης αναφέρεται στον διαχωρισμό των δεδομένων σε ομάδες, οι οποίες θα παρουσιάζουν κοινά χαρακτηριστικά και παρόμοιες ιδιότητες. Παράδειγμα αυτού του προβλήματος αποτελεί η ομαδοποίηση του καταναλωτικού κοινού σύμφωνα με μία σειρά χαρακτηριστικών όπως οι προτιμήσεις τους για συγκεκριμένα προϊόντα, η ηλικία τους, κ.α. Το πρόβλημα της μείωσης των διαστάσεων αναφέρεται στη σύμπτυξη δεδομένων με την αφαιρεση μεταβλητών χωρίς όμως να χάνεται η αρχική πληροφορία των δεδομένων. Η μείωση διαστάσεων δεδομένων μπορεί να οδηγήσει σε ευκολότερη αποθήκευση δεδομένων, ταχύτερη εκτέλεση υπολογισμών, κ.α.

2.1.3 Ενισχυτική Μάθηση

Στην Ενισχυτική Μάθηση τα μοντέλα έρχονται αντιμέτωπα με ένα δυναμικό περιβάλλον από το οποίο δέχονται πληροφορίες και καλούνται να αλληλεπιδράσουν μαζί του μαθαίνοντας διάφορες ενέργειες. Οι αλγόριθμοι αυτοί λειτουργούν με ένα σύστημα ανταμοιβής (reward system), το οποία τα ωθεί στην βέλτιστη επιλογή ενεργειών. Η οντότητα που καλείται να ”μάθει” ονομάζεται πράκτορας (agent) και τα υπόλοιπα στοιχεία των μοντέλων αποτελούν το περιβάλλον (environment). Ο πράκτορας κατά τη διάρκεια της μάθησης επιλέγει ενέργειες που θα εκτελέσει και στη συνέχεια ανταμείβεται από το περιβάλλον αναλογικά με το κατά πόσο η ενέργειά του ήταν προς τη σωστή ή τη λανθασμένη κατεύθυνση. Ο πράκτορας αποκτά εμπειρία από τις προηγούμενες ενέργειες του, η οποία τελικά οδηγεί στην βέλτιστη αλληλεπίδρασή του με το περιβάλλον. Το συγκεκριμένο είδος μάθησης χρησιμοποιείται συχνά σε παιχνίδια, όπου ο πράκτορας έχει σκοπό να κερδίσει έναν άνθρωπο ή έναν άλλο υπολογιστή σε παιχνίδια όπως το σκάκι [29], τα γλεκτρονικά παιχνίδια [30], κ.α.

2.1.4 Μεταφορική Μάθηση

Η μεταφορική μάθηση (Transfer Learning) αποτελεί μία πολύ διαδεδομένη τεχνική μηχανικής μάθησης για το λόγο ότι εξοικονομεί χρόνο και βοηθάει να αναπτυχθούν μοντέλα με καλύτερη απόδοση. Ουσιαστικά γίνεται η μεταφορά της γνώσης που έχει αποκτήσει ένα μοντέλο, το οποίο έχει εκπαιδευτεί σε ένα σύνολο δεδομένων, σε ένα άλλο μοντέλο για τη επίλυση ενός προβλήματος που είναι σχετικό με το πρώτο. Για παράδειγμα, αν το πρόβλημα είναι η αναγνώριση φορτηγών σε εικόνες, η μεταφορά μάθησης από ένα μοντέλο που αναγνωρίζει αυτοκίνητα μπορεί σίγουρα να επιταχύνει την διαδικασία εκπαίδευσης αλλά και να οδηγήσει σε καλύτερα αποτελέσματα. Άλλο ένα πλεονέκτημα της μεταφορικής μάθησης είναι τα καλύτερα αποτελέσματα ακόμα και με μικρότερο όγκο δεδομένων εκπαίδευσης για το νέο μοντέλο καθώς η μεταφορά γνώσης γίνεται από μοντέλα που έχουν εκπαιδευτεί σε πολύ μεγάλα σύνολα δεδομένων και συνήθως είναι διαθέσιμα στο διαδίκτυο.



Σχήμα 2.1: Διαφορά μεταξύ (a) παραδοσιακής μάθησης και (b) μεταφορικής μάθησης [31].

2.2 Τεχνητά Νευρωνικά Δίκτυα (ANN)

Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks - ANNs) αποτελούν μία κατηγορία αλγορίθμων υλοποίησης της θεωρίας της Μηχανικής Μάθησης και γενικότερα της Τεχνητής Νοημοσύνης. Αυτό το ισχυρό εργαλείο είναι πλέον ο πιο διαδεδομένος και σύγχρονος τρόπος ανάπτυξης μοντέλων Μάθησης για την επίλυση διάφορων προβλημάτων.

2.2.1 Τρόπος Λειτουργίας

Το τεχνητό νευρωνικό δίκτυο είναι μία αναπαράσταση της δομής του ανθρώπινου εγκεφάλου υλοποιημένο από έναν υπολογιστή. Αποτελείται από νευρώνες και συνάψεις μεταξύ νευρώνων δομημένα σε επίπεδα (layers), όπως ακριβώς συμβαίνει και στον ανθρώπινο εγκέφαλο. Ένα ANN μπορεί να έχει εκατομμύρια νευρώνες συνδεδεμένους με διάφορους τρόπους δίνοντας του την ικανότητα να αναλύει και να αποθηκεύει μεγάλο όγκο πληροφοριών. Τα βασικά στοιχεία που συνθέτουν ένα ANN είναι οι νευρώνες, οι συνάψεις, τα βάρη, η πόλωση και οι συναρτήσεις.

- **Νευρώνες (Neurons):** αποτελούν τη βασική μονάδα επεξεργασίας ενός νευρωνικού δικτύου, η οποία δέχεται δεδομένα εισόδου, εκτελεί κάποιους απλούς υπολογισμούς και έπειτα μεταδίδει το αποτέλεσμα στον επόμενο νευρώνα. Τα μεγάλα νευρωνικά δίκτυα έχουν οργανωμένους τους νευρώνες σε επίπεδα. Τα βασικά επίπεδα νευρώνων είναι το επίπεδο εισόδου (input layer), ένα ή περισσότερα κρυφά επίπεδα (hidden layers) και το επίπεδο εξόδου (output layer). Τα δεδομένα εισέρχονται από το εξωτερικό περιβάλλον στο επίπεδο εισόδου, γίνεται η επεξεργασία τους στα κρυφά επίπεδα και το αποτέλεσμα εμφανίζεται στο επίπεδο εξόδου.
- **Συνάψεις και Βάρη (Synapses and Weights):** αποτελούν τον τρόπο με τον οποίο συνδέονται οι νευρώνες και μεταδίδεται η πληροφορία. Σε κάθε σύναψη αντιστοιχεί και ένα βάρος, το οποίο με την εκπαίδευση του νευρωνικού δικτύου αλλάζει τιμές έτσι ώστε να πετυχαίνει τον σκοπό του.
- **Πόλωση (Bias):** ή νευρώνες πόλωσης επιτρέπουν τα βάρη να παίρνουν περισσότερες τιμές, οι οπίσης αποθηκεύονται. Ουσιαστικά, ένας νευρώνας πόλωσης δημιουργεί μία πλουσιότερη αναπαράσταση των δεδομένων εισόδου μέσω των βαρών. Η πόλωση δεν είναι αναγκαία αλλά πολύ χρήσιμη στα πιο πολύπλοκα νευρωνικά δίκτυα.
- **Συναρτήσεις (Functions):** είναι οι λειτουργίες που περιέχει κάθε νευρώνας, έτσι ώστε να επεξεργαστεί τα δεδομένα εισόδου και να δημιουργήσει μια έξοδο. Συγκεκριμένα αναφέρομαστε στη συνάρτηση ενεργοποίησης (activation function), η οποία όπως προδίδει και η ονομασία, καθορίζει αν θα "ενεργοποιηθεί" ο νευρώνας ή όχι, δηλαδή αν η έξοδος του θα είναι 1 (ενεργός) ή 0 (ανενεργός). Οι πιο γνωστές συναρτήσεις ενεργοποίησης είναι η γραμμική, η σιγμοειδής και η υπερβολική εφαπτομένη.

Το πιο απλό Νευρωνικό Δίκτυο ονομάζεται αλγόριθμος Perceptron και προτάθηκε για πρώτη φορά το 1957 από τον F. Rosenblatt [32]. Το δίκτυο Perceptron αποτελείται από ένα επίπεδο νευρώνων, το οποίο αποτελεί είσοδο και έξοδο του δικτύου. Ειδικότερα, το Perceptron περιγράφεται και ως δυαδικός ταξινομητής, δηλαδή μία συνάρτηση η οποία απεικονίζει την είσοδο x

(διάνυσμα πραγματικών τιμών) σε μία τιμή εξόδου $f(x)$ σύμφωνα με τη σχέση:

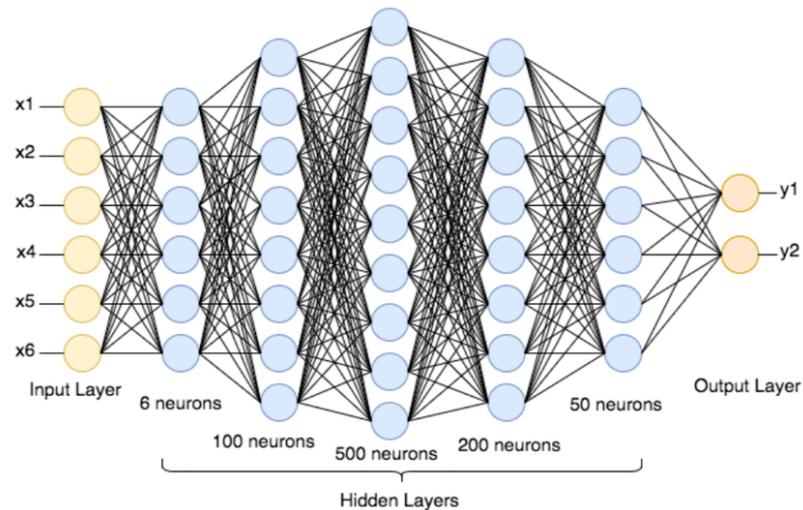
$$f(x) = \begin{cases} 1, & \text{if } w \cdot x + b > 0. \\ 0, & \text{otherwise.} \end{cases} \quad (2.1)$$

όπου w είναι το διάνυσμα βαρών και b είναι η πόλωση. Ο συγχεκριμένος αλγόριθμος μάθησης μπορεί να επιλύσει μόνο γραμμικά προβλήματα, συνεπώς ήταν αναγκαία η δημιουργία διαφορετικού αλγορίθμου για την επίλυση μη γραμμικών προβλημάτων. Αυτός ο αλγόριθμός είναι το Πολυεπίπεδο Perceptron (Multi-Layer Perceptron - MLP), το οποίο αποτελείται από τρία τουλάχιστον επίπεδα, το επίπεδο εισόδου, το κρυφό επίπεδο και το επίπεδο εξόδου.

2.2.2 Βαθιά Νευρωνικά Δίκτυα

Τα Βαθιά Νευρωνικά Δίκτυα και η Βαθιά Μάθηση, αναφέρονται σε δίκτυα που αποτελούνται από δύο ή παραπάνω κρυφά επίπεδα και η αρχιτεκτονική τους φαίνεται στο Σχήμα 2.2. Το πλεονέκτημα της βαθιάς μάθησης έναντι της παραδοσιακής μηχανικής μάθησης είναι η ικανότητα άντλησης χαρακτηριστικών από ανεπεξέργαστα δεδομένα [33]. Αυτό σημαίνει ότι η δύσκολη διαδικασία της δημιουργίας χαρακτηριστικών για την καλύτερη περιγραφή των δεδομένων γίνεται αυτόματα από το δίκτυο και όχι από τον προγραμματιστή, όπως γίνεται στη παραδοσιακή μέθοδο.

Επίσης, η βαθιά μάθηση μειώνει κατά πολύ τον ανθρώπινο παράγοντα και μπορεί να εξάγει χρήσιμες πληροφορίες από μεγάλο όγκο δεδομένων, χωρίς να γνωρίζει κάποιος από πριν ότι υπάρχουν, όπως γίνεται στη μη επιβλεπόμενη μάθηση. Η υλοποίηση της βαθιά μάθησης όμως απαιτεί μεγάλο όγκο δεδομένων, πολύ χρόνο και μεγάλη υπολογιστική, πράγματα όμως που καθημερινά γίνονται πιο προσβάσιμα με την εξέλιξη των υπολογιστών και του διαδικτύου.



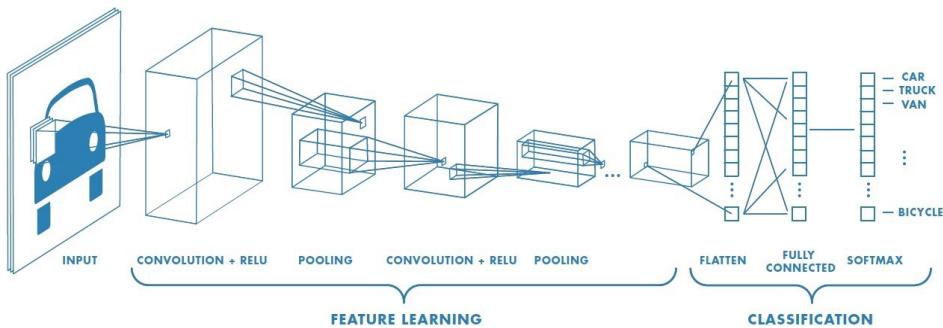
Σχήμα 2.2: Αναπαράσταση ενός Βαθύ Νευρωνικού Δικτύου ¹

¹ https://www.researchgate.net/figure/Deep-Neural-Network-architecture_fig1_330120030

2.3 Συνελικτικά Νευρωνικά Δίκτυα (CNN)

Τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNN) αποτελούν δίκτυα βαθιάς μάθησης τα οποία χρησιμοποιούνται κυρίως σε ανάπτυξη εφαρμογών που απαιτούν ανάλυση εικόνας και βίντεο. Το βασικό πλεονέκτημα των συνελικτικών δικτύων έναντι των παραδοσιακών πλήρως συνδεδεμένων δικτύων είναι η διατήρηση της χωρικής συσχέτισης που υπάρχει σε μία εικόνα.

Ο όρος "συνελικτικό" προέρχεται από την γραμμική μαθηματική πράξη της συνέλιξης, η οποία εμπεριέχει τον πολλαπλασιασμό ενός συνόλου βαρών με το διάνυσμα εισόδου, όπως και στα παραδοσιακά νευρωνικά δίκτυα. Η διαφορά είναι ότι στα συνελικτικά δίκτυα η δισδιάστατη είσοδος πολλαπλασιάζεται με έναν δισδιάστατο πίνακα από βάρη, ο οποίος ονομάζεται πυρήνας (kernel) ή φίλτρο (filter).



Σχήμα 2.3: Αρχιτεκτονική ενός Συνελικτικού Δικτύου ²

Ένα συνελικτικό νευρωνικό δίκτυο αποτελείται από δύο κύρια μέρη, όπως φαίνεται και στο Σχήμα 2.3. Στο πρώτο μέρος γίνεται η εξαγωγή των χαρακτηριστικών από την είσοδο και στο δεύτερο μέρος γίνεται η ταξινόμηση και η δημιουργία της εξόδου. Παρακάτω θα αναλύσουμε τα επίπεδα που συντελούν στο κάθε μέρος.

2.3.1 Συνελικτικό Επίπεδο (Convolutional Layer)

Το Συνελικτικό Επίπεδο είναι το βασικό δομικό στοιχείο ενός συνελικτικού δικτύου. Αρχικά, ας αναλύσουμε τη μαθηματική πράξη της συνέλιξης για σήματα διαχριτού χρόνου. Έχοντας τις συναρτήσεις $f, g \in \mathbf{Z}$, η διαχριτή συνέλιξή τους δίνεται από τη παρακάτω σχέση.

$$(f * g)[n] = \sum_{m=-\infty}^{\infty} f[m]g[n-m] = \sum_{m=-\infty}^{\infty} f[n-m]g[m] \quad (2.2)$$

Ο σκοπός του συνελικτικού επιπέδου είναι η ανίχνευση τοπικών χαρακτηριστικών από προηγούμενα επίπεδα και η προβολή τους σε ένα χάρτη χαρακτηριστικών (feature map). Ως αποτέλεσμα της συνέλιξης, η εικόνα χωρίζεται σε τοπικά δεκτικά πεδία τα οποία εν τέλει συμπλέζονται σε χάρτες χαρακτηριστικών μεγέθους $m_2 \times m_3$. Συνεπώς, οι χάρτες αυτοί αποθηκεύουν την

² <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks>

πληροφορία για το που βρίσκεται το χαρακτηριστικό στην εικόνα αλλά και πόσο καλά ανταποκρίνεται στο φίλτρο. Επομένως, το κάθε φίλτρο εκπαιδεύεται στο σημείο που αντιστοιχεί στη εικόνα.

Σε κάθε συνελικτικό επίπεδο, υπάρχει ένα σύνολο από m_1 φίλτρα. Το πλήθος των φίλτρων που εφαρμόζονται σε ένα επίπεδο είναι ανάλογο του βάθους που έχουν οι χάρτες χαρακτηριστικών εξόδου. Το κάθε φίλτρο ανιχνεύει ένα συγκεκριμένο χαρακτηριστικό σε κάθε περιοχή της εισόδου. Πιο αναλυτικά, η έξοδος $Y_i^{(l)}$ του επιπέδου l αποτελείται από $m_1^{(l)}$ χάρτες χαρακτηριστικών μεγέθους $m_2^{(l)} \times m_3^{(l)}$. Ο i^{th} χάρτης χαρακτηριστικών συμβολίζεται ως $Y_i^{(l)}$ και υπολογίζεται ως εξής

$$Y_i^{(l)} = B_i^{(l)} + \sum_{j=1}^{m_1^{(l-1)}} K_{i,j}^{(l)} * Y_j^{(l-1)} \quad (2.3)$$

όπου $B_i^{(l)}$ είναι ο πίνακας πόλωσης (bias) και $K_{i,j}^{(l)}$ είναι το φίλτρο μεγέθους $2h_1^{(l)} + 1 \times 2h_2^{(l)} + 1$ συνδέοντας έτσι το j^{th} χάρτη χαρακτηριστικών στο επίπεδο $(l-1)$ με τον i^{th} χάρτη στο επίπεδο l .

Το αποτέλεσμα της τοποθέτησης πολλών συνελικτικών επιπέδων στη σειρά και μαζί με τα επίπεδα που θα αναφέρουμε στη συνέχεια είναι η ταξινόμηση της πληροφορίας όπως γίνεται στην ανθρώπινη όραση. Αυτό σημαίνει ότι τα εικονοστοιχεία (pixels) της εικόνας μετατρέπονται σε ακμές, μετά σε μοτίβα και στο τέλος σε αντικείμενα.

2.3.2 Επίπεδο Ενεργοποίησης (Activation Layer)

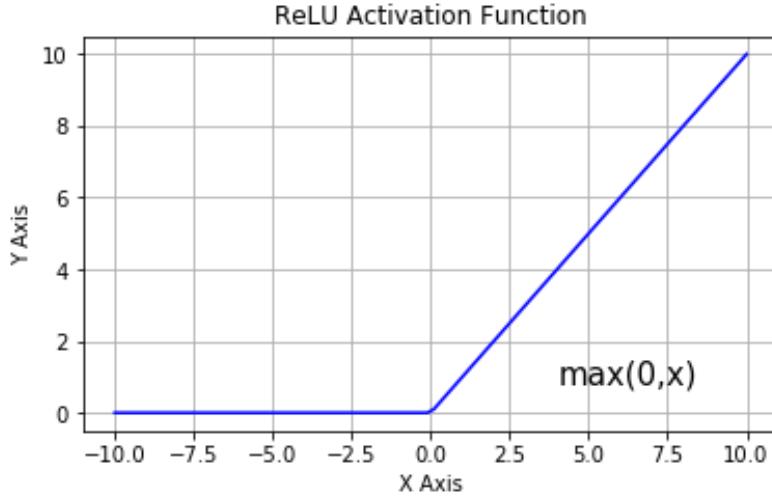
Στα συνελικτικά νευρωνικά δίκτυα το επίπεδο ενεργοποίησης μεταξύ συνελικτικών επιπέδων αποτελείται από την μονάδα *ReLU* (*Rectified Linear Unit*), η οποία εισάγει τη μη-γραμμικότητα και συγχρόνως είναι και η συνάρτηση ενεργοποίησης. Μία μονάδα ReLU με κατώφλι το 0 περιγράφεται από τη σχέση:

$$Y_i^{(l)} = \max(0, Y_i^{(l-1)}) \quad (2.4)$$

Τα πλεονεκτήματα της συγκεκριμένης μονάδας ενεργοποίησης στα συνελικτικά δίκτυα, σε σχέση με παραδοσιακές συναρτήσεις ενεργοποίησης όπως είναι η σιγμοειδής ή η υπερβολική εφαπτομένη, είναι:

- Οι ReLUs έχουν την ικανότητα να μεταδίδουν την κλίση μεταξύ των επιπέδων πιο αποδικά, με αποτέλεσμα να αποφεύγεται η εξαφάνιση κλίσης (vanishing gradient) που αποτελεί συχνό φαινόμενο στα βαθιά νευρωνικά δίκτυα.
- Οι ReLUs παρουσιάζουν μη θετικές τιμές κατωφλιού, το οποίο επιλύει το πρόβλημα της ακύρωσης και συμβάλει σε ένα σποραδικό όγκο ενεργοποίησης στην έξοδο τους. Η σποραδικότητα της εξόδου δημιουργεί ανθεκτικότητα σε μικρές διακυμάνσεις της εισόδου, που αποτελούν το θόρυβο [34].
- Οι ReLUs απαρτίζονται μόνο από απλές πράξεις όσο αναφορά το υπολογιστικό τους κόστος (χυρίως συγχρίσεις), συνεπώς είναι πιο αποδοτικές στην υλοποίηση τους.

³ <https://learnopencv.com/understanding-activation-functions-in-deep-learning/relu-activation-function-2/>



Σχήμα 2.4: Συνάρτηση Ενεργοποίησης ReLU ³

2.3.3 Επίπεδο Υποδειγματοληψίας (Pooling Layer)

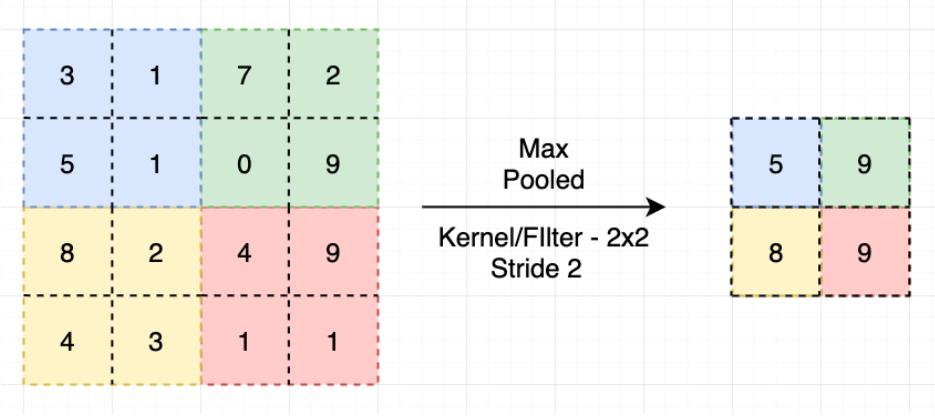
Το επίπεδο υποδειγματοληψίας (pooling or downsampling layer) είναι υπεύθυνο για τη μείωση του μεγέθους των χαρτών ενεργοποίησης (activation maps). Τα επίπεδα αυτά χρησιμοποιούνται σε πολλά σημεία μέσα σε ένα συνελικτικό δίκτυο, συνήθως μετά από κάποιο συνελικτικό επίπεδο ή επίπεδο ενεργοποίησης (βλ. Σχήμα 2.3), με σκοπό τη μείωση των υπολογιστικών απαιτήσεων σταδιακά κατά μήκος του δικτύου αλλά και την αποφυγή της πιθανότητας υπερπροσαρμογής (overfitting).

Το επίπεδο υποδειγματοληψίας l έχει δύο υπερπαραμέτρους, το μέγεθος του φίλτρου $F^{(l)}$ και το βήμα (stride) $S^{(l)}$. Ως είσοδο δέχεται έναν τρισδιάστατο πίνακα μεγέθους $m_1^{(l-1)} \times m_2^{(l-1)} \times m_3^{(l-1)}$ και παράγει στην έξοδο έναν πίνακα μεγέθους $m_1^{(l)} \times m_2^{(l)} \times m_3^{(l)}$ όπου:

$$\begin{aligned} m_1^{(l)} &= m_1^{(l-1)} \\ m_2^{(l)} &= (m_2^{(l-1)} - F^{(l)}) / S^{(l)} + 1 \\ m_3^{(l)} &= (m_3^{(l-1)} - F^{(l)}) / S^{(l)} + 1 \end{aligned} \quad (2.5)$$

Το επίπεδο υποδειγματοληψίας λειτουργεί ορίζοντας ένα παράθυρο μεγέθους $F^{(l)} \times F^{(l)}$ και ελαττώνοντας τα δεδομένα του παραθύρου σε μία μοναδική τιμή. Το παράθυρο μετακινείται κατά $S^{(l)}$ θέσεις κάθε φορά, όπως συμβαίνει και στο συνελικτικό επίπεδο, μέχρι να ελαττωθεί όλος ο όγκος εισόδου.

Οι πιο γνωστές μέθοδοι υποδειγματοληψίας αποτελούν η υποδειγματοληψία μέγιστης τιμής (max pooling) και η υποδειγματοληψία μέσης τιμής (average pooling). Στη πρώτη περίπτωση χρησιμοποιείται μόνο η υψηλότερη τιμή εντός του παραθύρου υποδειγματοληψίας και οι υπόλοιπες τιμές απορρίπτονται, ενώ στη δεύτερη περίπτωση χρησιμοποιείται ο μέσος όρος των τιμών του παραθύρου. Η μέθοδος υποδειγματοληψίας μέγιστης τιμής παρουσιάζει γρηγορότερη σύγκλιση και καλύτερη απόδοση σε σχέση με τη μέθοδο της μέσης τιμής και άλλες μεθόδους όπως την υποδειγματοληψία L^2 -norm [35], γεγονός που τη κάνει τη πιο διαδεδομένη μέθοδο στα σύγχρονα συνελικτικά δίκτυα.



Σχήμα 2.5: Παράδειγμα υποδειγματοληψίας μέγιστης τιμής με φίλτρο μεγέθους 2×2 και βήμα 2 ⁴.

2.3.4 Πλήρως Συνδεδεμένο Επίπεδο (Fully Connected Layer)

Τα πλήρως συνδεδεμένα επίπεδα σε ένα συνελικτικό νευρωνικό δίκτυο είναι πρακτικά ένα πολυεπίπεδο perceptron (Multi-layer Perceptron - MLP), συνήθως δύο ή τριών επιπέδων, το οποίο σκοπεύει στην αντιστοίχιση του $m_1^{(l-1)} \times m_2^{(l-1)} \times m_3^{(l-1)}$ πίνακα ενεργοποίησης που έχει προκύψει από το συνδυασμό των προηγούμενων επιπέδων σε μια κατανομή πιθανότητας για τη κάθε κλάση. Συνεπώς, η έξοδος του επιπέδου του MLP υποδικτύου θα έχει $m_1^{(l-i)}$ τιμές, όπου i το πλήθος των επιπέδων του MLP.

Η κύρια διαφορά από ένα παραδοσιακό πολυεπίπεδο perceptron είναι ότι στην είσοδο δέχεται έναν τρισδιάστατο πίνακα και όχι ένα διάνυσμα. Συνεπώς το πλήρως συνδεδεμένο επίπεδο ορίζεται ως:

Έστω το $l - 1$ είναι πλήρως συνδεδεμένο επίπεδο, τότε

$$y_i^{(l)} = f(z_i^{(l)}) \quad \text{όπου} \quad z_i^{(l)} = \sum_{j=1}^{m_1^{(l-1)}} w_{i,j}^{(l)} y_j^{(l-1)} \quad (2.6)$$

αλλιώς,

$$y_i^{(l)} = f(z_i^{(l)}) \quad \text{όπου} \quad z_i^{(l)} = \sum_{j=1}^{m_1^{(l-1)}} \sum_{r=1}^{m_2^{(l-1)}} \sum_{s=1}^{m_3^{(l-1)}} w_{i,j,r,s}^{(l)} (Y_i^{(l-1)})_{r,s} \quad (2.7)$$

Ο στόχος του πλήρως συνδεδεμένου επιπέδου είναι η σωστή ρύθμιση των παραμέτρων $w_{i,j}^{(l)}$ ή $w_{i,j,r,s}^{(l)}$ για τη δημιουργία της στοχαστικής πιθανότητας που αντιπροσωπεύει κάθε μία από τις κλάσεις, βασισμένη στους χάρτες ενεργοποίησης που παρήγαγαν μία ακολουθία συνελικτικών επιπέδων, επιπέδων ενεργοποίησης και επιπέδων υποδειγματοληψίας.

2.3.5 Επίπεδο Ενεργοποίησης Softmax

Το επίπεδο ενεργοποίησης είναι παρόμοιο με αυτό που αναφέραμε προηγουμένως με τη διαφορά ότι τώρα χρησιμοποιείται διαφορετική συνάρτηση ενεργοποίησης, η Softmax. Το επίπεδο

⁴ <https://ai.plainenglish.io/pooling-layer-beginner-to-intermediate-fa0dbdce80eb>

αυτό εισάγεται μετά το πλήρως συνδεδεμένο επίπεδο και έχει τη λειτουργία να αλλάζει τις τιμές της εξόδου του προηγούμενο επιπέδου έτσι ώστε το άθροισμα του να είναι η μονάδα. Με αυτό τον τρόπο εκφράζεται η κατανομή πιθανότητας για κάθε κλάση του προβλήματος ταξινόμησης. Η συνάρτηση Softmax ορίζεται ως:

$$f_j(z) = \frac{e^{z_j}}{\sum_k e^{z_k}} \quad (2.8)$$

όπου z_i οποιοσδήποτε πραγματικός αριθμός.

2.3.6 Επίπεδο Κανονικοποίησης Παρτίδας (Batch Normalization Layer)

Το επίπεδο Κανονικοποίησης Παρτίδας είναι μία τεχνική που συμβάλει στη γρηγορότερη σύγκλιση και στην πιο αποδοτική εκπαίδευση δικτύων με πολλά κρυφά επίπεδα [36]. Ειδικότερα, μειώνει την εμφάνιση του προβλήματος που αναφέρεται ως "internal covariate shift" και ορίζεται ως η αλλαγή της κατανομής των χαρτών ενεργοποίησης λόγω της αλλαγής των παραμέτρων του δικτύου κατά τη διάρκεια της εκπαίδευσης. Η κανονικοποίηση παρτίδας μπορεί να εφαρμοστεί μετά από συνελικτικά επίπεδα, επίπεδα υποδειγματοληψίας αλλά και πλήρως συνδεδεμένα επίπεδα.

Η λειτουργία του επιπέδου αυτού είναι να αφαιρεί από την είσοδο το μέσο όρο της παρτίδας (batch) και να διαιρεί με την τυπική απόκλιση της παρτίδας, το αποτέλεσμα είναι ότι η έξοδος του επιπέδου έχει μέσο όρο μηδέν και διακύμανση μονάδα. Με μαθηματικούς όρους η κανονικοποίηση παρτίδας ορίζεται ως:

Αν $x \in B$ η είσοδος της κανονικοποίησης παρτίδας (BN) με το σύνολο της παρτίδας, τότε

$$BN(x) = \gamma \cdot \frac{x - \hat{\mu}_B}{\hat{\sigma}_B} + \beta \quad (2.9)$$

όπου $\hat{\mu}_B$ η μέση τιμή της παρτίδας και $\hat{\sigma}_B$ η τυπική απόκλιση. Επίσης οι παράμετροι αλλαγής κλίμακας γ και πόλωσης β έχουν διαστάσεις ίσες με την είσοδο x και αποτελούν εκπαιδεύσιμες παράμετροι του δικτύου.

2.3.7 Επίπεδο Dropout

Το επίπεδο Dropout ουσιαστικά είναι μία τεχνική προσθήκης θορύβου ανάμεσα στα επίπεδα των βαθιών νευρωνικών δικτύων, η οποία συμβάλει στην αποφυγή της υπερπροσαρμογής του δικτύου κατά την εκπαίδευση. Η αρχική ιδέα προέρχεται από τον Christopher Bishop, ο οποίος αποδεικνύει ότι μία συνάρτηση είναι ομαλή και ταυτόχρονα απλή όταν είναι ανεκτική στον θόρυβο της εισόδου [37]. Την ιδέα αυτή υλοποίησαν έξυπνα οι Srivastava et al. [38] το 2014, προσθέτωντας θόρυβο μεταξύ των κρυφών επιπέδων των δικτύων.

Συγκεκριμένα, το επίπεδο Dropout αφαιρεί κάποιους νευρώνες μεταξύ των επιπέδων με μία πιθανότητα p και στους νευρώνες που παραμένουν η ενδιάμεση ενεργοποίηση h αντικαταστέται με την τυχαία μεταβλητή h' όπως φαίνεται στη συνέχεια:

$$h' = \begin{cases} 0, & \text{with probability } p. \\ \frac{h}{1-p}, & \text{otherwise.} \end{cases} \quad (2.10)$$

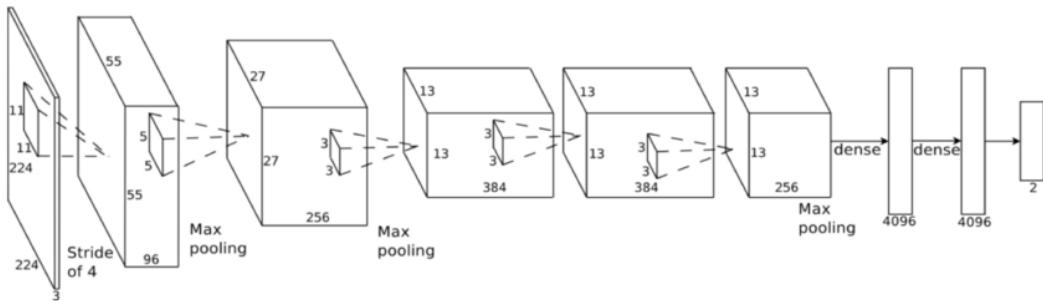
2.4 Σύγχρονα Συνελικτικά Νευρωνικά Δίκτυα

Γνωρίζοντας πλέον τα βασικά στοιχεία που συντελούν στην ανάπτυξη ενός συνελικτικού νευρωνικού δίκτυου, στη παράγραφο αυτή θα αναφερθούμε σε γνωστές αρχιτεκτονικές δίκτυων που χρησιμοποιούνται κατά κόρον για την επίλυση προβλημάτων αναγνώρισης προτύπων σε εικόνες σήμερα.

2.4.1 AlexNet

Το συνελικτικό δίκτυο 8 επιπέδων, που ονομάστηκε *AlexNet* από τον δημιουργό του Alex Krizhevsky, κέρδισε τον διαγωνισμό ταξινόμησης εικόνων στο σύνολο δεδομένων *ImageNet* [39] το 2012 με φανομενικά μεγάλη διαφορά [40]. Το σύνολο δεδομένων *ImageNet* είναι μία μεγάλη βάση δεδομένων εικόνων, η οποία δημιουργήθηκε για την έρευνα αναγνώρισης αντικειμένων σε εικόνες. Περισσότερες από 14 εκατομμύρια εικόνες έχουν κατηγοριοποιηθεί σε σχέση με το αντικείμενο που απεικονίζουν σε πάνω από 20 χιλιάδες κατηγορίες. Από το 2010 και έπειτα διοργανώνεται κάθε χρόνο ο διαγωνισμός οπτικής αναγνώρισης αντικειμένων του *ImageNet* (ImageNet Large Scale Visual Recognition Challenge - ILSVRC). Το 2012 λοιπόν, το *AlexNet* είχε το χαμηλότερο ποσοστό σφάλματος, μόλις 15.3%, το οποίο ήταν 10.8% χαμηλότερο από το αμέσως επόμενο διαγωνίζομενο.

Το *AlexNet* αποτελείται από πέντε συνελικτικά επίπεδα, τρία επίπεδα υποδειγματοληψίας, δύο πλήρως συνδεδεμένα κρυφά επίπεδα και ένα πλήρως συνδεδεμένο επίπεδο εξόδου, η ακριβής αρχιτεκτονική του φαίνεται στο Σχήμα 2.6. Επίσης χρησιμοποιεί την ReLU ως συνάρτηση ενεργοποίησης και όχι τη σιγμοειδή όπως γινόταν μέχρι τότε. Για την εκπαίδευση του μοντέλου ήταν αναγκαία η χρήση δύο Μονάδων Επεξεργασίας Δεδομένων (Graphics Processing Unit - GPU), έτσι ώστε να είναι εφικτή η αποθήκευση των χιλιάδων παραμέτρων του μοντέλου.



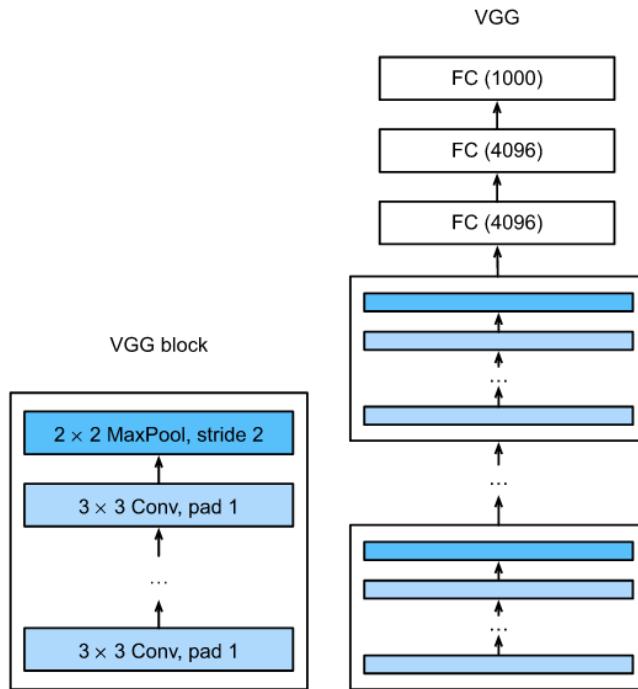
Σχήμα 2.6: Η Αρχιτεκτονική του συνελικτικού δίκτυου *AlexNet* ⁵.

2.4.2 Δίκτυο VGG

Το δίκτυο VGG είναι ένα βαθύ συνελικτικό δίκτυο, το οποίο προτάθηκε από την ομάδα Εικονικής Γεωμετρίας (Visual Geometry Group) του πανεπιστημίου της Οξφόρδης από όπου πήρε και την ονομασία του [41]. Συγκεκριμένα, το μοντέλο πήρε μέρος στο διαγωνισμό οπτικής αναγνώρισης αντικειμένων του *ImageNet* το 2014 και αναδείχθηκε πρώτο με μόλις 6.8% ποσοστό

⁵ https://www.researchgate.net/figure/AlexNet-CNN-architecture-layers_fig1_318168077

σφάλματος (top-5 error rate) στο σύνολο δοκιμής (test set). Η βελτίωση της απόδοσης του μοντέλου σε σχέση με το AlexNet οφείλεται στην αντικατάσταση των μεγάλων σε διαστάσεις φίλτρων των πρώτων δύο συνελικτικών επιπέδων του (11×11 και 5×5 αντίστοιχα) με πολλαπλά συνελικτικά επίπεδα με φίλτρα διαστάσεων 3×3 . Ειδικότερα, το δίκτυο VGG αποτελείται από μπλοκς επιπέδων (VGG blocks), τα οποία απαρτίζονται από μία σειρά συνελικτικών επιπέδων, επιπέδων ενεργοποίησης (ReLUs) και επιπέδων υποδειγματοληψίας όπως φαίνεται στο Σχήμα 2.7. Το τελευταίο μέρος του δικτύου αποτελείται από τρία πλήρως συνδεδεμένα επίπεδα όπως ακριβώς και το AlexNet.



Σχήμα 2.7: Η Αρχιτεκτονική του συνελικτικού δικτύου VGG⁶.

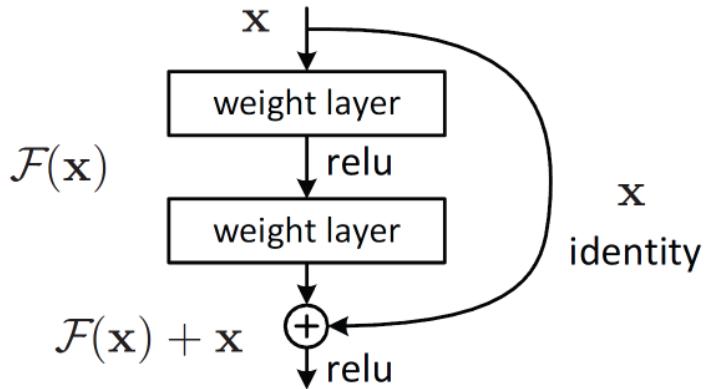
2.4.3 Διαφορικά Δίκτυα (Residual Networks)

Τα βαθιά συνελικτικά νευρωνικά δίκτυα έχουν αποδειχτεί πολύ αποτελεσματικά για την εξαγωγή χαρακτηριστικών από εικόνες και το πλήθος των κρυφών επιπέδων φαίνεται να παίζει σημαντικό ρόλο στην "εκφραστικότητα" των μοντέλων. Ωστόσο, η αύξηση του βάθους των δικτύων οδήγησε στο πρόβλημα υποβιβασμού της απόδοσής τους, με αποτέλεσμα η ακρίβεια του μοντέλου να φτάνει σε έναν κορεσμό και στη συνέχεια να μειώνεται απότομα και να αυξάνεται το σφάλμα εκπαίδευσης. Λύση στο παραπάνω πρόβλημα έδωσε το Διαφορικό (Residual) δίκτυο ή ResNet, το οποίο εισήγαγε ερευνητική ομάδα της Microsoft το 2015 [42]. Το ResNet κατέλαβε την πρώτη θέση στο διαγωνισμό του ImageNet το 2015 με μόλις 3.57% ποσοστό σφάλματος στο σύνολο δοκιμής, αρκετά χαμηλότερο από εκείνο του VGG.

Ο τρόπος αντιμετώπισης του προβλήματος υποβιβασμού του δικτύου με την αύξηση των επιπέδων γίνεται με την εισήγηση του πλαισίου της βαθιάς διαφορικής μάθησης. Αντίθετα με τον

⁶ https://d2l.ai/_images/vgg.svg

παραδοσιακό τρόπο μάθησης ενός συνελικτικού δικτύου, στον οποίο τα διάφορα επίπεδα προσπαθούν άμεσα να προσαρμοστούν στην επιθυμητή υποκείμενη απεικόνιση, στα διαφορικά δίκτυα η προσαρμογή γίνεται στην διαφορική απεικόνιση. Αναλυτικότερα, αν η αρχική υποκείμενη απεικόνιση συμβολίζεται ως $H(x)$, τότε η διαφορική απεικόνιση ορίζεται ως $F(x) := H(x) - x$, όπου x η είσοδος του εκάστοτε επιπέδου. Συνεπώς η αρχική απεικόνιση γίνεται πλέον $F(x) + x$. Η υλοποίηση της σχέσης $F(x) + x$ σε ένα δίκτυο γίνεται με συνδέσεις συντόμευσης (shortcut connections). Στο Σχήμα 2.8 φαίνεται ένα διαφορικό μπλοκ (residual block), στο οποίο η σύνδεση συντόμευσης γίνεται πολλαπλασιάζοντας την είσοδο x με τον μοναδιαίο πίνακα.



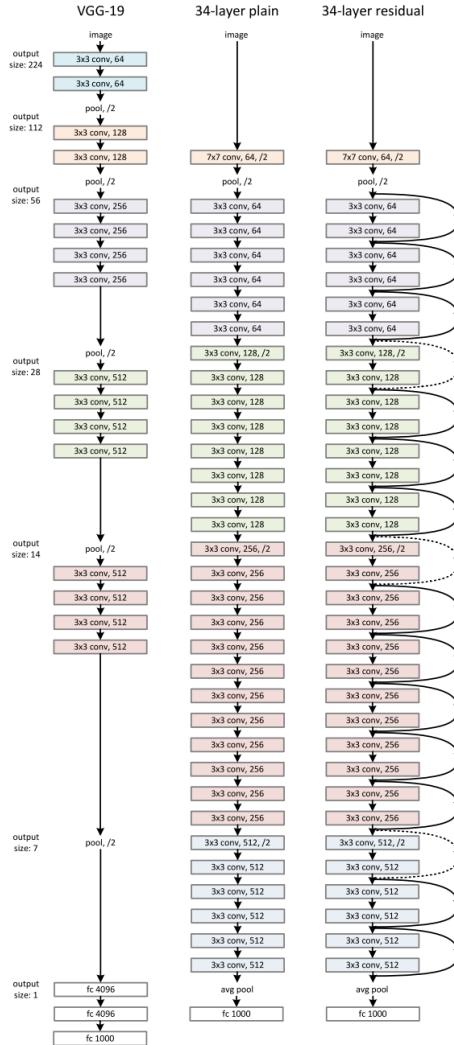
Σχήμα 2.8: Η υλοποίηση της διαφορικής μάθησης μέσω του διαφορικού μπλοκ (Residual Block)[42].

Χρησιμοποιώντας αυτή τη τεχνική, τα ResNets πέτυχαν δύο σημαντικούς στόχους, πρώτον δημιούργησαν δίκτυα που είναι ευκολότερα στην βελτιστοποίηση από τα αντίστοιχα απλά βαθιά δίκτυα, στα οποία το σφάλμα εκπαίδευσης αυξάνεται με την αύξηση των επιπέδων και δεύτερον η αύξηση του βάθους στα διαφορικά δίκτυα οδηγεί σε καλύτερη απόδοση του δικτύου, καλύτερη από όλα τα προηγούμενα γνωστά συνελικτικά δίκτυα.

Η αρχιτεκτονική ενός ResNet, βασίζεται σε εκείνη του VGG, καθώς αποτελείται από σειρά συνελικτικών επιπέδων με φίλτρα μικρής διάστασης (3×3), επίπεδα ενεργοποίησης ReLU και επίπεδα υποδειγματοληψίας. Στο στάδιο της ταξινόμησης υπάρχει ένα επίπεδο που εκτελεί ολική υποδειγματοληψία μέσης τιμής (Average Pooling) και στη συνέχεια ένα πλήρως συνδεδεμένο επίπεδο με διάσταση εξόδου 1000 ακολουθούμενο από ένα επίπεδο ενεργοποίησης softmax για την ταξινόμηση των κλάσεων του ImageNet. Το χαρακτηριστικό που διαχωρίζει το ResNet από το αντίστοιχο απλό δίκτυο είναι οι συνδέσεις συντόμευσης μεταξύ των επιπέδων. Συγκεκριμένα, όπως φαίνεται και στο Σχήμα 2.9, αυτές οι συνδέσεις γίνονται ανά δύο συνελικτικά επίπεδα και υλοποιούνται πολλαπλασιάζοντας την είσοδο με τον μοναδιαίο πίνακα όταν οι διαστάσεις παραμένουν ίδιες (συνδέσεις με συνεχόμενη γραμμή) ή όταν οι διαστάσεις αλλάζουν χρησιμοποιούνται δύο τεχνικές που μεταβάλλουν τη διάσταση της εισόδου έτσι ώστε να είναι δυνατή η πρόσθεση στοιχείο προς στοιχείο (συνδέσεις με διακεκομένη γραμμή). Η πρώτη τεχνική είναι η αύξηση της διάστασης εισόδου εισάγοντας μηδενικά στις άκρες (zero-padding) και η δεύτερη είναι με την εφαρμογή μίας γραμμικής απεικόνισης W_s στην είσοδο x με τρόπο που φαίνεται στη σχέση 2.11, όπου y είναι η έξοδος του επιπέδου.

$$y = F(x, \{W_i\}) + W_s x \quad (2.11)$$

Το πλεονέκτημα της πρώτης τεχνικής είναι ότι δεν εισάγει νέες παραμέτρους στο μοντέλο με αποτέλεσμα να είναι λιγότερο περίπλοκο και να χρειάζεται λιγότερο χρόνο και λιγότερη υπολογιστική ισχύ για την εκπαίδευσή του.

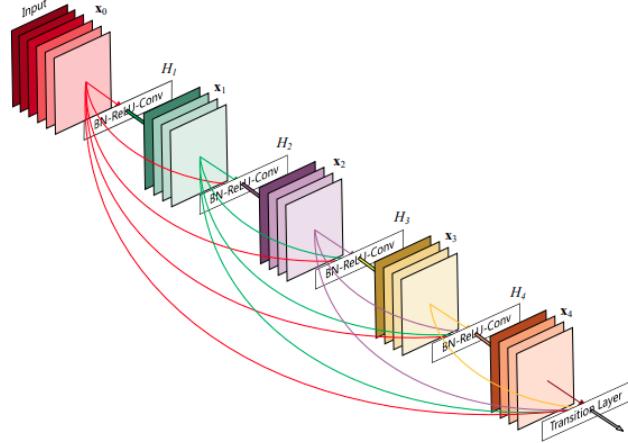


Σχήμα 2.9: Αρχιτεκτονικές των μοντέλων Vgg-19 (αριστερά), και του απλού και διαφορικού μοντέλου του ResNet-34[42].

2.4.4 Πυκνά Συνδεδεμένα Δίκτυα (DenseNet)

Τα Πυκνά Συνδεδεμένα Δίκτυα (DenseNets) βασίζονται στην αρχιτεκτονική των Διαφορικών Δικτύων, καθώς αποτελούνται από παρόμοια μπλοκ επιπέδων αλλά διαφέρουν στον τρόπο διασύνδεσης. Η λογική τους είναι η επιπλέον βελτίωση της ροής της πληροφορίας κατά μήκος του δικτύου, συνεπώς προτείνουν την άμμεση σύνδεση όλων των επιπέδων με όλα τα επόμενα όπως φαίνεται στο Σχήμα 2.10. Παρατηρώντας το σχήμα κατανοούμε καλύτερα την επιλογή της ονομασίας του δικτύου σε πυκνά συνδεδεμένου.

Αναλυτικότερα, το l^{th} επίπεδο δέχεται ως είσοδο τους χάρτες χαρακτηριστικών από όλα τα



Σχήμα 2.10: Αρχιτεκτονική ενός DenseNet 5 επιπέδων [43].

προηγούμενα επίπεδα, x_0, \dots, x_{l-1} , όπως φαίνεται παρακάτω:

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}]) \quad (2.12)$$

όπου $[x_0, x_1, \dots, x_{l-1}]$ συμβολίζει τη συγκέντρωση των χαρτών χαρακτηριστικών των προηγούμενων επιπέδων σε έναν τένσορα. Συγχρίνοντας τη διασύνδεση των επιπέδων των DenseNets με τα ResNets, παρατηρούμε ότι η βασική διαφορά είναι ότι στα πρώτα γίνεται συγκέντρωση των χαρακτηριστικών από όλα τα προηγούμενα επίπεδα ενώ στα δεύτερα γίνεται πρόσθεση των πινάκων των χαρακτηριστικών σε κάθε επίπεδο.

Κεφάλαιο 3

Δεδομένα

Στο παρόν Κεφάλαιο θα παρουσιάσουμε τους τρόπους αναπαράστασης του ανθρώπινου συναισθήματος, οι οποίοι έχουν αναπτυχθεί ώστε να είναι πιο εύκολη και εύστοχη η περιγραφή τους. Στη συνέχεια, θα αναλύσουμε τα χαρακτηριστικά και τον τρόπο συλλογής του συνόλου δεδομένων Aff-Wild2, το οποίο θα χρησιμοποιηθεί στη συνέχεια ως είσοδος των μοντέλων βαθιάς μάθησης που θα αναπτύξουμε. Έπειτα, θα περιγράψουμε την διαδικασία που ακολουθήθηκε ώστε τα δεδομένα του συνόλου να μετατραπούν σε μορφή συμβατή ώστε να γίνει η τροφοδότησή τους στα μοντέλα.

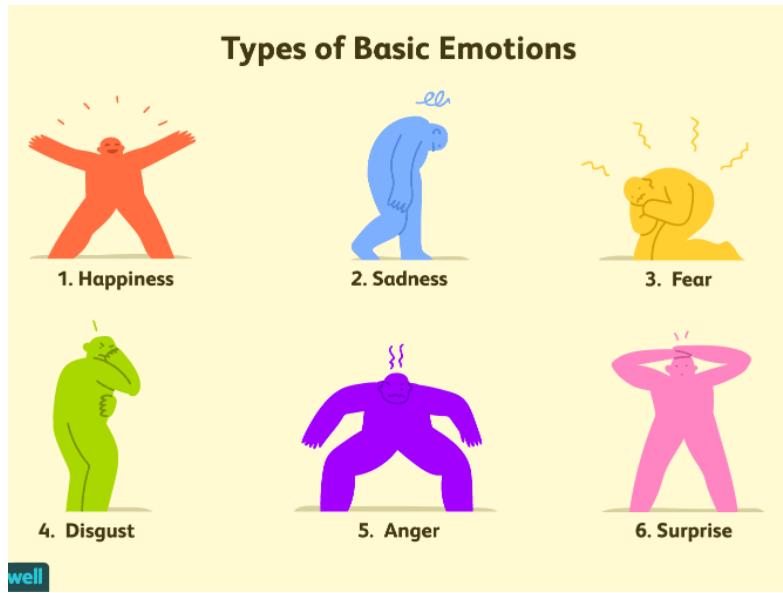
3.1 Τρόποι Αναπαράστασης Συναισθήματος

Η αναπαράσταση του ανθρώπινου συναισθήματος αποτελεί κυρίαρχο θέμα έρευνας της επιστήμης της ψυχολογίας, καθώς έχουν αναπτυχθεί διάφορα μοντέλα που έχουν ως στόχο την περιγραφή του ανθρώπινου συναισθήματος παρόλο τη μεγάλη ποικιλομορφία του. Τα δύο μοντέλα που θα μας απασχολήσουν στην παρούσα εργασία, είναι το κατηγορικό μοντέλο, το οποίο χωρίζει τα συναισθήματα σε 7 βασικές κατηγορίες και το διαστατικό μοντέλο, το οποίο περιγράφει οποιοδήποτε συναίσθημα με την τοποθέτησή του σε ένα δισδιάστατο χώρο.

3.1.1 Κατηγορικό Μοντέλο

Το μοντέλο αναπαράστασης συναισθημάτων, το οποίο κατηγοριοποιεί όλα τα συναισθήματα σε επτά βασικές κατηγορίες είναι το πιο ευραίως διαδεδομένο στις έρευνες αναγνώρισης συναισθήματος. Οι 7 αυτές κατηγορίες είναι ο Θυμός, η Αηδία, ο Φόβος, η Χαρά, η Λύπη, η Έκπληξη και η Ουδετερότητα [44, 45]. Υπάρχουν στοιχεία ότι τα 7 αυτά συναισθήματα συναντώνται σε όλους τους ανθρώπους ανεξάρτητα την ηλικία, την εθνικότητα, το φύλο και οτιδήποτε μπορεί να διαχωρίσει τους ανθρώπους. Ωστόσο, τα συναισθήματα αυτά εκφράζονται από τους ανθρώπους σε διάφορες εντάσεις και με διάφορους τρόπους στη καθημερινή ζωή, με αποτέλεσμα η κάθε κατηγορία να περιλαμβάνει μία πολύ μεγάλη γκάμα συναισθημάτων. Το γεγονός αυτό δειχνεί ότι το μοντέλο αυτό δεν μπορεί να περιγράψει με ακρίβεια ένα συναίσθημα αλλά το εντάσσει σε ένα ευρύτερο πλαίσιο συναισθημάτων.

¹ <https://www.verywellmind.com/an-overview-of-the-types-of-emotions-4163976>



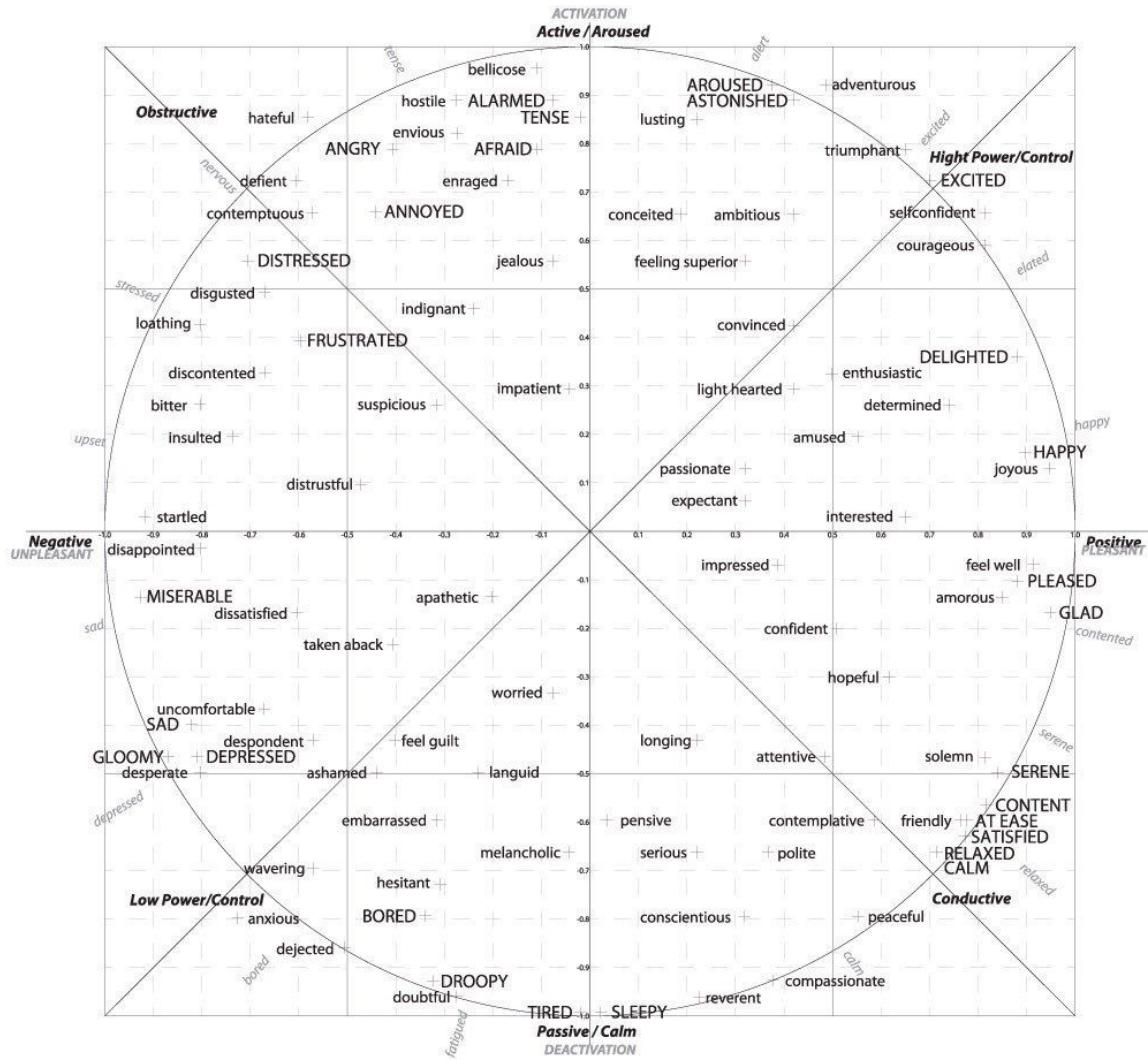
Σχήμα 3.1: Εικονική αναπαράσταση του μοντέλου των 7 βασικών συναισθημάτων¹.

3.1.2 Διαστατικό Μοντέλο

Η ανάγκη για πιο ακριβή περιγραφή των ανθρώπινων συναισθημάτων έστρεψε πολλές έρευνες σχετικές με την αναγνώριση συναισθημάτων στην υιοθετήση του διαστατικού μοντέλου αναπαράστασης συναισθημάτων [46, 47] παρέχοντας τη δυνατότητα αναπαράστασης λιγότερο έντονων συναισθημάτων, τα οποία συναντώνται συχνότερα στην καθημερινή αλληλεπίδραση ανθρώπου υπολογιστή. Το μοντέλο αυτό περιγράφει κάθε ανθρώπινο συναίσθημα απεικονίζοντάς το σε ένα χώρο δύο διαστάσεων, όπου ο πρώτος άξονας χαρακτηρίζει το σθένος (valence) και ο δεύτερος άξονας χαρακτηρίζει τη διέγερση (arousal). Το σθένος δείχνει κατά πόσο ένα συναίσθημα είναι θετικό ή αρνητικό και η διέγερση κατά πόσο έχει παθητικό ή ενεργητικό χαρακτήρα. Οι δύο αυτές μεταβλητές παίρνουν συνεχείς τιμές στο διάστημα $[-1, 1]$, όπου για το σθένος η τιμή -1 χαρακτηρίζει πλήρως αρνητικό το συναίσθημα ενώ το 1 πλήρως θετικό, και για τη διέγερση η τιμή -1 συμβολίζει το πλήρως παθητικό συναίσθημα και η τιμή 1 το πλήρως ενεργητικό συναίσθημα. Ο δισδιάστατος αυτό χώρος ονομάζεται ο τροχός των συναισθημάτων (Σχήμα 3.2) όπως αναφέρει στο [48] ο ψυχολόγος Robert Plutchik το 1980.

3.2 Το σύνολο δεδομένων Aff-Wild2

Η αυτόματη αναγνώριση ανθρώπινης συναισθηματικής συμπεριφοράς από υπολογιστές λαμβάνοντας οπτικά αλλά και ακουστικά σήματα είναι σημαντική για την καθημερινή αλληλεπίδραση ανθρώπου υπολογιστή. Στις περισσότερες περιπτώσεις, η αλληλεπίδραση αυτή διεξάγεται σε συνθήκες πραγματικού κόσμου, όπου το περιβάλλον, οι συνθήκες βιωτεοσκόπησης και ο εκάστοτε άνθρωπος ποικίλουν. Συνεπώς, θα ήταν εύλογο να συμπεράνουμε ότι η βάση δεδομένων στην οποία θα εκπαιδευτεί ένα τέτοιο μοντέλο αναγνώρισης ανθρώπινου συναισθήματος θα πρέπει να προσομοιώνει όσο το δυνατόν καλύτερα τις συνθήκες ενός πραγματικού κόσμου. Τις τελευ-



Σχήμα 3.2: Ο δισδιάστατος τροχός των συνασθημάτων.

ταίες δύο δεκαετίες, οι έρευνες σχετιζόμενες με την ανάλυση συναισθηματικής συμπεριφοράς έχουν επικεντρωθεί στη δημιουργία βάσεων δεδομένων όπου ζητήθηκε από τους συμμετέχοντες να υποδυθούν κάποιο από τα επτά βασικά συναισθήματα και το οποίο καταγράφηκε κάτω από πλήρως ελεγχόμενες συνθήκες. Μερικές από τις πιο ευρέως γνωστές βάσεις δεδομένων αυτού του είδους είναι οι βάσεις δεδομένων Cohn-Kanade, η οποία αρχικά παρουσιάστηκε το 2001 [49] και επεκτάθηκε το 2010 [50], MMI [51], Multi-PIE [52], BU-3D[53] και BU-4D[54]. Ένα άλλο είδος βάσεων δεδομένων που αναπτύχθηκε για την αναγνώριση συναισθήματος είναι εκείνες που υιοθέτησαν το διαστατικό μοντέλο αναπαράστασης συναισθήματος, το οποίο περιγράφηκε παραπάνω. Οι πιο γνωστές είναι οι SEMAINE [55], RECOLA [56], SEWA² και AFEW-VA [57].

Οι βάσεις δεδομένων που προαναφέρθηκαν, παρουσιάζουν κάποιους περιορισμούς σε σχέση με την αναπαράσταση της ποικιλομορφίας του πραγματικού κόσμου, όπως η καταγραφή τους σε ελεγχόμενο περιβάλλον, το περιορισμένο πλήθος ανθρώπων που απεικονίζονται, οι καλές συνθήκες καταγραφής με συγκεκριμένη οπτική και ομοιόμορφο φωτισμό αλλά και η μειωμένη χρονική διάρκεια δράσης κάθε ατόμου. Η βάση δεδομένων Aff-Wild [58] και η εμπλουτισμένη έκδοσή του, Aff-Wild2 [25], σκοπεύουν στην αντιμετώπιση των παραπάνω περιορισμών, δημιουργώντας μία

² <https://sewaproject.eu/>

βάση δεδομένων με μεγάλη συλλογή από βίντεο που έχουν καταγραφεί σε πραγματικές συνθήκες ή αλλιώς ”in-the-wild”, όπως αναφέρεται χαρακτηριστικά.

Οι βάσεις δεδομένων Aff-Wild και Aff-Wild2 αποτελούνται κατά κύριο λόγο από βίντεο που έχουν αναρτηθεί στη διαδικτυακή πλατφόρμα YouTub³και απεικονίζουν ανθρώπους διάφορων εθνικοτήτων και ηλικιών να αντιδρούν σε κάποιο ερεθίσμα, όπως για παράδειγμα σε μία απρόσμενη τροπή μίας τηλεοπτικής σειράς ή ταινίας. Στο Σχήμα 3.3 παρουσιάζονται μία σειρά από στιγμιότυπα των βίντεο που συμπεριλαμβάνονται στη βάση δεδομένων Aff-Wild2, όπως γίνεται αντιληπτό εκτός από τις διάφορες εθνικότητες και ηλικιακές ομάδες των ατόμων, τα βίντεο ποικίλουν και σε σχέση με το φωτισμό, το φόντο, την οπτική γωνία και την ανάλυση της καταγραφής με αποτέλεσμα να προσομοιώνουν όσο το δυνατόν καλύτερα τις πραγματικές συνθήκες στις οποίες μπορεί να κληθεί ένας υπολογιστής να αναγνωρίσει συναισθήματα ανθρώπων.



Σχήμα 3.3: Διάφορα στιγμιότυπα των βίντεο του Aff-Wild2

Συγκεκριμένα, η βάση δεδομένων Aff-Wild2 αποτελείται από 564 βίντεο με μέσο όρο ταχύτητας δειγματοληψίας 30fps (frames per second - καρέ το δευτερόλεπτο) και συνολική διάρκεια σχεδόν 60 ώρες. Στα βίντεο έχουν τοποθετηθεί ετικέτες για τα τρία μοντέλα αναπαράστασης συναισθημάτων, το διαστατικό μοντέλο όπου οι ετικέτες έχουν συνεχή ποσοτική τιμή για τους άξονες του σθένους και της διέγερσης (VA Set), το μοντέλο με τα επτά βασικά συναισθήματα όπου οι ετικέτες παίρνουν ποιοτικές τιμές (Expr Set) και το μοντέλο με τις οκτώ βασικές κινήσεις του προσώπου (AU Set), στο οποίο δεν έχουμε αναφερθεί καθώς δεν θα μας απασχολήσει στη συνέχεια. Το καθένα από τα παραπάνω υποσύνολα της βάσης χωρίζονται σε σύνολο εκπαίδευσης (training set), σύνολο αξιολόγησης (validation set) και σύνολο δοκιμής (test set), με τρόπο ώστε να μην υπάρχει το ίδιο βίντεο σε πάνω από ένα από τα σύνολα αυτά. Στον Πίνακα 3.1 φαίνονται αναλυτικά τα στοιχεία που περιέχει η βάση δεδομένων Aff-Wild2.

3.2.1 Πρόβλημα Παλινδρόμισης Σθένους/Διέγερσης (VA Task)

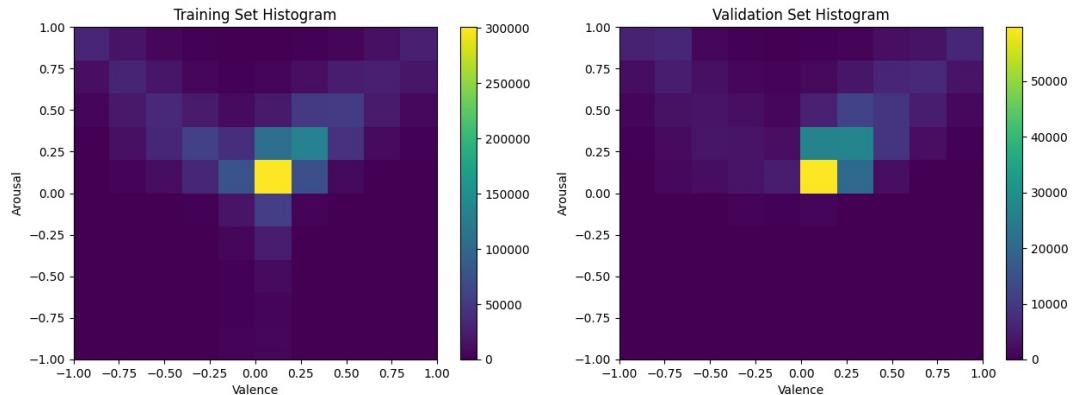
Οι ετικέτες σθένους και διέγερσης αποσκοπούν στην εκπαίδευση και στην αξιολόγηση ενός μοντέλου που λύνει ένα πρόβλημα παλινδρόμησης, καθώς οι ετικέτες του σθένους και της διέγερσης αποτελούν συνεχόμενες ποσοτικές τιμές στο διάστημα $[-1, 1]$. Για τη δημιουργία των

³ <https://www.youtube.com/>

Aff-Wild2	Πλήθος Βίντεο	Πλήθος Καρέ (frames)
VA Training Set	351	1.560.468
VA Validation Set	71	273.811
Expr Training Set	253	922.030
Expr Validation Set	70	319.324

Πίνακας 3.1: Στατιστικά Στοιχεία Aff-Wild2

ετικετών αυτών συμμετείχαν τέσσερις σχολιαστές, οι οποίοι παρακολούθησαν όλα τα βίντεο και με τη βοήθεια ενός ψηφιακού εργαλείου προσθήκης ετικέτας προσδιόρισαν τις τιμές του σθένους και της διέγερσης. Οι τελικές ετικέτες αποτελούν το μέσο όρο των τιμών που έδωσαν οι τέσσερις σχολιαστές. Η κατανομή των τιμών των ετικετών των συνόλων εκπαίδευσης και αξιολόγησης σε σχέση με το πλήθος των καρέ φαίνεται στο δισδιάστατο ιστόγραμμα στο Σχήμα 3.4. Παρατηρώντας το Ιστόγραμμα, συμπεραίνουμε πως το μεγαλύτερο μέρος των καρέ έχουν τιμές σθένους και διέγερσης στο διάστημα $[0, 0.25]$, επίσης οι θετικές τιμές και για τις δύο ποσότητες επικρατούν έναντι των αρνητικών. Συνεπώς, το σύνολο δεδομένων είναι μη ισορροπημένο και το γεγονός αυτό μπορεί να αποτελέσει δυσκολία κατά τη διαδικασία εκπαίδευσης του μοντέλου.



Σχήμα 3.4: Δισδιάστατο Ιστόγραμμα των τιμών του σθένους και της διέγερσης στα σύνολα εκπαίδευσης (αριστέρα) και αξιολόγησης (δεξιά).

3.2.2 Πρόβλημα Ταξινόμησης Βασικών Συναισθημάτων (Expressions Task)

Οι ετικέτες που αφορούν τα επτά βασικά συναισθήματα σκοπεύουν στην εκπαίδευση και στην αξιολόγηση ενός μοντέλου ταξινόμησης, καθώς οι ετικέτες αποτελούν ποιοτικές τιμές. Αναλυτικά, τρεις σχολιαστές παρακολούθησαν τα βίντεο και χαρακτήρισαν το συναίσθημα που απεικονίζεται σε κάθε frame κατατάσσοντάς το σε μία από τις επτά βασικές κατηγορίες ή κλάσεις συναισθημάτων, αν το εικονιζόμενο συναίσθημα δεν ανήκε σε κάποια από τις κατηγορίες του δινόταν μια άλλη τιμή. Οι ετικέτες που επιλέχτηκαν είναι αυτές για τις οποίες συμφωνούσαν και οι τρεις σχολιαστές. Οι ετικέτες πάρονταν τις τιμές 0, 1, 2, 3, 4, 5, 6 οι οποίες αντιστοιχούν

στα συναισθήματα Ουδετερότητα (Neutral), Θυμός (Anger), Αηδία (Disgust), Φόβος (Fear), Χαρά (Happiness), Λύπη (Sadness), Έκπληξη (Surprise). Όταν το εικονιζόμενο συναισθήμα δεν ανήκει σε κάποια από τις παραπάνω κλάσεις παίρνει την τιμή -1. Στο Σχήμα 3.5 φαίνεται η κατανομή των κλάσεων σε σχέση με το πλήθος των frames που τους αντιστοιχούν για τα σύνολα εκπαίδευσης και αξιολόγησης. Είναι φανερό ότι το σύνολο δεδομένων δεν είναι ισορροπημένο και αυτό μπορεί να αποτελέσει εμπόδιο κατά την εκπαίδευση του μοντέλου ταξινόμησης. Ωστόσο, η κατανομή των δεδομένων, δηλαδή η μεγάλη συχνότητα απεικόνισης του Ουδέτερου συναισθήματος, είναι απολύτως λογική και προσομοιώνει την πραγματική ζωή.



Σχήμα 3.5: Ιστόγραμμα της κατανομής των κλάσεων των επτά βασικών Συναισθημάτων των συνόλων εκπαίδευσης (πάνω) και αξιολόγησης (κάτω).

3.3 Προεπεξεργασία Δεδομένων

Η βάση δεδομένων Aff-Wild2 παρέχει δεδομένα σε μορφή βίντεο όπως προαναφέραμε. Η πλειοψηφία των βίντεο προσφέρουν πληροφορία σε οπτική αλλά και ακουστική μορφή. Συνεπώς, είναι εύλογο να εκμεταλλευτούμε και τα δύο αυτά είδη σημάτων έτσι ώστε να εκπαιδεύσουμε με μεγαλύτερη επιτυχία τα μοντέλα παλινδρόμησης του σθένους και της διέγερσης και ταξινόμησης των επτά βασικών συναισθημάτων. Βέβαια, για την τροφοδότηση των οπτικοακουστικών

δεδομένων στα μοντέλα μηχανικής μάθησης που θα αναλύσουμε στη συνέχεια, είναι αναγκαία η προεπεξεργασία τους. Τα βήματα προεπεξεργασίας που ακολουθήθηκαν αναλύονται στις παρακάτω υποενότητες.

3.3.1 Προεπεξεργασία Οπτικών Δεδομένων

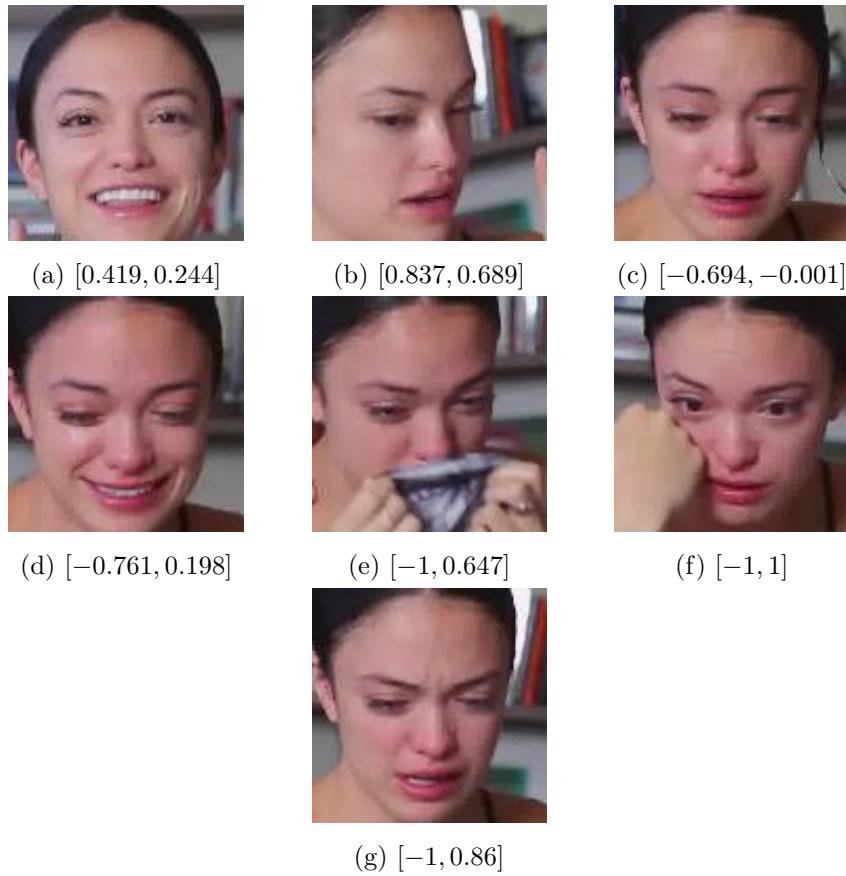
Αρχικά πραγματοποιήθηκε η μετατροπή των βίντεο σε μία σειρά από εικόνες χρησιμοποιώντας το λογισμικό Menpo [59]. Σε κάθε εικόνα των βίντεο έγινε η ανίχνευση των προσώπων τοποθετώντας πλαίσια οριοθέτησης (bounding boxes) στη περιοχή του προσώπου κάνοντας χρήση του ανιχνευτή προσώπων SSH (Single Stage Headless) [60]. Ο ανιχνευτής προσώπων SSH βασίζεται σε ένα δίκτυο ResNet και έχει εκπαιδευτεί στο σύνολο δεδομένων WiderFace [61]. Έπειτα, έγινε η εξαγωγή των πέντε βασικών σημείων του κάθε προσώπου (τα κέντρα των ματιών, η μύτη και οι δύο άκρες του στόματος), τα οποία στη συνέχεια χρησιμοποιήθηκαν για την σωστή ευθυγράμμιση του προσώπου. Κατά τη διάρκεια των δύο παραπάνω βημάτων προεπεξεργασίας, αφαιρέθηκαν frames στα οποία δεν είχαν τοποθετηθεί με επιτυχία τα πλαίσια οριοθέτησης του προσώπου ή δεν είχε γίνει σωστή ανίχνευση των πέντε βασικών σημείων του προσώπου. Ο έλεγχος αυτός πραγματοποιήθηκε με δύο τρόπους, ο πρώτος τρόπος ήταν με την αναζήτηση σημαντικών αλλαγών στη θέση είτε των πλαισίων οριοθέτησης είτε των πέντε σημείων του προσώπου μεταξύ συνεχόμενων καρέ και ο δεύτερος τρόπος ήταν ο χειροκίνητος έλεγχος των εικόνων από τους σχολιαστές. Οι τελικές εικόνες των προσώπων έχουν διαστάσεις $112 \times 112 \times 3$ και είναι στη μορφή αρχείου JPEG.

Στο Σχήμα 3.6 παρουσιάζονται κάποια τελικά οπτικά δεδομένα του συνόλου δεδομένων VA, μετά τη διαδικασία προεπεξεργασίας που περιγράφηκε παραπάνω. Όπως φαίνεται η ανίχνευση του προσώπου είναι δυνατή ακόμα και όταν το πρόσωπο εμφανίζεται υπό γωνία ή όταν υπάρχουν άλλα στοιχεία μπροστά στο πρόσωπο, όπως για παράδειγμα τα χέρια. Το γεγονός αυτό είναι πολύ σημαντικό καθώς στη πραγματική ζωή είναι συχνό φαινόμενο η αλλαγή οπτικής γωνίας και η τοποθέτηση των χεριών στο πρόσωπο όταν εκφράζουμε διάφορα συναισθήματα, με αποτέλεσμα η μη ανίχνευσή τους να αποτελούσε μειονέκτημα για την εκπαίδευση ενός μοντέλου αναγνώρισης συναισθημάτων.

Παραδείγματα του συνόλου των βασικών συναισθημάτων φαίνονται στο Σχήμα 3.7 με λεζάντα την αντίστοιχη ετικέτα της κλάσης του συναισθήματος που απεικονίζουν. Παρατηρούμε ότι εδώ τη ποικιλομορφία των ατόμων που απεικονίζεται αλλά και τη διαφορά στην ανάλυση των εικόνων, το φωτισμό και το φόντο.

3.3.2 Προεπεξεργασία Ακουστικών Δεδομένων

Η ένταση και ο τόνος της φωνής αποτελεί σημαντικό παράγοντα για την επιτυχή αναγνώριση των ανθρώπινων συναισθημάτων, καθώς κατά την εμφάνισή τους η αλλαγή της έκφρασης του προσώπου συνοδεύεται πολλές φορές από επιφωνήματα ή αλλαγή στον τρόπο ομιλίας. Για την εξαγωγή των χαρακτηριστικών του ήχου και την τροφοδότησή τους σε ένα συνελικτικό νευρωνικό δίκτυο είναι αναγκαία η μετατροπή του ηχητικού σήματος σε μία οπτική αναπαράσταση. Ο τρόπος που επιλέγεται να γίνει η μετατροπή αυτή είναι η δημιουργία φασματογραφημάτων



Σχήμα 3.6: Παραδείγματα συνεχόμενων frames μετά την ανίχνευση και την ευθυγράμμιση του προσώπου με τις αντίστοιχες ετικέτες σθένους και διέγερσης ([σθένος, διέγερση]).

(spectrograms).

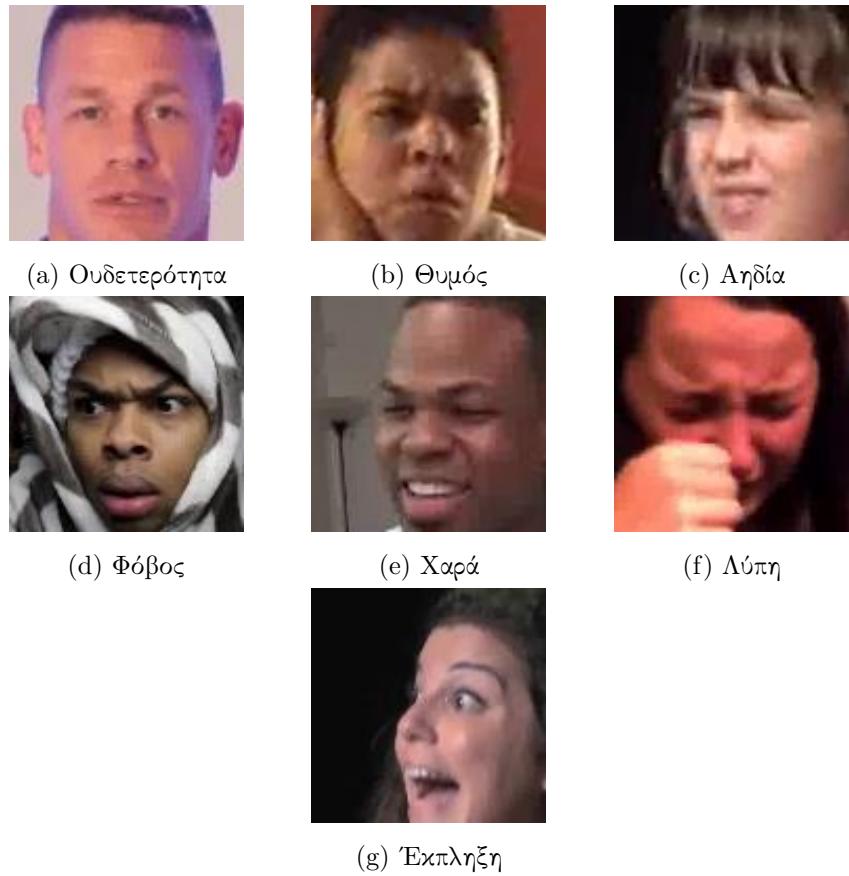
3.3.2.1 Φασματογράφημα Mel (Mel Spectrogram)

Το φασματογράφημα αποτελεί μία οπτική αναπαράσταση του φάσματος των συχνοτήτων ενός ηχητικού σήματος σε σχέση με το χρόνο. Η πιο συνήθης μορφή του είναι μία εικόνα δύο διαστάσεων όπου ο οριζόντιος άξονας είναι ο χρόνος και ο κάθετος η συχνότητα. Η ένταση του σήματος σε ένα συγκεκριμένο σημείο χρόνου και συχνότητας αναπαριστάται με διάφορους χρωματισμούς. Στο Σχήμα 3.8 φαίνεται η αναπαράσταση του ηχητικού σήματος σειρήνας πλοίου σε διάγραμμα χρόνου-πλάτους και χρόνου-συχνότητας, το τελευταίο αποτελεί το φασματογράφημα.

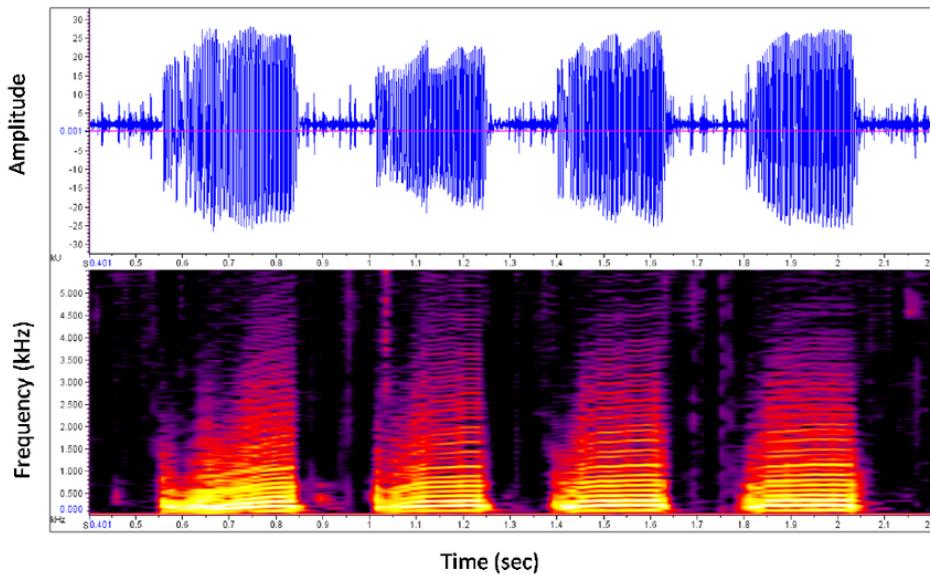
Η δημιουργία ενός φασματογραφήματος γίνεται ψηφιακά με τον υπολογισμό του μετασχηματισμού Fourier για μικρά χρονικά παράθυρα ενός ηχητικού σήματος, τα οποία συνήθως επικαλύπτονται. Συγκεκριμένα, η παραπάνω διαδικασία ονομάζεται Βραχυπρόθεσμος Μετασχηματισμός Fourier (STFT - short-time Fourier transform) (βλ. Σχήμα 3.9) και το φασματογράφημα αποτελεί το τετράγωνο του μέτρου του μετασχηματισμού αυτού, όπως φαίνεται στη παρακάτω σχέση:

$$\text{spectrogram}(s(t), w) = |\text{STFT}(s(t), w)|^2 \quad (3.1)$$

⁴ <https://www.researchgate.net/figure/Spectrograms-and-Oscillograms>

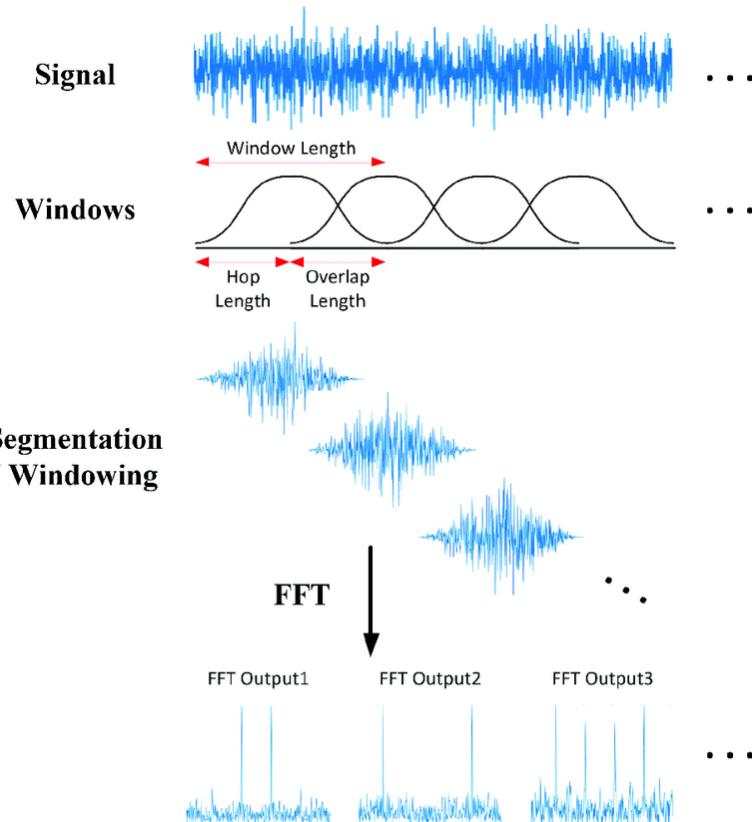


Σχήμα 3.7: Παραδείγματα οπτικών δεδομένων του συνόλου Basic Expressions μετά την ανίχνευση και την ευθυγράμμιση του προσώπου με τις αντίστοιχες ετικέτες για τα βασικά συνασθήματα.



Σχήμα 3.8: Αναπαράσταση ηχητικού σήματος στους άξονες χρόνου-πλάτους και το αντίστοιχο φασματογράφημα⁴.

όπου $s(t)$ ηχητικό σήμα στο πεδίο του χρόνου και w το χρονικό παράθυρο του μετασχηματισμού.



Σχήμα 3.9: Διαδικασία Βραχυπρόθεσμου Μετασχηματισμού Fourier (STFT)⁵.

Για την καλύτερη παρατήρηση των φασματογραφημάτων από τον άνθρωπο γίνεται μετατροπή του άξονα των συχνοτήτων από γραμμικό σε λογαριθμικό καθώς και οι τιμές του πλάτους μετατρέπονται σε decibels (dB). Με αυτό τον τρόπο τα φασματογραφήματα απεικονίζουν καλύτερα τους ήχους που μπορεί να συλλάβει το ανθρώπινο αυτί, οι οποίοι είναι συγκεντρωμένοι σε ένα μικρό εύρος συχνοτήτων και πλατών. Επίσης, οι άνθρωποι δεν συλλαμβάνουν ήχους με γραμμική κλίμακα, δηλαδή έχουν μεγαλύτερη ευαισθησία και ικανότητα να διακρίνουν ήχους χαμηλών συχνοτήτων παρά υψηλών. Για τον λόγο αυτό έχει εφευρεθεί μία κλίμακα, γνωστή ως κλίμακα Mel, η οποία μετατρέπει τον άξονα των συχνοτήτων με τέτοιο τρόπο ώστε να ακολουθεί την ευαισθησία του ανθρώπινου αυτιού. Μετά τις παραπάνω μετατροπές ουσιαστικά δημιουργούμε το Φασματογράφημα Mel (Mel Spectrogram), το οποίο παρέχει όλα τα χρήσιμα χαρακτηριστικά του ακουστού από τον άνθρωπο ήχου και αυτό θα χρησιμοποιήσουμε για την εκπαίδευση του νευρωνικού δικτύου αναγνώρισης συναισθήματος στη συνέχεια.

3.3.2.2 Εξαγωγή Φασματογραφημάτων Mel

Με στόχο τη δημιουργία των φασματογραφημάτων, αρχικά εξάγουμε το ηχητικό σήμα από τα βίντεο, μετατρέποντάς τα σε αρχεία WAV. Όλα τα αρχεία ήχου έχουν ρυθμό δειγματοληψίας $44100Hz$. Έπειτα χωρίζουμε το ηχητικό σήμα από κάθε βίντεο σε παράθυρα των 2 δευτερολέπτων με βήμα ενός δευτερολέπτου. Το κάθε παράθυρο ευθυγραμμίζεται με τέτοιο τρόπο ώστε το κέντρο του να αντιστοιχεί σε ένα frame του βίντεο και στην αντίστοιχη ετικέτα. Μετά το διαχωρισμό των παραθύρων, εξάγουμε το αντίστοιχο φασματογράφημα Mel που αναπαριστά το κάθε

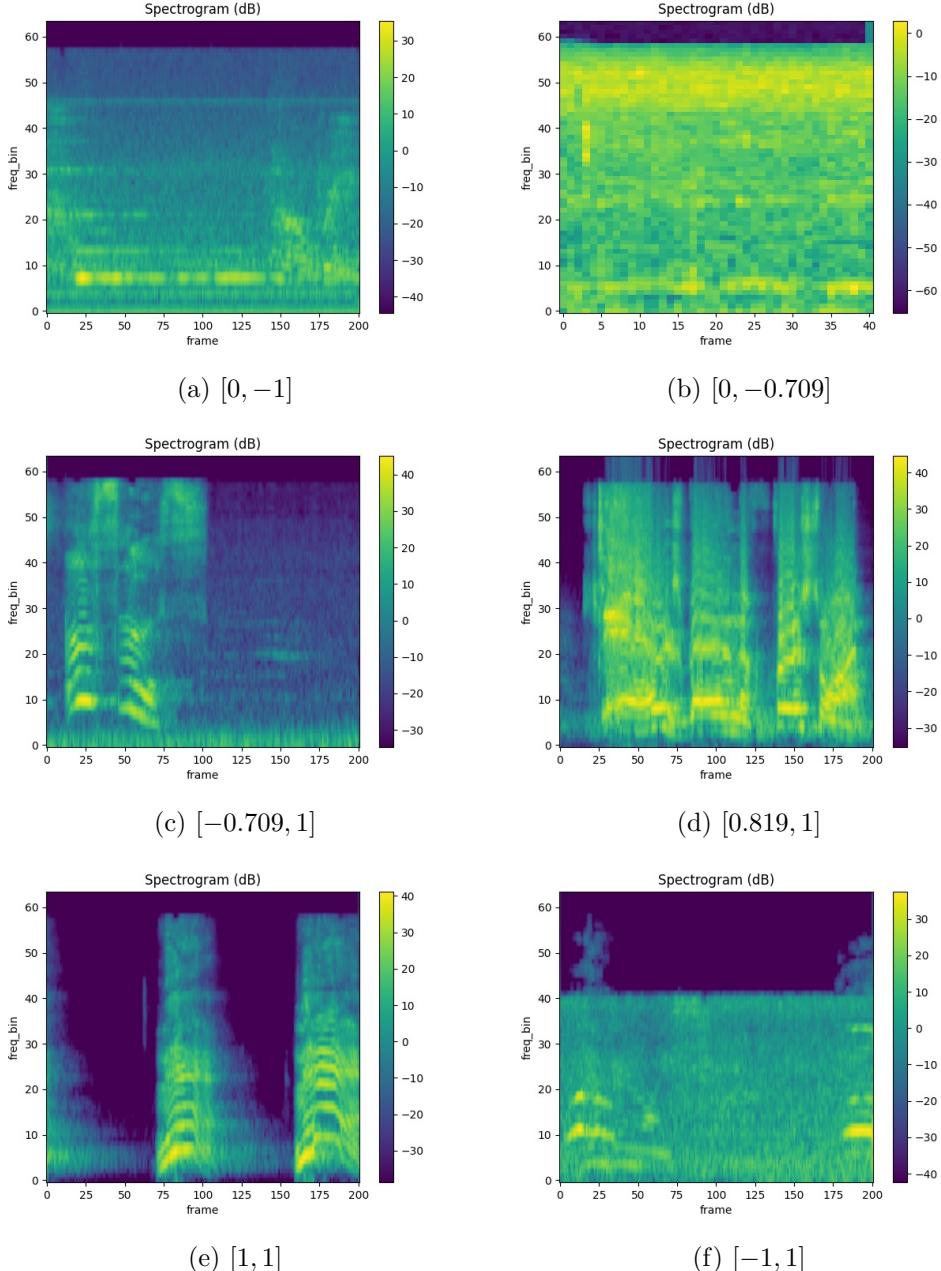
⁵ https://www.researchgate.net/figure/Short-time-Fourier-transform-STFT-overview_fig1_346243843

ένα από τα παράθυρα. Για τη δημιουργία των φασματογραφημάτων χρησιμοποιούμε το πακέτο επεξεργασίας ηχητικών σημάτων TorchAudio, το οποίο είναι μέρος της βιβλιοθήκης PyTorch [62]. Συγκεκριμένα, η συνάρτηση που χρησιμοποιήθηκε για την μετατροπή του ηχητικού σήματος σε φασματογράφημα Mel ονομάζεται *MelSpectrogram()* και οι παράμετροι είναι οι εξής:

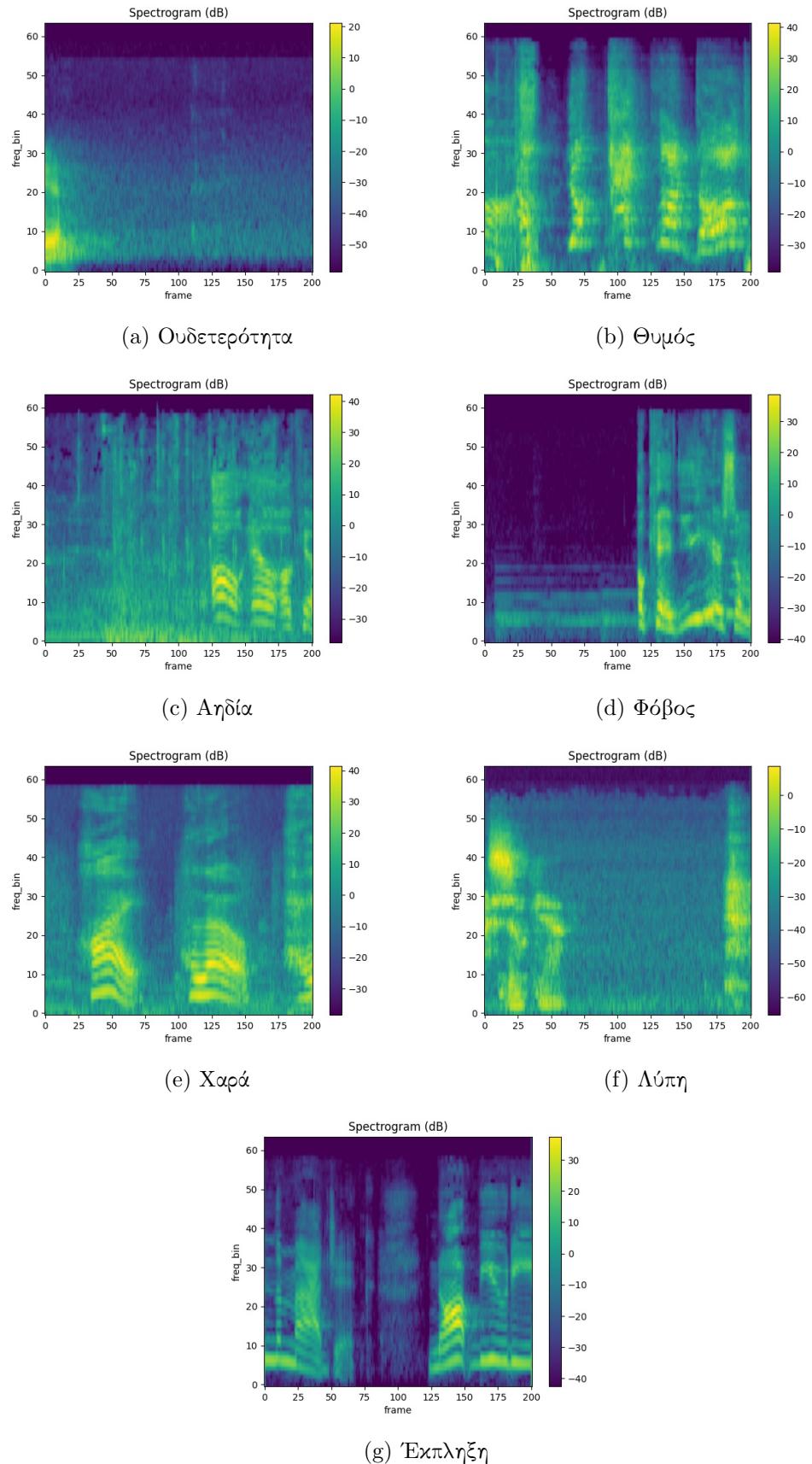
- μήκος παραθύρου STFT: $w_{win} = 20 \text{ ms}$
- μήκος βήματος STFT: $t_{stride} = 10 \text{ ms}$
- μέγεθος FFT: $n_{fft} = 1024 \text{ bins}$
- αριθμός ζωνών του φίλτρου Mel: $n_{mels} = 64$

Μετά την εξαγωγή του φασματογραφήματος με τη χρήση της παραπάνω συνάρτησης, κάνουμε τη μετατροπή του πλάτους στη κλίμακα dB χρησιμοποιώντας την συνάρτηση *AmplitudeToDB()*, βάζοντας άνω όριο τα 80 dB, τα οποία συμπεριλαμβάνουν όλους τους ήχους που είναι χρήσιμοι για την αναγνώριση των συναισθημάτων [63]. Τα τελικά φασματογραφήματα είναι σε μορφή τένσορα και έχουν διαστάσεις $1 \times 201 \times 64$.

Στα παρακάτω σχήματα (Σχήμα 3.10 και Σχήμα 3.11) παρουσιάζονται κάποια παραδείγματα των φασματογραφημάτων που δημιουργήσαμε με τις ετικέτες που τους αντιστοιχούν. Όσο αναφορά τα φασματογραφήματα με τις ετικέτες σθένους και διέγερσης, παρατηρούμε ότι με τη χαμηλότερη τιμή διέγερσης έχουμε και λιγότερες εντάσεις και με χαμηλότερο πλάτος (dB), γεγονός λογικό καθώς όσο πιο παθητικό είναι ένα συναίσθημα συνήθως εκφράζεται με λιγότερο δυνατούς ήχους, ενώ όσο η τιμή της διέγερσης αυξάνεται παρατηρούμε μεγαλύτερες διακυμάνσεις τόσο στη συχνότητα όσο και στην ένταση του ήχου. Στη περίπτωση των φασματογραφημάτων με ετικέτες τα βασικά συναισθήματα, διαπιστώνουμε ότι συναισθήματα λιγότερης έντασης, όπως η Ουδετερότητα και η Λύπη, έχουν χαμηλή ένταση ήχου ή και καθόλου ήχο. Συναισθήματα, όμως, όπως η Χαρά, ο Θυμός και η Έκπληξη, εμφανίζουν υψηλές εντάσεις ήχων (40dB) και διαφορετικές συχνότητες.



Σχήμα 3.10: Παραδείγματα Φασματογραφημάτων με τις αντίστοιχες επικέτες σθένους και διέγερσης ([σθένος, διέγερση]).



Σχήμα 3.11: Παραδείγματα Φασματογραφημάτων με τις αντίστοιχες ετικέτες για τα βασικά συναισθήματα.

Κεφάλαιο 4

Μεθοδολογία

Στο Κεφάλαιο αυτό θα αναλύσουμε τα μοντέλα τα οποία θα χρησιμοποιηθούν για την αναγνώριση των συναισθημάτων στο πλαίσιο της παρούσας διπλωματικής, καθώς και απαραίτητα συστατικά για την εκπαίδευσή και την αξιολόγησή τους, όπως είναι οι μετρικές αξιολόγησης και οι συναρτήσεις κόστους.

4.1 Μετρικές Αξιολόγησης Μοντέλων

Η επιλογή των μετρικών αξιολόγησης των μοντέλων είναι πολύ σημαντική καθώς έτσι μπορούμε να παραχολουθήσουμε την εκπαίδευση και να κάνουμε τις απαραίτητες αλλαγές των παραμέτρων των μοντέλων έτσι ώστε να πετύχουμε το καλύτερο δυνατό αποτέλεσμα. Επιπρόσθετα, είναι σημαντική η επιλογή μετρικών που αρμόζουνε στο πρόβλημα και χρησιμοποιούνται και από άλλες ερευνητικές ομάδες έτσι ώστε να μπορεί να γίνεται σύγχριση της επίδοσης των μοντέλων στο πλαίσιο της ευρύτερης ερευνητικής δραστηριότητας.

4.1.1 Μετρική Αξιολόγησης Προβλήματος Παλινδρόμησης

Για την αξιολόγηση των μοντέλων του προβλήματος παλινδρόμησης των τιμών του σθένους και της διέγερσης των συναισθημάτων θα χρησιμοποιήσουμε το Συντελεστή Συμφωνίας Συσχέτισης (Concordance Correlation Coefficient - CCC) [64], ο οποίος χρησιμοποιείται συχνά για τη μέτρηση της επίδοσης μοντέλων αναγνώρισης συναισθημάτων. Ο συντελεστής CCC εκτιμά τη συμφωνία μεταξύ δύο χρονοσειρών αλλάζοντας τη κλίμακα του συντελεστή συσχέτισής τους με τη μέση τετραγωνική διαφορά τους. Η σχέση από την οποία υπολογίζεται ο CCC είναι:

$$\rho_c = \frac{2\sigma_x\sigma_y\rho_{xy}}{\sigma_x^2 + \sigma_y^2 + (\mu_x + \mu_y)^2} \quad (4.1)$$

όπου σ_x, σ_y οι διακυμάνσεις των ετικετών του σθένους ή της διέγερσης και των προβλεπόμενων από το μοντέλο τιμών αντίστοιχα, ρ_{xy} ο Συντελεστής Συσχέτισης Pearson (Pearson CC) και μ_x, μ_y οι μέσες τιμές.

Ο συντελεστής CCC παίρνει τιμές στο διάστημα $[-1, 1]$, με την τιμή 1 να συμβολίζει την απόλυτη συμφωνία δύο χρονοσειρών και την τιμή -1 την απόλυτη ασυμφωνία. Στη περίπτωση της αξιολόγησης του μοντέλου παλινδρόμησης επιθυμούμε τιμές όσο πιο κοντά στη μονάδα έτσι

ώστε να υπάρχει συμφωνία μεταξύ των επικετών και των προβλέψεων. Η συνολική αξιολόγηση του μοντέλου γίνεται με το μέσο όρο των συντελεστών CCC του σθένους (CCC-V) και της διέγερσης (CCC-A).

4.1.2 Μετρική Αξιολόγησης Προβλήματος Ταξινόμησης

Η αξιολόγηση των μοντέλων ταξινόμησης των επτά βασικών συναισθημάτων τόσο κατά την εκπαίδευση όσο και στην τελική αξιολόγηση γίνεται με το συνδυασμό των μετρικών της ακρίβειας (Accuracy) και του F1-Score. Συγκεκριμένα η μετρική που χρησιμοποιείται δίνεται από τη σχέση:

$$Metric = 0.67 * F1-Score + 0.33 * Accuracy \quad (4.2)$$

Η ακρίβεια αποτελεί την πιο συχνά χρησιμοποιούμενη μετρική για προβλήματα ταξινόμησης και υπολογίζεται από τη σχέση:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.3)$$

όπου:

- **TP:** *True Positive*, το σύνολο των προβλέψεων που είναι σωστές και ανήκουν στην κλάση για την οποία υπολογίζουμε την ακρίβεια.
- **TN:** *True Negative*, το σύνολο των προβλέψεων που είναι σωστές και δεν ανήκουν στην κλάση για την οποία υπολογίζουμε την ακρίβεια.
- **FP:** *False Positive*, το σύνολο των προβλέψεων που είναι λανθασμένες και ανήκουν στην κλάση για την οποία υπολογίζουμε την ακρίβεια.
- **FN:** *False Negative*, το σύνολο των προβλέψεων που είναι λανθασμένες και δεν ανήκουν στην κλάση για την οποία υπολογίζουμε την ακρίβεια.

Η ακρίβεια εκτός από τη σχέση 4.3, μπορεί να υπολογιστεί και πρακτικά προσθέτοντας όλες τις σωστές προβλέψεις για όλες τις κλάσεις και διαιρώντας με το συνολικό αριθμό των προβλέψεων. Η συγκεκριμένη μετρική δεν μπορεί να δείξει την απόδοση του μοντέλου για κάθε κλάση ξεχωριστά και ειδικά στη περίπτωση σημαντικά μη ισορροπημένου συνόλου δεδομένων, όπου μία κλάση μπορεί να έχει μεγάλο πλήθος παραδειγμάτων, δεν είναι αρκετά ενδεικτική για την επιτυχία του μοντέλου σε κλάσεις με μικρότερο αριθμό παραδειγμάτων. Για τον λόγο αυτό επιλέγουμε να δώσουμε χαμηλότερη βαρύτητα σε αυτή τη μετρική για την αξιολόγηση των μοντέλων μας.

Η μετρική F1-Score αξιολογεί την απόδοση ενός μοντέλου ταξινόμησης συνδυάζοντας τις μετρικές Precision (Ακρίβεια) και Recall (Ανάκληση). Οι ορισμοί των δύο αυτών μετρικών φαίνονται στη συνέχεια.

$$Precision = \frac{TP}{TP + FP} \quad (4.4)$$

$$Recall = \frac{TP}{TP + FN} \quad (4.5)$$

Όταν έχουμε τη περίπτωση της ταξινόμησης πολλαπλών κλάσεων οι μετρικές Precision και Recall υπολογίζονται ξεχωριστά για κάθε κλάση με τη λογική του "ενός εναντίον όλων".

Κάθε φορά που γίνεται ο υπολογισμός για μία κλάση αυτή έχει την ετικέτα της θετικής κλάσης (positive) ενώ όλες οι υπόλοιπες μαζί έχουν την ετικέτα της αρνητικής κλάσης (negative).

Η μετρική F1-score είναι η αρμονική μέση τιμή των μετρικών Precision και Recall και υπολογίζεται από την παρακάτω σχέση:

$$F1\text{-Score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (4.6)$$

Το F1-Score παίρνει τιμές στο διάστημα $[0, 1]$ με την τιμή 1 να δείχνει την καλύτερη απόδοση και την τιμή 0 τη χειρότερη. Ο τύπος F1-Score που χρησιμοποιείται εδώ λέγεται Macro F1-Score και ουσιαστικά υπολογίζεται από τις μη σταθμισμένες τιμές των Precision και Recall, οι οποίες δεν εξαρτώνται από το πλήθος των παραδειγμάτων κάθε κλάσης. Το F1-Score μπορεί να υπολογιστεί για κάθε κλάση ξεχωριστά αλλά και συνολικά παίρνοντας το μέσο όρο των επιμέρους τιμών. Η μετρική αυτή λοιπόν δεν επηρεάζεται από τον πληθυσμό της κάθε κλάσης και μας δίνει μία καλύτερη εικόνα της απόδοσης του ταξινομητή ακόμα και για τις κλάσεις που έχουν μικρό πλήθος παραδειγμάτων, συνεπώς σε μη ισορροπημένα σύνολα δεδομένων είναι εύλογο να του δίνεται μεγαλύτερη βαρύτητα από την απλή ακρίβεια που αναλύσαμε παραπάνω.

4.2 Συναρτήσεις Κόστους (Loss Functions)

Οι συναρτήσεις κόστους παίζουν πολύ σημαντικό ρόλο στην εκπαίδευση των νευρωνικών δικτύων, επειδή επιβραβεύουν το δίκτυο όταν κάνει σωστές προβλέψεις και του επιβάλλουν πονές όταν κάνει λανθασμένες. Συνεπώς, επιλέγοντας τις κατάλληλες συναρτήσεις κόστους για το εκάστοτε πρόβλημα συμβάλει κατά πολύ στην επιτυχημένη εκπαίδευση του δικτύου και την καλύτερη απόδοσή του.

4.2.1 Συνάρτηση Κόστους Προβλήματος Παλινδρόμησης

Η συνάρτηση κόστους που χρησιμοποιούμε για την εκπαίδευση των μοντέλων πρόβλεψης των τιμών του σθένους και της διέγερσης βασίζεται στη μετρική αξιολόγησης των αντίστοιχων μοντέλων CCC, η οποία περιγράφεται εκτενώς στην υποενότητα 4.1.2. Όπως έχουμε προαναφέρει το μοντέλο αξιολογείται με βάση το μέσο όρο των συντελεστών CCC για το σθένος και τη διέγερση, συνεπώς με παρόμοιο τρόπο δημιουργούμε και τη συνάρτηση κόστους, η οποία φαίνεται στη συνέχεια:

$$L_{total} = 1 - \frac{\rho_v + \rho_a}{2} \quad (4.7)$$

όπου οι μεταβλητές ρ_v και ρ_a συμβολίζουν τον συντελεστή CCC για το σθένος και τη διέγερση αντίστοιχα και υπολογίζονται από τη σχέση 4.1.

Στις περισσότερες περιπτώσεις προβλημάτων παλινδρόμησης χρησιμοποιείται ως συνάρτηση κόστους η μέθοδος ελαχίστων τετραγώνων (Mean Square Error - MSE), αλλά παρατηρώντας τα αποτελέσματα για τις δύο συναρτήσεις κόστους τις οποίες χρησιμοποίησαν στη δημοσίευση [65] (Πίνακες 7 και 9) είναι προφανές ότι η επίδοση των μοντέλων για το συγκεκριμένο πρόβλημα είναι

μακράν καλύτερη με τη συνάρτηση κόστους CCC. Για το λόγο αυτό, στη παρούσα διπλωματική θα χρησιμοποιήσουμε μόνο τη συνάρτηση κόστους που είναι βασισμένη στο συντελεστή CCC.

4.2.2 Συνάρτηση Κόστους Προβλήματος Ταξινόμησης

Η Συνάρτηση Κόστους Εγκάρσιας Εντροπίας (Cross Entropy Loss Function) αποτελεί συνήθη μέθοδος για την εκπαίδευση μοντέλων ταξινόμησης πολλαπλών κλάσεων. Με την μέθοδο αυτή συγχρίνεται η πιθανότητα πρόβλεψης της εξόδου του μοντέλου σε σχέση με την αναμενόμενη κλάση και υπολογίζεται το κόστος με βάση την απόσταση μεταξύ τους, το οποίο στη συνέχεια επιβάλει την αντίστοιχη ποινή στο μοντέλο. Το κόστος υπολογίζεται για κάθε κλάση ξεχωριστά και στη συνέχεια χρησιμοποιείται ο απλός ή ο σταθμισμένος μέσος όρος για την τελική εκτίμηση του κόστους του μοντέλου. Συγκεκριμένα, το κόστος για κάθε κλάση υπολογίζεται ως εξής:

$$loss(x, class) = -\log\left(\frac{e^{x[class]}}{\sum_j e^{x[j]}}\right) = -x[class] + \log\left(\sum_j e^{x[j]}\right) \quad (4.8)$$

όπου x είναι η έξοδος του μοντέλου πριν εφαρμοστεί η συνάρτηση Softmax και έχει διάσταση (*minibatch, C*), όπου C είναι το πλήθος των κλάσεων, με $x[class]$ συμβολίζεται η πιθανότητα πρόβλεψης της αναμενόμενης κλάσης και με $x[j]$ η πιθανότητες για τις υπόλοιπες κλάσεις.

Όταν το σύνολο δεδομένων είναι μη ισορροπημένο, όπως συμβαίνει και στη περίπτωσή μας, δίνεται η δυνατότητα να υπολογίσουμε το κόστος σύμφωνα με τη κατανομή των κλάσεων, παρέχοντας μία τιμή που να δείχνει το βάρος κάθε κλάσης. Τότε το κόστος κάθε κλάσης υπολογίζεται με την παρακάτω σχέση:

$$loss(x, class) = weight[class]\left(-x[class] + \log\left(\sum_j e^{x[j]}\right)\right) \quad (4.9)$$

όπου $weight[class]$ συμβολίζει το βάρος της αναμενόμενης κλάσης. Το συνολικό κόστος για τις 7 κλάσεις των βασικών συναισθημάτων όταν χρησιμοποιούμε τα βάρη είναι ο σταθμισμένος μέσος όρος και υπολογίζεται ως εξής:

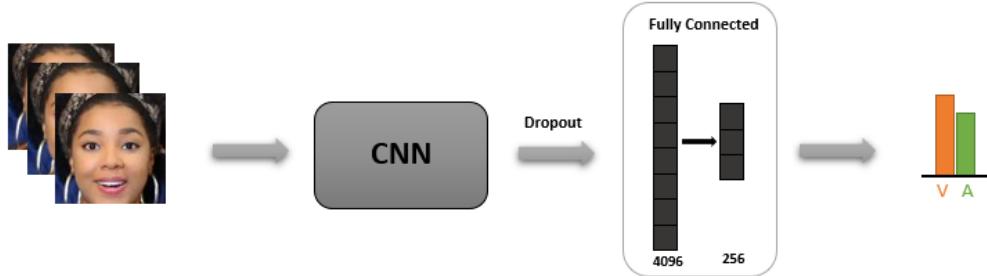
$$L_{CE} = \frac{\sum_{i=1}^7 loss(i, class[i])}{\sum_{i=1}^7 weight[class[i]]} \quad (4.10)$$

4.3 Προτεινόμενα Μοντέλα Οπτικής Αναγνώρισης

Στην ενότητα αυτή θα παρουσιαστούν τα μοντέλα που προτείνονται για τους δύο τύπους προβλημάτων αναγνώρισης συναισθήματος, τα οποία έχουν ως είσοδο μόνο το οπτικό σήμα, δηλαδή τις εικόνες των προσώπων που έχουν προκύψει από την προεπεξεργασία των βίντεο του συνόλου δεδομένων Aff-Wild2. Η διαδικασία της προεπεξεργασίας των δεδομένων έχει αναλυθεί στην υποενότητα 3.3.1. Για την επίλυση των δύο προβλημάτων αναγνώρισης συναισθήματος, της παλινδρόμησης σθένους/διέγερσης και της ταξινόμησης των βασικών συναισθημάτων, αναπτύσσουμε δύο διαφορετικά μοντέλα τα οποία αποτελούνται από ένα συνελικτικό νευρωνικό δίκτυο

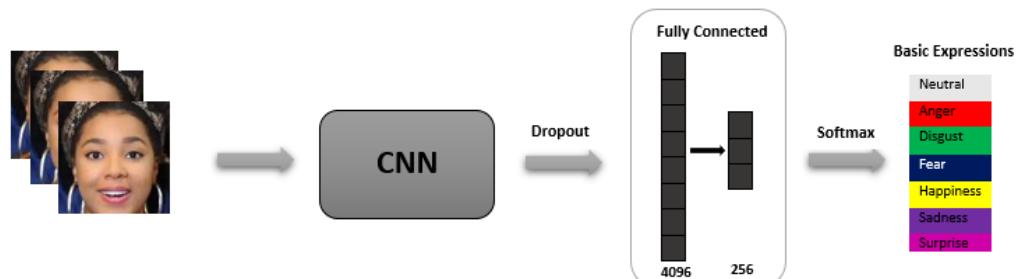
και ένα συνδυασμό πλήρως συνδεδεμένων επιπέδων που οδηγεί στις προβλέψεις. Τα συνελικτικά δίκτυα που επιλέγονται να χρησιμοποιηθούν αναλύονται στις παρακάνω υποενότητες.

Οπτικό Μοντέλο VA Στη περίπτωση του προβλήματος παλινδρόμησης, το μοντέλο μας δέχεται RGB εικόνες με διαστάσεις $3 \times 224 \times 224$ ως είσοδο, στη συνέχεια γίνεται η εξαγωγή των χαρακτηριστικών των εικόνων μέσω ενός σύγχρονου συνελικτικού δικτύου από το οποίο αφαιρείται το τελευταίο πλήρως συνδεδεμένο επίπεδο και έπειτα τοποθετούνται στη σειρά δύο πλήρως συνδεδεμένα επίπεδα με διαστάσεις εξόδου 256 και 2. Το τελευταίο πλήρως συνδεδεμένο επίπεδο κάνει τη πρόβλεψη για τις δύο τιμές του σθένους και της διέγερσης. Επίσης, μεταξύ του συνελικτικού δικτύου και των 2 τελευταίων πλήρως συνδεδεμένων επιπέδων τοποθετείται ένα επίπεδο Dropout (βλ. Ενότητα 2.3.7) για την αποφυγή του προβλήματος υπερπροσαρμογής (overfitting) του δικτύου. Στο Σχήμα 4.1 περιγράφεται και σχηματικά το προτεινόμενο μοντέλο.



Σχήμα 4.1: Αρχιτεκτονική Προτεινόμενου Οπτικού Μοντέλου για το πρόβλημα Παλινδρόμησης Σθένους/Διέγερσης (VA).

Οπτικό Μοντέλο Basic Expressions Για την επίλυση του προβλήματος ταξινόμησης των 7 βασικών συναισθημάτων αναπτύσσουμε το μοντέλο το οποίο παρουσιάζεται στο Σχήμα 4.2. Η λογική, όπως φαίνεται και στο Σχήμα, είναι ίδια με εκείνη του οπτικού μοντέλου VA με τη διαφορά ότι το τελευταίο πλήρως συνδεδεμένο επίπεδο έχει διάσταση εξόδου 7, όπου οι 7 τιμές δείχνουν την πιθανότητα η εικόνα εισόδου να απεικονίζει κάθε μία από τις 7 κλάσεις. Για την τελική πρόβλεψη της κλάσης προστίθεται ένα επίπεδο ενεργοποίησης Softmax (βλ. Ενότητα 2.3.5), το οποίο δείχνει τη κλάση με τη μεγαλύτερη πιθανότητα.



Σχήμα 4.2: Αρχιτεκτονική Προτεινόμενου Οπτικού Μοντέλου για το πρόβλημα Ταξινόμησης των 7 Βασικών Συναισθημάτων (Basic Expressions).

4.3.1 VGG-16

Το συνελικτικό νευρωνικό δίκτυο VGG-16 επιλέγεται ως βάση για την ανάπτυξη των μοντέλων αναγνώρισης συναισθήματος από εικόνες. Τα δίκτυα VGG ζεκινούν από τα 11 επίπεδα και φτάνουν τα 19. Το δίκτυο VGG-16, το οποίο όπως γίνεται και αντιληπτό από το όνομα, αποτελείται από 16 επίπεδα συνολικά από τα οποία τα 13 είναι συνελικτά και τα 3 πλήρως συνδεδεμένα. Συγκεχριμένα, το δίκτυο που θα χρησιμοποιήσουμε εδώ είναι το δίκτυο VGG-16 με Κανονικοποίηση Παρτίδας (βλ. Ενότητα 2.3.6), το οποίο βοηθάει το δίκτυο στην εκπαίδευση και την αποφυγή του προβλήματος "internal covariate shift". Στον Πίνακα 4.1 παρουσιάζεται αναλυτικά η αρχιτεκτονική του δικτύου που χρησιμοποιούμε για την εκπαίδευση των οπτικών μοντέλων VA και Basic Expressions. Κάθε συνελικτικό επίπεδο που απαρτίζει το δίκτυο ακολουθείται από ένα επίπεδο κανονικοποίησης παρτίδας και ένα επίπεδο ενεργοποίησης ReLU, τα οποία δεν έχουν συμπεριληφθεί στον πίνακα για λόγους χώρου. Επίσης, στον πίνακα αναφέρεται και ο συνολικός αριθμός εκπαιδεύσιμων παραμέτρων, ο οποίος ανέρχεται στα 135 εκατομμύρια.

Layers		Feature Map	Output Size	Kernel Size	Stride
1 - 2	2 × Conv	64	224 × 224 × 64	3 × 3	1
	Max Pooling	64	112 × 112 × 64	3 × 3	2
3 - 4	2 × Conv	128	112 × 112 × 128	3 × 3	1
	Max Pooling	128	56 × 56 × 128	3 × 3	2
5 - 7	3 × Conv	256	56 × 56 × 256	3 × 3	1
	Max Pooling	256	28 × 28 × 256	3 × 3	2
8 - 10	3 × Conv	512	28 × 28 × 512	3 × 3	1
	Max Pooling	512	14 × 14 × 512	3 × 3	2
11 - 13	3 × Conv	512	14 × 14 × 128	3 × 3	1
	Max Pooling	512	7 × 7 × 512	3 × 3	2
	Adaptive Avg Pooling	512	7 × 7 × 512		
14	FC		4096		
15	FC		4096		
16	FC		256		
17	FC		2 or 7		
Total Parameters		135,319,623			

Πίνακας 4.1: Αρχιτεκτονική του συνελικτικού δικτύου VGG-16 με αναλυτική παρουσίαση των επιπέδων του, όπως υλοποιείται από το Pytorch.

4.3.2 ResNet-50

Το δεύτερο συνελικτικό δίκτυο που επιλέγουμε να εκπαιδεύσουμε για τα δύο προβλήματα της αναγνώρισης συναισθήματος είναι της οικογένεια των Διαφορικών Δικτύων, τα οποία αναλύθηκαν στην Ενότητα 2.4.3. Τα διαφορικά δίκτυα ζεκινάνε από 18 επίπεδα και φτάνουν μέχρι και 152 επίπεδα. Εμείς για το παρόν πρόβλημα, επιθυμούμε ένα αρκετά βαθύ νευρωνικό δίκτυο καθώς

το σύνολο των εικόνων που έχουμε ως είσοδο είναι αρκετά μεγάλο και οι εικόνες έχουν πολλές πληροφορίες αλλά όχι τόσο βαθύ ώστε να οδηγούμαστε σε υπερπροσαρμογή των παραμέτρων του δικτύου. Συνεπώς, επιλέγουμε μία μέση λύση με ένα δίκτυο 50 επιπέδων, το ResNet-50. Στον Πίνακα 4.2 παρουσιάζεται αναλυτικά η αρχιτεκτονική του δικτύου και τα επίπεδα που το αποτελούν. Κάθε διαφορικό μπλοκ επιπέδων περιέχει 3 συνελικτικά επίπεδα με τα αντίστοιχα μεγέθη εξόδου, τα οποία φαίνονται στον Πίνακα αναλυτικά, και κάθε συνελικτικό επίπεδο του μπλοκ ακολουθείται από επίπεδο κανονικοποίησης παρτίδας και επίπεδο ενεργοποίησης ReLU όπως και στη περίπτωση του VGG-16. Τέλος, το πλήθος των εκπαίδευσιμων παραμέτρων για το δίκτυο είναι λίγο πάνω από τα 24 εκατομμύρια, το οποίο είναι πολύ μικρότερο από το δίκτυο VGG-16 (135 εκατομμύρια), γεγονός παράδοξο καθώς το ResNet-50 αποτελείται από πολύ περισσότερα επίπεδα. Το μικρότερο πλήθος παραμέτρων του διαφορικού δικτύου του δίνει σίγουρα πλεονέκτημα, καθώς το συγκεκριμένο χρειάζεται λιγότερο χρόνο και μικρότερη υπολογιστική ισχύ για την εκπαίδευση του.

Layers		Feature Map	Output Size	Kernel Size	Stride
1	Conv	64	112 × 112 × 64	7 × 7	2
	Max Pooling	64	56 × 56 × 64	3 × 3	2
2 - 10	3× Res Block	64	56 × 56 × 64	1 × 1	1
		64	56 × 56 × 64	3 × 3	1
		256	56 × 56 × 256	1 × 1	1
11 - 22	4× Res Block	128	56 × 56 × 128	1 × 1	1
		128	28 × 28 × 128	3 × 3	1
		512	28 × 28 × 512	1 × 1	1
23 - 40	6× Res Block	256	28 × 28 × 256	1 × 1	1
		256	14 × 14 × 256	3 × 3	1
		1024	14 × 14 × 1024	1 × 1	1
41 - 49	6× Res Block	512	14 × 14 × 512	1 × 1	1
		512	7 × 7 × 512	3 × 3	1
		2048	7 × 7 × 2048	1 × 1	1
	Adaptive Avg Pooling	2048	1 × 1 × 2048		
50	FC		256		
51	FC		2 or 7		
Total Parameters		24,034,375			

Πίνακας 4.2: Αρχιτεκτονική του συνελικτικού δικτύου ResNet-50 με αναλυτική παρουσίαση των επιπέδων του, όπως υλοποιείται από το Pytorch.

4.3.3 DenseNet-121

Το τρίτο συνελικτικό δίκτυο που επιλέγουμε να χρησιμοποιήσουμε για το πρόβλημα αναγνώρισης συναυτισθήματος είναι της οικογένειας των Πυκνά Συνδεδεμένων Δικτύων (βλ. Ενότητα 2.4.4). Τα λεγόμενα DenseNets που έχουν προταθεί στην αντίστοιχη δημοσίευση [43],

ζεκινούν από τα 121 επίπεδα και φτάνουν τα 264. Εδώ θα χρησιμοποιήσουμε το δίκτυο με τα λιγότερα επίπεδα, δηλαδή το DenseNet-121, καθώς είναι αρκετά βαθύ ώστε να μπορέσει να εξάγει τα χαρακτηριστικά που χρειαζόμαστε για την αναγνώριση των συνοισθημάτων από εικόνες. Στον Πίνακα 4.3 παρουσιάζονται αναλυτικά τα επίπεδα που αποτελούν το δίκτυο DenseNet-121 και τον συνολικό αριθμό εκπαιδεύσιμων παραμέτρων. Το δίκτυο DenseNet-121 αποτελείται από τα λεγόμενα Dense Blocks, τα οποία απαρτίζονται από δύο συνελικτικά επίπεδα και ακολουθούνται από επίπεδα κανονικοποίησης παρτίδας και επίπεδα ενεργοποίησης ReLU. Μεταξύ των Dense Blocks τοποθετούνται τα επίπεδα μετάβασης (Transition Layers), τα οποία αποτελούνται από ένα συνελικτικό επίπεδο και ένα επίπεδο υποδειγματοληψίας με την τεχνική του μέσου όρου (Average Pooling) με μέγεθος πυρήνα 2×2 . Ο αριθμός των εκπαιδεύσιμων παραμέτρων είναι σχεδόν 8 εκατομμύρια και είναι πολύ μικρότερος αριθμός σε σχέση με τα δίκτυα VGG-16 και ResNet-50, παρόλο που το συγκεκριμένο δίκτυο έχει τα περισσότερα επίπεδα.

Layers		Feature Map	Output Size	Kernel Size	Stride
1	Conv	64	$112 \times 112 \times 64$	7×7	2
	Max Pooling	64	$56 \times 56 \times 64$	3×3	2
2 - 13	6× Dense Block	128	$56 \times 56 \times 128$	1×1	1
		32	$56 \times 56 \times 32$	3×3	1
14	Conv	128	$56 \times 56 \times 128$	1×1	1
	Avg Pool	128	$28 \times 28 \times 128$	2×2	2
15 - 38	12× Dense Block	128	$28 \times 28 \times 128$	1×1	1
		32	$28 \times 28 \times 32$	3×3	1
39	Conv	256	$28 \times 28 \times 256$	1×1	1
	Avg Pool	256	$14 \times 14 \times 256$	2×2	2
40 - 87	24× Dense Block	128	$14 \times 14 \times 128$	1×1	1
		32	$14 \times 14 \times 32$	3×3	1
89	Conv	512	$14 \times 14 \times 512$	1×1	1
	Avg Pool	512	$7 \times 7 \times 512$	2×2	2
90 - 121	16× Dense Block	128	$7 \times 7 \times 128$	1×1	1
		32	$7 \times 7 \times 32$	3×3	1
122	FC		256		
123	FC		2 or 7		
Total Parameters		7,978,856			

Πίνακας 4.3: Αρχιτεκτονική του συνελικτικού δικτύου DenseNet-121 με αναλυτική παρουσίαση των επιπέδων του, όπως υλοποιείται από το Pytorch.

4.4 Προτεινόμενα Μοντέλα Ακουστικής Αναγνώρισης

Στην ενότητα αυτή θα αναλύσουμε τη μέθοδο ανάπτυξης των μοντέλων αναγνώρισης συναισθήματος με είσοδο το ακουστικό σήμα. Στην Ενότητα 3.3.2 παρουσιάστηκε η διαδικασία προεπεξεργασίας των ακουστικών σημάτων των βίντεο και η μετατροπή τους σε φασματογραφήματα. Τα φασματογραφήματα αυτά αποτελούν την είσοδο των δύο μοντέλων παλινδρόμησης του σθένους και της διέγερσης και ταξινόμησης των βασικών συναισθημάτων. Όπως και στην περίπτωση των οπτικών μοντέλων και εδώ αναπτύσσουμε δύο διαφορετικά μοντέλα για το κάθε πρόβλημα και τα παρουσιάζουμε εν συνεχείᾳ.

Ακουστικό Μοντέλο VA Ακολουθώντας την ίδια λογική των οπτικών μοντέλων, χρησιμοποιούμε ένα σύγχρονο συνελικτικό δίκτυο για την εξαγωγή των χαρακτηριστικών ώστε να γίνει η πρόβλεψη των τιμών του σθένους και της διέγερσης. Όλα τα συνελικτικά δίκτυα δέχονται ως είσοδο εικόνες με 3 κανάλια (RGB), τα φασματογραφήματα όμως αποτελούνται από ένα κανάλι, συνεπώς για να εισάγουμε τα φασματογραφήματα στα συνελικτικά δίκτυα μετατρέπουμε το πρώτο συνελικτικό επίπεδο των δικτύων με τρόπο ώστε να δέχεται εικόνες ενός καναλιού. Επιπρόσθετα, τα συνελικτικά δίκτυα είναι προεκπαιδευμένα στο ImageNet και για να μην χάσουμε αυτή τη πληροφορία αντιγράφουμε τα αντίστοιχα βάρη από το παλιό στο νέο συνελικτικό επίπεδο που δημιουργούμε. Για την τελική πρόβλεψη, αφαιρούμε το τελευταίο πλήρως συνδεδεμένο επίπεδο και προσθέτουμε δύο πλήρως συνδεδεμένα με μεγέθη εξόδου 256 και 2, όπου 2 είναι η τιμές της πρόβλεψης για το σθένος και τη διέγερση. Κι εδώ χρησιμοποιούμε το επίπεδο Dropout για την αποφυγή της υπερπροσαρμογής. Το μοντέλο έχει την ίδια αρχιτεκτονική με το αντίστοιχο οπτικό που φαίνεται στο Σχήμα 4.1 με τη διαφορά ότι δέχεται ως είσοδο τα φασματογραφήματα που προκύπτουν από τα βίντεο του Aff-Wild2.

Ακουστικό Μοντέλο Basic Expressions Το ακουστικό μοντέλο ταξινόμησης των 7 βασικών συναισθημάτων διαφέρει σε σχέση με αυτό της παλινδρόμησης στο τελευταίο πλήρως συνδεδεμένο επίπεδο, το οποίο τώρα έχει 7 τιμές, κάθε μία από τις οποίες δείχνει την πιθανότητα το φασματογράφημα εισόδου να ανήκει στη κάθε μία από τις 7 κλάσεις συναισθημάτων. Για τον τελικό προσδιορισμό της κλάσης χρησιμοποιείται το επίπεδο ενεργοποίησης Softmax. Η αρχιτεκτονική του ακουστικού αυτού δικτύου είναι ίδιο με εκείνο του οπτικού που παρουσιάζεται στο Σχήμα 4.2 με διαφορά ότι η είσοδος τώρα είναι τα φασματογραφήματα.

4.4.1 VGG-11

Το πρώτο συνελικτικό δίκτυο που θα χρησιμοποιηθεί για την εξαγωγή των χαρακτηριστικών από τα φασματογραφήματα είναι της οικογένειας VGG και συγκεκριμένα είναι αυτό με τα λιγότερα επίπεδα. Ο λόγος που επιλέγουμε ένα λιγότερο βαθύ δίκτυο είναι ότι τα φασματογραφήματα αποτελούν εικόνες ενός καναλιού και παρέχουν λιγότερη πληροφορία από ότι μία φωτογραφία ενός προσώπου, συνεπώς η εξαγωγή χαρακτηριστικών μπορεί να γίνει και από ένα λιγότερο πολύπλοκο δίκτυο που θα οδηγήσει σε μία πιο σωστή εκπαίδευση. Συγκεκριμένα χρησιμοποιείται το δίκτυο VGG-11 με επίπεδα κανονικοποίησης παρτίδας και η αρχιτεκτονική του φαίνεται

αναλυτικά στον Πίνακα 4.4. Στον Πίνακα φαίνονται και το πλήθος των παραμέτρων του δικτύου, το οποίο είναι σχεδόν 126 εκατομμύρια, πολύ μεγάλος αριθμός για ένα σχετικά απλό δίκτυο όπως το VGG-11.

Layers		Feature Map	Output Size	Kernel Size	Stride
1	Conv	64	$224 \times 224 \times 64$	3×3	1
	Max Pooling	64	$112 \times 112 \times 64$	3×3	2
2	Conv	128	$112 \times 112 \times 128$	3×3	1
	Max Pooling	128	$56 \times 56 \times 128$	3×3	2
3 - 4	$2 \times Conv$	256	$56 \times 56 \times 256$	3×3	1
	Max Pooling	256	$28 \times 28 \times 256$	3×3	2
5 - 6	$2 \times Conv$	512	$28 \times 28 \times 512$	3×3	1
	Max Pooling	512	$14 \times 14 \times 512$	3×3	2
7 - 8	$2 \times Conv$	512	$14 \times 14 \times 128$	3×3	1
	Max Pooling	512	$7 \times 7 \times 512$	3×3	2
	Adaptive Avg Pooling	512	$7 \times 7 \times 512$		
9	FC		4096		
10	FC		4096		
11	FC		256		
12	FC		2 or 7		
Total Parameters		125,822,471			

Πίνακας 4.4: Αρχιτεκτονική του συνελικτικού δίκτυου VGG-11 με αναλυτική παρουσίαση των επιπέδων του, όπως υλοποιείται από το Pytorch.

4.4.2 ResNet-18

Το δεύτερο συνελικτικό δίκτυο που επιλέγουμε να χρησιμοποιήσουμε για την εκπαίδευση του ακουστικού μοντέλου είναι ένα το διαφορικό συνελικτικό δίκτυο ResNet-18, το οποίο είναι το διαφορικό δίκτυο με τα λιγότερα επίπεδα. Επιλέγουμε το λιγότερο βαθύ διαφορικό δίκτυο καθώς τα φασματογραφήματα αποτελούν λιγότερο περίπλοκες εικόνες και συνεπώς ένα λιγότερο βαθύ νευρωνικό μπορεί να εξάγει με μεγαλύτερη επιτυχία τα αναγκαία χαρακτηριστικά για την αναγνώριση των συναισθημάτων. Αναλυτικά, η αρχιτεκτονική του ResNet-18 φαίνεται στον Πίνακα 4.5. Το ResNet-18 αποτελείται από 11 εκατομμύρια εκπαιδεύσιμες παραμέτρους, αρκετά λιγότερες από το VGG-11.

Layers		Feature Map	Output Size	Kernel Size	Stride
1	Conv	64	$112 \times 112 \times 64$	7×7	2
	Max Pooling	64	$56 \times 56 \times 64$	3×3	2
2 - 5	2× Res Block	64	$56 \times 56 \times 64$	3×3	1
		64	$56 \times 56 \times 64$	3×3	1
6 - 9	4× Res Block	128	$28 \times 28 \times 128$	3×3	1
		128	$28 \times 28 \times 128$	3×3	1
10 - 13	2× Res Block	256	$14 \times 14 \times 256$	3×3	1
		256	$14 \times 14 \times 256$	3×3	1
14 - 17	2× Res Block	512	$7 \times 7 \times 512$	3×3	1
		512	$7 \times 7 \times 512$	3×3	1
	Adaptive Avg Pooling	2048	$1 \times 1 \times 2048$		
18	FC		256		
19	FC		2 or 7		
Total Parameters			11,309,639		

Πίνακας 4.5: Αρχιτεκτονική του συνελικτικού δικτύου ResNet-18 με αναλυτική παρουσίαση των επιπέδων του, όπως υλοποιείται από το Pytorch.

4.5 Συνδυαστικό Μοντέλο Οπτικοακουστικής Αναγνώρισης

Στη ενότητα αυτή θα παρουσιάσουμε τα μοντέλα που προτείνονται για την αναγνώριση συναισθήματος από οπτικά και ακουστικά σήματα ταυτόχρονα. Αρχικά θα αναλύσουμε τις διάφορες τεχνικές που χρησιμοποιούνται για την ενσωμάτωση δύο τύπων εισόδου και στη συνέχεια θα επιλέξουμε μία από αυτές για την ανάπτυξη του συνδυαστικού μοντέλου για την επίλυση των προβλημάτων της παλινδρόμησης του σθένους και της διέγερσης και της ταξινόμησης των 7 βασικών συναισθημάτων.

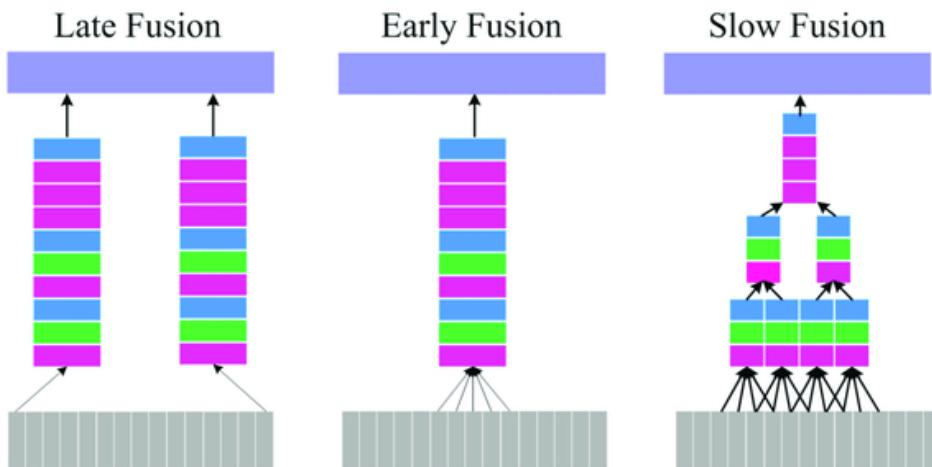
4.5.1 Μέθοδοι Ενσωμάτωσης Δεδομένων

Ο συνδυασμός δύο ή και περισσότερων τρόπων για την επίλυση προβλήματα αναγνώρισης προτύπων ονομάζεται πολυτροπική (multimodal) μέθοδος. Οι τρόποι αυτοί αναγνώρισης αναφέρονται στα διάφορα είδη σημάτων εισόδου, όπως είναι οι εικόνες (οπτικό σήμα), ο ήχος (ακουστικό σήμα), το κείμενο κ.α. Η ενσωμάτωση των διάφορων αυτών εισόδων για την εκπαίδευση ενός νευρωνικού δικτύου μπορεί να γίνει με τις εξής μεθόδους:

- **Early Fusion:** Η μέθοδος αυτή συνδυάζει τους διάφορους τύπους δεδομένων σε αρχικό επίπεδο, δηλαδή πριν ακόμα γίνει κάποια εξαγωγή χαρακτηριστικών. Αρχικά, γίνεται η συγκέντρωση όλων των δεδομένων εισόδου και στη συνέχεια πραγματοποιείται η τροφοδότησή τους σε κάποιο νευρωνικό δίκτυο ώστε να γίνει η ταξινόμηση. Η μέθοδος αυτή προϋποθέτει την προεπεξεργασία των δεδομένων εισόδου ώστε να έρθουν στην ίδια ακριβώς μορφή με αυτή που δέχεται το νευρωνικό δίκτυο.

- **Slow Fusion:** Η μέθοδος αυτή συνδυάζει τα δεδομένα σε όλο το μήκος του δικτύου, δημιουργώντας διάφορους κλάδους, οι οποίοι τελικά καταλήγουν σε ένα πλήρως συνδεδεμένο επίπεδο που κάνει την τελική ταξινόμηση. Η συγκεκριμένη μέθοδος είναι αρκετά περίπλοκη λόγω των διάφορων κλάδων του δικτύου και έχει ως αποτέλεσμα την πιο αργή εκπαίδευση και τη μεγαλύτερη κατανάλωση μνήμης.
- **Late Fusion:** Η μέθοδος late fusion διενεργεί το συνδυασμό των δεδομένων στο τελευταίο στάδιο της ταξινόμησης, δηλαδή σε επίπεδο χαρακτηριστικών. Κάθε είδος εισόδου έχει το αντίστοιχο δίκτυο εξαγωγής χαρακτηριστικών δηλαδή τον δικό του κλάδο στο συνδυαστικό μοντέλο. Μετά την εξαγωγή των χαρακτηριστικών από κάθε είδος εισόδου γίνεται η συνένωσή τους με έναν αριθμό πλήρως συνδεδεμένων επιπέδων έτσι ώστε να γίνει η τελική πρόβλεψη του μοντέλου.

Οι παραπάνω μέθοδοι φαίνονται και παραστατικά στο Σχήμα 4.3. Η κάθε μέθοδος έχει πλεονεκτήματα και μειονεκτήματα, συνεπώς η επιλογή της μεθόδου πρέπει να γίνεται με βάση το είδος των δεδομένων του προβλήματος αλλά και των νευρωνικών δικτύων που θα χρησιμοποιηθούν. Στη περίπτωση του παρόντος προβλήματος αναγνώρισης συναισθήματος από δεδομένα εικόνας και ήχου, επιλέγουμε να χρησιμοποιήσουμε τη μέθοδο late fusion. Ο λόγος επιλογής της είναι ότι οι φωτογραφίες των προσώπων και τα φασματογραφήματα είναι δύο εντελώς διαφορετικά σήματα εισόδου, καθώς έχουν διαφορετικό πλήθος καναλιών και γενικώς διαφέρουν αρκετά σε σχέση με τα χαρακτηριστικά τους. Επίσης, αρχικός σκοπός ήταν η δοκιμή διαφορετικών συνελικτικών δικτύων για κάθε είδος εισόδου, το οποίο επιτρέπει μόνο η συγκεκριμένη μέθοδος.

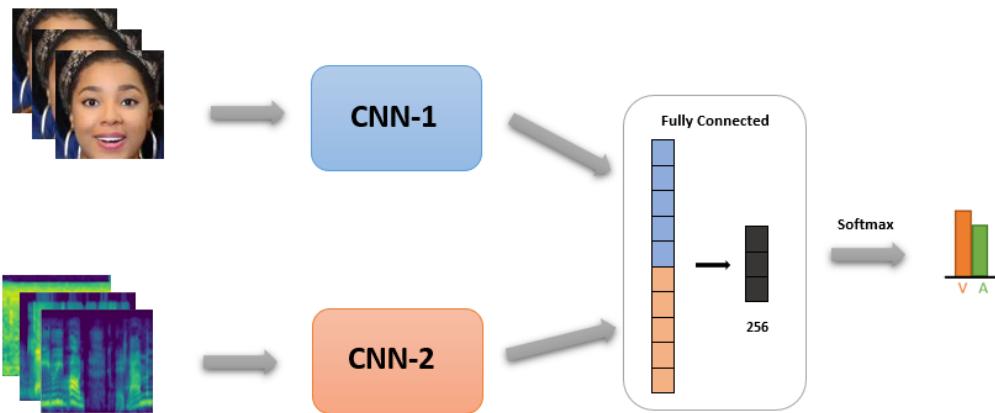


Σχήμα 4.3: Μέθοδοι Ενσωμάτωσης Δεδομένων και οι αντίστοιχες αρχιτεκτονικές των δικτύων [66].

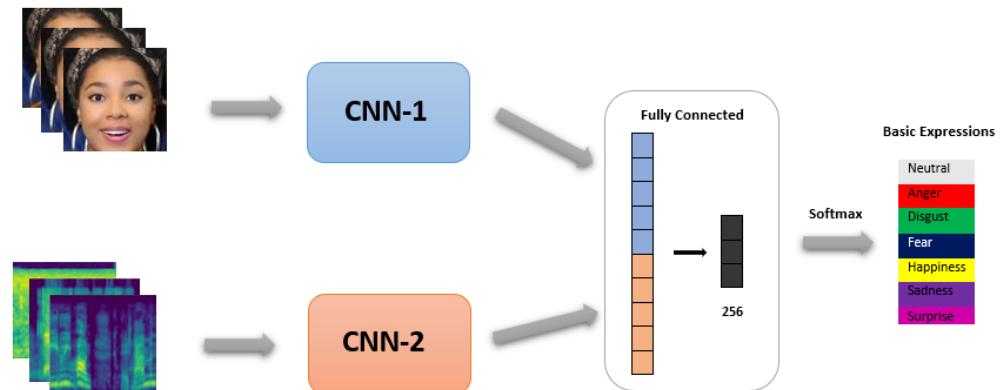
4.5.2 Συνδυαστικά Μοντέλα Δύο Κλάδων

Ο συνδυασμός των οπικών και ακουστικών μοντέλων για τα προβλήματα παλινδρόμησης του σθένους και τη διέγερσης και ταξινόμησης των 7 βασικών συναισθημάτων γίνεται με τη μέθοδο late fusion όπως προαναφέρθηκε νωρίτερα. Συγκεκριμένα, αναπτύσσονται μοντέλα με δύο κλάδους και ο κάθε κλάδος αναφέρεται σε ένα είδος εισόδου, οπικό ή ακουστικό. Στη

συνέχεια γίνεται η εξαγωγή των χαρακτηριστικών από δύο διαφορετικά συνελικτικά δίκτυα και τα διανύσματα εξόδου του κάθε δικτύου συνενώνονται σε ένα μεγαλύτερο διάνυσμα, το οποίο τροφοδοτείται σε ένα πλήρως συνδεδεμένο επίπεδο με αριθμό νευρώνων ίσο με το μέγεθος του συνδυαστικού διανύσματος. Έπειτα προσθέτουμε άλλα δύο πλήρως συνδεδεμένα επίπεδα μεγέθους 256 και 2 ή 7 ανάλογα με το πρόβλημα. Τα δύο μοντέλα παρουσιάζονται και σχηματικά στα Σχήματα 4.4 και 4.5.



Σχήμα 4.4: Αρχιτεκτονική του συνδυαστικού μοντέλου παλινδρόμησης του σθένους και της διέγερσης (VA).



Σχήμα 4.5: Αρχιτεκτονική του συνδυαστικού μοντέλου ταξινόμησης των 7 βασικών συναισθημάτων.

Για τον κλάδο του μοντέλου που εισάγονται οι εικόνες των προσώπων θα χρησιμοποιηθούν τα συνελικτικά δίκτυα που αναπτύχθηκαν στην Ενότητα 4.3 και είναι τα VGG-16, ResNet-50 και DenseNet-121. Ενώ για τον κλάδο που δέχεται τα φασματογραφήματα θα χρησιμοποιηθούν τα συνελικτικά δίκτυα που αναπτύχθηκαν στην Ενότητα 4.4 και είναι τα VGG-11 και ResNet-18. Ο σκοπός είναι να γίνει κάθε πιθανός συνδυασμός των παραπάνω συνελικτικών δικτύων και να

συγχριθούν οι επιδόσεις του κάθε μοντέλου. Η εκπαίδευση και η αξιολόγηση των μοντέλων που αναλύθηκαν σε αυτό το κεφάλαιο παρουσιάζεται στο Κεφάλαιο 5.

Κεφάλαιο 5

Εκπαίδευση και Αξιολόγηση των Μοντέλων

Στο παρόν Κεφάλαιο θα παρουσιαστεί η πειραματική διαδικασία που ακολουθήθηκε ώστε να γίνει η εκπαίδευση των μοντέλων που αναλύθηκαν στο Κεφάλαιο 4. Επίσης, θα γίνει η αξιολόγηση των οπτικών, ακουστικών και συνδυαστικών μοντέλων για τα δύο είδη προβλημάτων αναγνώρισης συναίσθήματος και τέλος θα αναλυθούν τα αποτελέσματα αυτών λαμβάνοντας υπόψιν παράγοντες που συμβάλουν στην απόδοσή τους.

5.1 Εκπαίδευση των Μοντέλων

Για την επιτυχή εκπαίδευση των μοντέλων και την τελική επιλογή των υπερπαραμέτρων έγιναν μία σειρά από πειράματα χρησιμοποιώντας διάφορα υπολογιστικά συστήματα. Παρακάτω γίνεται αναλυτική παρουσίαση των ενεργειών και των πειραμάτων που πραγματοποιήθηκαν σε κάθε υπολογιστικό σύστημα και το λογισμικό που χρησιμοποιήθηκε. Ακόμη, αναγράφεται με συγκεκριμένα βήματα η συνολική διαδικασία εκπαίδευσης των μοντέλων και αναλύονται οι επιλογές κάθε παραμέτρου.

5.1.1 Υπολογιστικά Συστήματα

Η αρχική ανάπτυξη του κώδικα υλοποίησης των μοντέλων αναγνώρισης συναίσθήματος έγινε χρησιμοποιώντας τη πλατφόρμα ανάπτυξης και εκτέλεσης κώδικα Python, [Google Colaboratory](#). Λόγω των περιορισμών μνήμης του συστήματος έγινε εκπαίδευση των μοντέλων με ένα μικρό υποσύνολο του συνόλου δεδομένων ώστε να γίνει η διόρθωση του κώδικα (debugging).

Στη συνέχεια, έγινε χρήση του υπολογιστικού συστήματος του Εργαστηρίου Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου. Το σύστημα αυτό διαθέτει 2 Μονάδες Επεξεργασίας Γραφικών (GPU) Nvidia GeForce GTX 1080 με μνήμη 8 GB έκαστη. Σε αυτό το σύστημα έγινε η εκτέλεση της διαδικασίας της προεπεξεργασίας του συνόλου δεδομένων Aff-Wild2 και τα πειράματα για τον κατάλληλο προσδιορισμό των υπερπαραμέτρων των επιμέρους μοντέλων.

Η τελική εκπαίδευση των μοντέλων με το πλήρες σύνολο εκπαίδευσης, το οποίο αποτελείται από πάνω από 3 εκατομμύρια εικόνες συνολικά εκτελέστηκε στο εθνικό υπερυπολογιστικό

σύστημα ***ARIS*** του Εθνικού Δικτύου Υποδομών Τεχνολογίας και Έρευνας (ΕΔΥΤΕ). Η συγκεκριμένη υπερυπολογιστική υποδομή αποτελείται από 44 κόμβους GPU (GPU Nodes), ο καθένας από τους οποίους διαθέτει 2 GPU Nvidia Tesla k40m μη μνήμη 12 GB έκαστη. Συνεπάς, μας δόθηκε η δυνατότητα ταυτόχρονης εκτέλεσης του κάθικα υλοποίησης των διάφορων μοντέλων και η γρηγορότερη εκπαίδευσή τους.

5.1.2 Λογισμικό

Η ανάπτυξη του κάθικα έγινε με γλώσσα Python 3.8 και συγκεκριμένα χρησιμοποιήθηκε η βιβλιοθήκη Pytorch [62] για την υλοποίηση των νευρωνικών δικτύων. Το κάθικε ένα από τα συνελικτικά δίκτυα που χρησιμοποιούνται στην δημιουργία των μοντέλων αναγνώρισης συνασθήματος ακολουθούν την υλοποίηση του Pytorch και είναι προεκπαιδευμένα στο ImageNet. Άλλες βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι **Pandas**, **Scikit-learn**, **Imbalanced-learn** και **Seaborn**.

5.1.3 Διαδικασία Εκπαίδευσης

Η διαδικασία εκπαίδευσης των μοντέλων αποτελείται από τα εξής βήματα:

- **Βήμα 1:** Διάβασμα των αρχείων όπου καταγράφονται η διαδρομή (path) κάθε εικόνα στο σύστημα αρχείων και η αντίστοιχη ετικέτα. Τα αρχεία αναφέρονται στο σύνολο εκπαίδευσης και στο σύνολο αξιολόγησης.
- **Βήμα 2:** Φόρτωση των εικόνων του συνόλου δεδομένων ανά παρτίδα (batch). Το μέγεθος παρτίδας (batch size) διαφέρει για κάθε μοντέλο καθώς έχει σχέση με τη μνήμη που διαθέτει η κάθικ GPU, όπου γίνεται η φόρτωση της παρτίδας και του μοντέλου. Η μνήμη που καταλαμβάνει κάθικ μοντέλο εξαρτάται από το συνελικτικό δίκτυο που χρησιμοποιείται. Όπως είδαμε στις Ενότητες 4.3 και 4.4 το κάθικ συνελικτικό δίκτυο έχει διαφορετικό αριθμό παραμέτρων. Όταν το πλήθος των παραμέτρων είναι μεγάλο το μοντέλο καταλαμβάνει μεγάλο χώρο στη μνήμη και για αυτό το λόγο χρησιμοποιούμε μικρότερο batch size. Το γεγονός αυτό επιβραδύνει το χρόνο εκπαίδευσης του μοντέλου.
- **Βήμα 3:** Μετατροπή των εικόνων εισόδου ανά παρτίδα έτσι ώστε να γίνει η τροφοδότηση τους στο δίκτυο. Συγκεκριμένα γίνεται η μετατροπή του μεγέθους των RGB εικόνων σε $3 \times 224 \times 224$ και των φασματογραφημάτων σε $1 \times 224 \times 224$. Έπειτα γίνεται η κανονικοποίηση των τιμών του κάθικ pixel σύμφωνα με το μέσο όρο και την τυπική απόκλιση των εικόνων. Και τέλος, πραγματοποιείται επαύξηση των δεδομένων (data augmentation) με το τυχαίο οριζόντιο flip της εικόνας με πιθανότητα 50%.
- **Βήμα 4:** Επιλογή και αρχικοποίηση του αλγορίθμου βελτιστοποίησης (optimizer) με τις κατάλληλες υπερπαραμέτρους. Εδώ επιλέγουμε τον αλγόριθμο βελτιστοποίησης Adam [67]. Στην αρχικοποίηση του αλγορίθμου προσδιορίζουμε την τιμή του ρυθμού μάθησης (learning rate) και του παράγοντα μείωσης του βάρους (weight decay). Οι τιμές αυτές διαφέρουν για κάθικ μοντέλο και θα αναφερθούν στη συνέχεια.
- **Βήμα 5:** Εκπαίδευση των μοντέλων στο σύνολο δεδομένων εκπαίδευσης για ένα πλήθος εποχών (epoch). Το μοντέλο κάθικ εποχή τροφοδοτείται με όλο το σύνολο δεδομένων και

προσαρμόζει τις παραμέτρους του. Συνεπώς με κάθε εποχή βελτιώνεται στην πρόβλεψη των σωστών ετικετών εκτός αν έχουμε το φαινόμενο της υπερπροσαρμογής, όπως το μοντέλο σταματάει να βελτιώνεται. Το πλήθος των εποχών εδώ είναι 10.

- *Bήμα 6:* Αξιολόγηση του μοντέλου σύμφωνα με το σύνολο αξιολόγησης. Μετά το πέρας κάθε εποχής το μοντέλο πραγματοποιεί προβλέψεις για το σύνολο αξιολόγησης. Η απόδοσή του σε αυτό μετράται με τις μετρικές αξιολόγησης που αναλύσαμε στο Κεφάλαιο 4 και με αυτό τον τρόπο επιλέγεται το καλύτερο μοντέλο, το οποίο και αποθηκεύεται.

5.2 Αξιολόγηση Μοντέλων

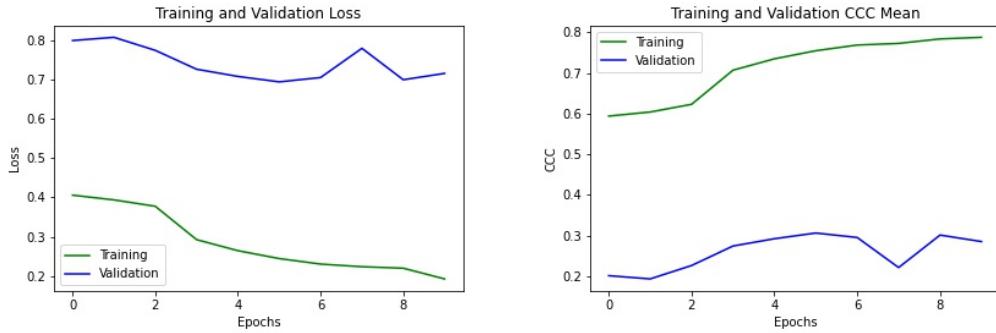
Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα των μοντέλων όπως αυτά αξιολογήθηκαν με βάση τις μετρικές αξιολόγησης του κάθε προβλήματος (βλ. Ενότητα 4.1). Επίσης, αναφέρονται οι υπερπαράμετροι που επιλέχθηκαν για την εκπαίδευση του κάθε μοντέλου και απεικονίζονται οι καμπύλες εκπαίδευσης και αξιολόγησης του κόστους και της αντίστοιχης μετρικής αξιολόγησης για τα καλύτερα μοντέλα κάθε κατηγορίας.

5.2.1 Αξιολόγηση Μοντέλων VA

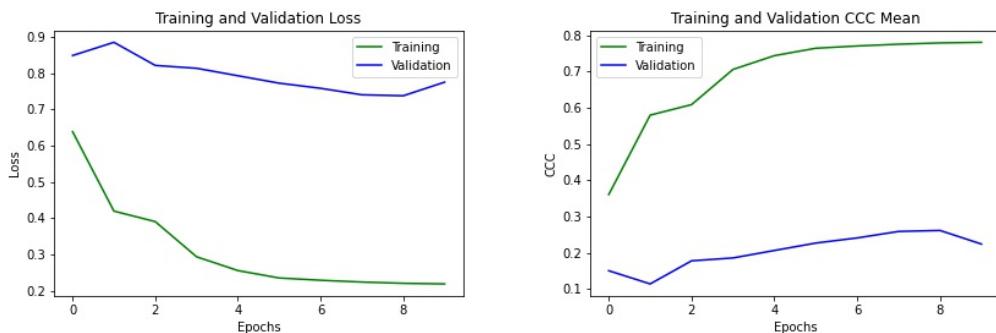
Τα μοντέλα τα οποία έχουμε αναπτύξει για την επίλυση του προβλήματος παλινδρόμησης του σθένους και της διέγερσης των συναισθημάτων χωρίζονται σε σχέση με το είδος της πληροφορίας εισόδου που δέχονται σε τρεις κατηγορίες. Οι κατηγορίες είναι τα οπτικά, τα ακουστικά και τα συνδυαστικά μοντέλα (βλ. Κεφάλαιο 4), τα οποία έχουν ως βάση συγκεκριμένα συνελικτικά νευρωνικά δίκτυα.

Το καθένα από αυτά τα μοντέλα έχουν εκπαίδευτεί στο σύνολο εκπαίδευσης για 10 εποχές και έχει γίνει η αξιολόγησή του στο σύνολο αξιολόγησης. Συγκεκριμένα, όσο αναφορά τις υπερπαραμέτρους εκπαίδευσης, για τα οπτικά και τα ακουστικά μοντέλα ο ρυθμός εκπαίδευσης τέθηκε στο 0.001 και μειώνονταν 0.0001 μετά από 7 εποχές ενώ για τα συνδυαστικά μοντέλα τέθηκε στο 0.005 και μειώνόταν στο 0.0005 μετά από 5 εποχές. Επίσης, η υπερπαράμετρος weight decay τέθηκε στο 0.0001 για όλα τα μοντέλα. Στα Σχήματα 5.1, 5.2 και 5.3 παρουσιάζονται οι καμπύλες του κόστους και της μετρικής αξιολόγησης CCC συναρτήσει των εποχών εκπαίδευσης. Παρατηρώντας τα σχήματα αυτά, βλέπουμε ότι οι τιμές του κόστους στη φάση της εκπαίδευσης είναι πολύ χαμηλότερο από εκείνο της αξιολόγησης, αυτό δείχνει ότι το δίκτυο έχει μειωμένη ικανότητα γενίκευσης στο σύνολο αξιολόγησης.

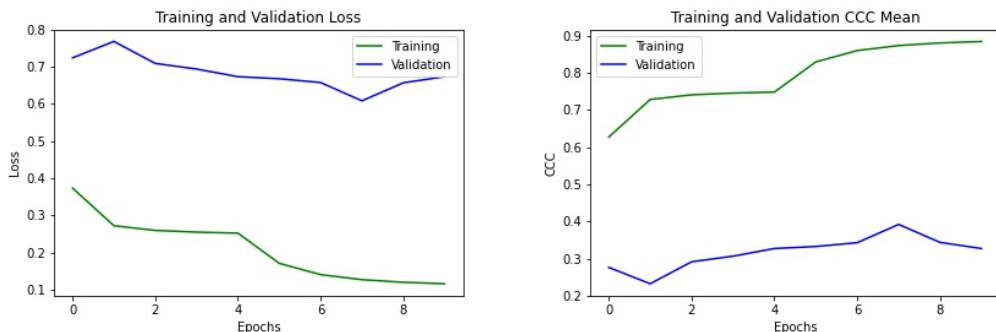
Στο Πίνακα 5.1 παρουσιάζονται αναλυτικά τα αποτελέσματα της αξιολόγησης όλων των μοντέλων που εκπαίδεύηκαν για το πρόβλημα αναγνώρισης συναισθήματος μέσω των τιμών του σθένους και της διέγερσης. Για τα μοντέλα που δέχονται μόνο τις εικόνες του συνόλου δεδομένων επιλέχθηκαν τα συνελικτικά δίκτυα VGG-16, ResNet-50 και DenseNet-121 (βλ. Ενότητα 4.3), με την καλύτερη απόδοση συνολικά να έχει το DenseNet-121 ενώ υψηλότερη μετρική CCC για τη διέγερση έχει το ResNet-50. Τα μοντέλα που δέχονται ως είσοδο μόνο τα φασματογραφήματα που δημιουργήθηκαν από τον ήχο των βίντεο του Aff-Wild2 έχουν ως βάση τα συνελικτικά δίκτυα VGG-11 και ResNet-18 (βλ. Ενότητα 4.4), από τα οποία καλύτερη απόδοση έχει το VGG-11. Για τα μοντέλα που δέχονται ταυτόχρονα τις εικόνες των προσώπων και τα φασματογραφήματα,



Σχήμα 5.1: Καμπύλες Κόστους και Μετρικής Αξιολόγησης CCC με βάση το DenseNet-121.



Σχήμα 5.2: Καμπύλες Κόστους και Μετρικής Αξιολόγησης CCC με βάση το VGG-11.



Σχήμα 5.3: Καμπύλες Κόστους και Μετρικής Αξιολόγησης CCC με βάση το ResNet-50 και το ResNet-18.

έγινα δύο συνδυασμοί συνελικτικών δικτύων. Ο πρώτος συνδυασμός είναι το ResNet-50 για τα οπτικά και το ResNet-18 για τα ακουστικά δεδομένα, ο οποίος έχει τη καλύτερη απόδοση συνολικά και στη κατηγορία του σθένους. Ο δεύτερος συνδυασμός είναι το DenseNet-121 για τα οπτικά και το VGG-11 για τα ακουστικά δεδομένα και έχει τη καλύτερη απόδοση στη κατηγορία της διέγερσης. Το VGG-16 δεν χρησιμοποιήθηκε στα συνδυαστικά μοντέλα καθώς είναι αυτό με τις περισσότερους παραμέτρους (135 εκατομμύρια) και συνεπώς η εκπαίδευση του απαιτεί μεγάλη υπολογιστική ισχύ και είναι πολύ χρονοβόρα.

Ανακεφαλαιώνοντας, το μοντέλο με την καλύτερη απόδοση για το συγκεκριμένο πρόβλημα σύμφωνα με τη μετρική αξιολόγησης CCC είναι το συνδυαστικό μοντέλο δύο κλάδων, το οποίο αποτελείται από το συνελικτικό δίκτυο ResNet-50 για τα οπτικά δεδομένα και το ResNet-18 για

Models	Modality		CCC		
	Visual	Audio	Valence	Arousal	Mean
VGG-16	✓	-	0.0738	0.3137	0.194
ResNet-50	✓	-	0.0716	0.4229	0.247
DenseNet-121	✓	-	0.3100	0.3026	0.306
VGG-11	-	✓	0.1391	0.3848	0.262
ResNet-18	-	✓	0.1011	0.3129	0.207
ResNet-50 & ResNet-18	✓	✓	0.3373	0.4469	0.392
DenseNet-121 & VGG-11	✓	✓	0.2357	0.4871	0.361

Πίνακας 5.1: Αποτελέσματα των μοντέλων παλινδρόμησης του σθένους(valence) και της διέγερσης(arousal) στο σύνολο αξιολόγησης. Με έντονη γραφικάσειρα επισημαίνονται οι καλύτερες τιμές της κάθε μετρικής.

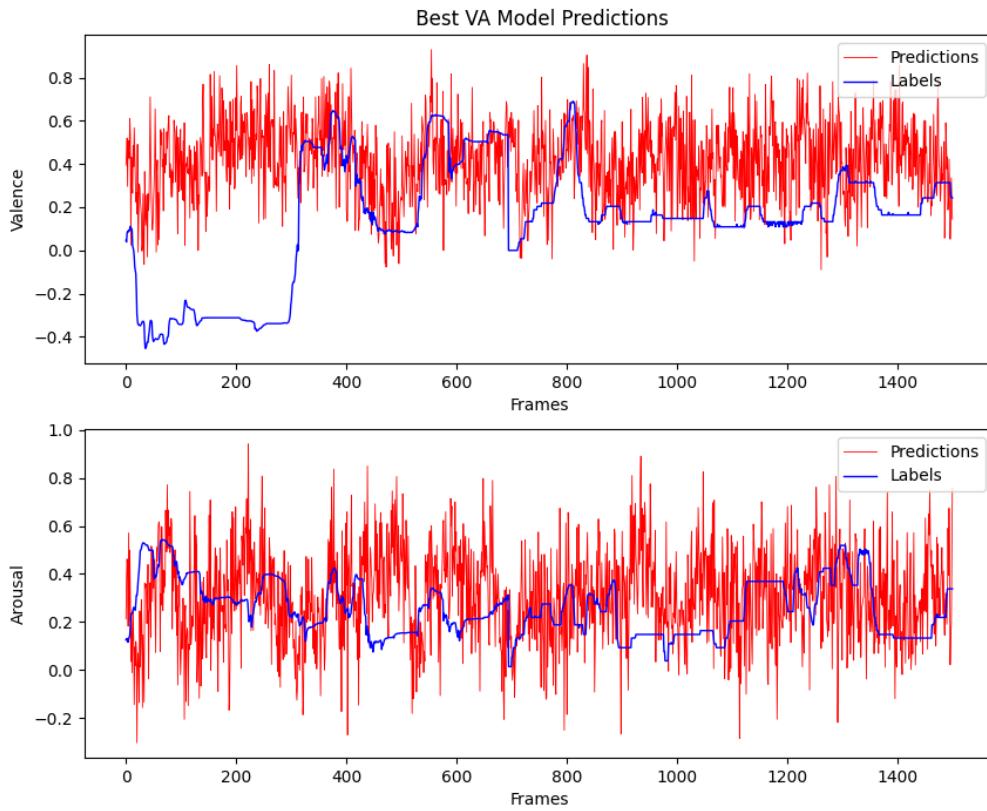
τα ακουστικά. Στο Σχήμα 5.4 φαίνονται ταυτόχρονα οι ετικέτες του σθένους και της διέγερσης και οι προβλέψεις των τιμών αυτών από το παραπάνω μοντέλο για ένα βίντεο του συνόλου αξιολόγησης. Είναι εμφανές ότι οι προβλέψεις του μοντέλου είναι αρκετά ασταθείς αλλά δεν είναι πολύ μακρινές από τις πραγματικές τιμές.

5.2.2 Αξιολόγηση Μοντέλων Basic Expressions

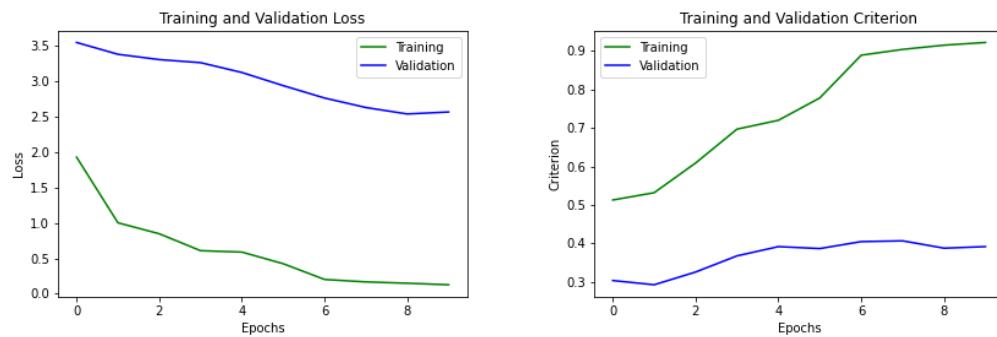
Παρόμοια με τη περίπτωση του προβλήματος παλινδρόμησης, για το πρόβλημα της ταξινόμησης των 7 βασικών συναισθημάτων έχουμε αναπτύξει τρία είδη μοντέλων, τα οπτικά, τα ακουστικά και τα συνδυαστικά. Κάθε ένα από τα παραπάνω μοντέλα βασίζεται σε ένα συνελικτικό δίκτυο.

Η εκπαίδευση των μοντέλων και σε αυτή τη περίπτωση έγινε σε 10 εποχές και επιλέχθηκαν οι αντίστοιχοι υπερπαράμετροι. Ο ρυθμός μάθησης τέθηκε στο 0.0001 και μειωνόταν στο 0.00001 μετά από 7 εποχές για τα οπτικά μοντέλα, ενώ για τα ακουστικά και τα συνδυαστικά τέθηκε στο 0.0005 και μειωνόταν στο 0.00005 μετά από 5 εποχές. Σε όλα τα μοντέλα η υπερπαράμετρος weight decay τέθηκε στο 0.0001. Στα Σχήματα 5.5, 5.6 και 5.7 απεικονίζονται οι καμπύλες του κόστους και του κριτήριου αξιολόγησης για τις φάσεις της εκπαίδευσης και της αξιολόγησης συναρτήσει των εποχών. Εδώ οι τιμές του κόστους για το σύνολο αξιολόγησης είναι αρκετά μεγάλος καθώς χρησιμοποιούμε τη συνάρτηση κόστους Cross Entropy Loss, την οποία αναφέραμε στην Ενότητα 4.2.2, πολλαπλασιάζοντας όμως με τα βάρη της κάθε ακλάσης καθώς έχουμε μη ισορροπημένο σύνολο δεδομένων. Συνεπώς, οι ακλάσεις που έχουν μικρό πλήθος παραδειγμάτων έχουν μεγάλο κόστος ώστε το δίκτυο να βελτιωθεί στην πρόβλεψή τους.

Στον Πίνακα 5.2 γίνεται η παρουσίαση των αποτελεσμάτων των μοντέλων σε σχέση με τις μετρικές αξιολόγησης της ακρίβειας, του F1-Score και του κριτηρίου που συνδυάζει αυτές τις δύο τιμές. Τα συνελικτικά δίκτυα που επιλέχθηκαν είναι ίδια με τη περίπτωση του προβλήματος παλινδρόμησης. Παρατηρώντας τον πίνακα, το οπτικό μοντέλο με βάση το συνελικτικό δίκτυο DenseNet-121 αποδίδει τη μεγαλύτερη ακρίβεια, η οποία είναι 52,15%, και το μεγαλύτερο F1-

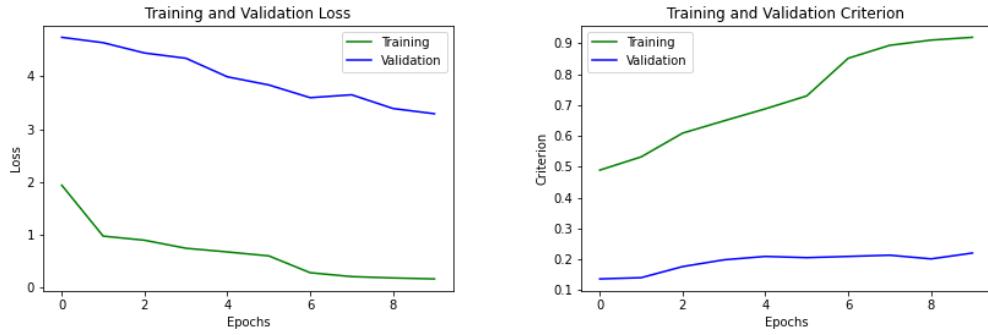


Σχήμα 5.4: Παράδειγμα προβλέψεων και ετικετών του σθένους και της διέγερσης ενός βίντεο του συνόλου αξιολόγησης.

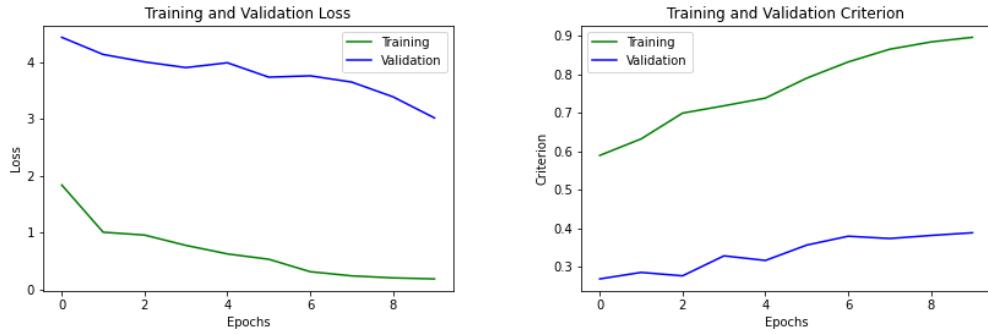


Σχήμα 5.5: Καμπύλες Κόστους και Κριτηρίου Αξιολόγησης με βάση το DenseNet-121.

Score, το οποίο είναι 35,05%. Ο συνδυασμός των δύο με βάση το χριτήριο αξιολόγησης είναι 40,7%. Δεύτερο σε απόδοση είναι το συνδυαστικό μοντέλο δύο χλάδων, το οποίο έχει ως βάση το DenseNet-121 για τα οπτικά δεδομένα και το VGG-11 για τα ακουστικά δεδομένα, με την τιμή του χριτηρίου να ανέρχεται στο 38,8%. Τέλος, το VGG-16 ενώ έχει καλύτερη απόδοση από το ResNet-50 στα οπτικά δεδομένα, δεν χρησιμοποιείται στα συνδυαστικά μοντέλα λόγω του μεγάλου πλήθους εκπαιδεύσιμων παραμέτρων που οδηγεί σε πολύ μεγάλο χρόνο εκπαίδευσης.



Σχήμα 5.6: Καμπύλες Κόστους και Κριτηρίου Αξιολόγησης με βάση το VGG-11.



Σχήμα 5.7: Καμπύλες Κόστους και Κριτηρίου Αξιολόγησης με βάση το ResNet-50 και το ResNet-18.

Models	Modality		Metric		
	Visual	Audio	Accuracy	F1-Score	Criterion
VGG-16	✓	-	0.5093	0.3169	0.380
ResNet-50	✓	-	0.5075	0.3030	0.370
DenseNet-121	✓	-	0.5215	0.3505	0.407
VGG-11	-	✓	0.3185	0.1714	0.220
ResNet-18	-	✓	0.3472	0.1547	0.218
ResNet-50 & ResNet-18	✓	✓	0.4099	0.2726	0.318
DenseNet-121 & VGG-11	✓	✓	0.4808	0.3306	0.388

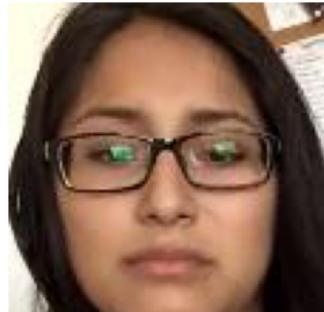
Πίνακας 5.2: Αποτελέσματα των μοντέλων ταξινόμησης των 7 βασικών συναισθημάτων στο σύνολο αξιολόγησης. Με έντονη γραμματοσειρά επισημαίνονται οι καλύτερες τιμές για κάθε μετρική.

Στα Σχήματα 5.8 και 5.9 παρουσιάζονται κάποιες από τις εικόνες προσώπων του συνόλου αξιολόγησης με τις αντίστοιχες προβλέψεις του μοντέλου με τη καλύτερη απόδοση. Όπως γίνεται εύκολα αντιληπτό, η πιθανότητα να είναι σωστή μία πρόβλεψη είναι 40.7%, σύμφωνα με το κριτήριο αξιολόγησης. Συνεπώς, στο σύνολο των 8 εικόνων, υπάρχει η πιθανότητα τριών σωστών προβλέψεων, όπως συμβαίνει και σε αυτή τη περίπτωση των παραδειγμάτων. Παρατηρούμε ότι το μοντέλο έχει κάνει σωστές προβλέψεις για τη κλάση της Ουδετερότητας (Neutral), η οποία είναι και η κλάση με το μεγαλύτερο αριθμό παραδειγμάτων στο σύνολο δεδομένων.

Predicted: Happiness
Label: Neutral



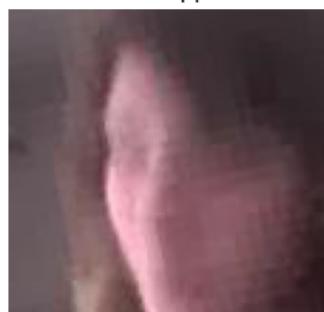
Predicted: Sadness
Label: Sadness



Predicted: Neutral
Label: Neutral



Predicted: Sadness
Label: Happiness



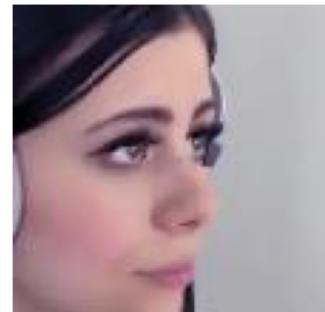
Predicted: Sadness
Label: Neutral



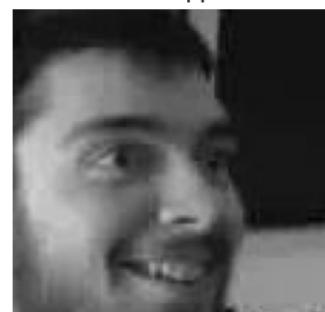
Predicted: Neutral
Label: Neutral



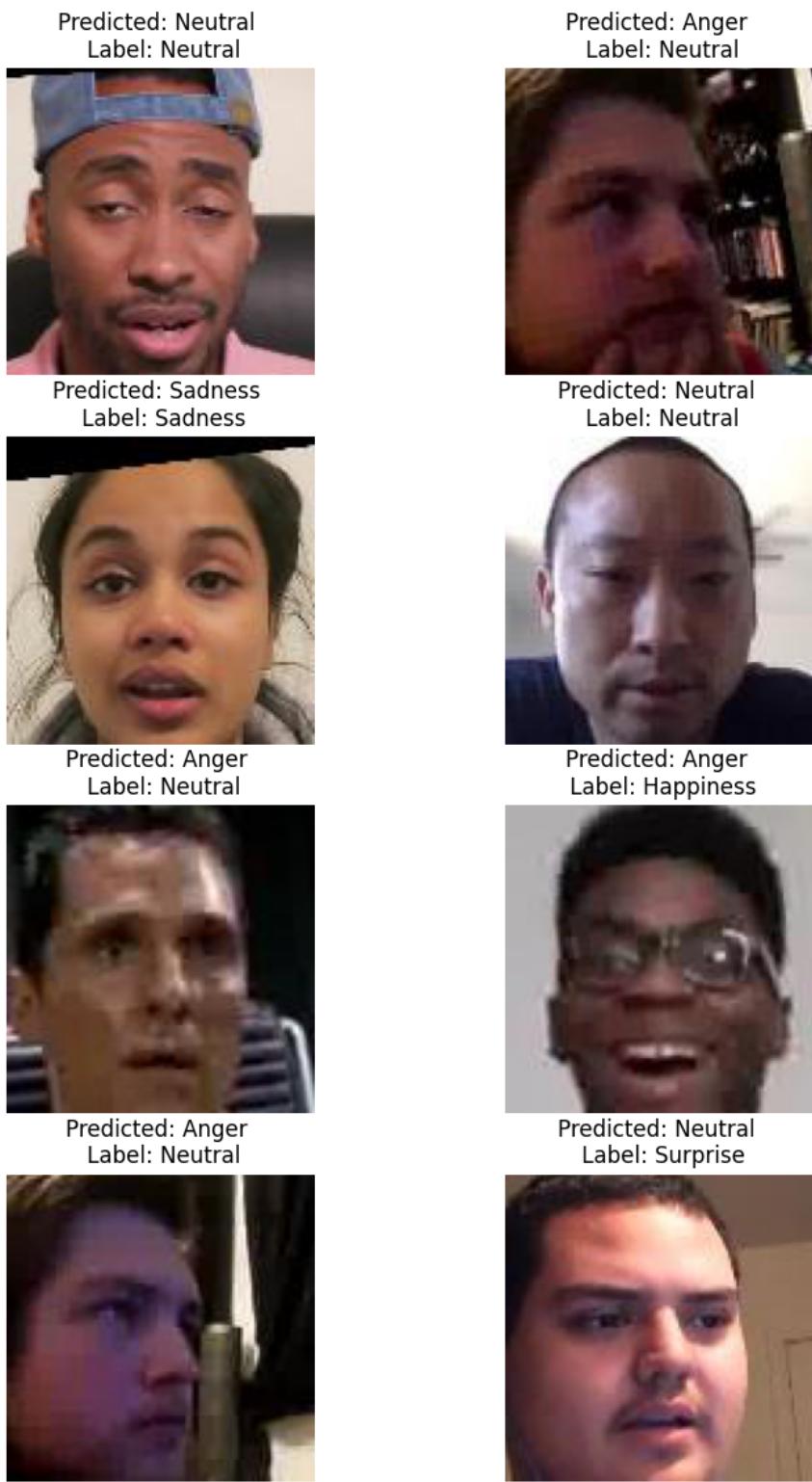
Predicted: Sadness
Label: Neutral



Predicted: Sadness
Label: Happiness



Σχήμα 5.8: Παραδείγματα προβλέψεων και ετικετών εικόνων από βίντεο του συνόλου αξιολόγησης.



Σχήμα 5.9: Παραδείγματα προβλέψεων και ετικετών εικόνων από βίντεο του συνόλου αξιολόγησης.

5.3 Ανάλυση Αποτελεσμάτων

Για την ακριβή αξιολόγηση των αποτελεσμάτων των μοντέλων αναγνώρισης συναισθήματος που αναπτύξαμε είναι αναγκαίο πρώτα να αναφερθούμε στους παράγοντες από τους οποίους εξαρτώνται τα αποτελέσματα αυτά. Το είδος των δεδομένων με τα οποία έγινε η εκπαίδευση των μοντέλων, τα συνελικτικά δίκτυα που επιλέχθηκαν για την εξαγωγή χαρακτηριστικών αλλά και η ίδια η διαδικασία της εκπαίδευσης είναι παράγοντες που επηρέασαν τα αποτελέσματα και αναλύονται στις επόμενες παραγράφους. Λαμβάνοντας υπόψιν τους παράγοντες αυτούς, στη συνέχεια αναλύουμε την απόδοση των μοντέλων των δύο προβλημάτων ξεχωριστά.

Δεδομένα Το σύνολο δεδομένων Aff-Wild2, όπως έχουμε αναφέρει και στην Ενότητα 3.2, αποτελείται από βίντεο τα οποία έχουν δημιουργηθεί "in-the-wild", δηλαδή σε πραγματικές συνθήκες όπου οι άνθρωποι έχουν φυσικές αντιδράσεις και η ποιότητα των βίντεο καθώς και οι συνθήκες φωτισμού και οπτικής γωνίας διαφέρουν. Επίσης, σχετικά με το ηχητικό σήμα των βίντεο, υπάρχουν πολλές περιπτώσεις όπου ο ήχος προέρχεται από άλλη πηγή και όχι από το πρόσωπο που απεικονίζεται. Γεγονός που καταστεί το ακουστικό σήμα λιγότερο αξιόπιστο στη συμβολή του για την αναγνώριση των συναισθημάτων. Οι συγκεκριμένοι παράγοντες δυσκολεύουν κατά πολύ το μοντέλο αναγνώρισης συναισθημάτων. Επίσης το σύνολο δεδομένων είναι σε μεγάλο βαθμό μη ισορροπημένο και για τα δύο είδη ετικετών δημιουργώντας δυσκολία στην εκπαίδευση των μοντέλων.

Συνελικτικά Δίκτυα Τα συνελικτικά δίκτυα που επιλέχθηκαν για την εξαγωγή των χαρακτηριστικών ήταν προεκπαιδευμένα στο ImageNet, όπως έχει προαναφερθεί. Το ImageNet όμως περιέχει εικόνες από πράγματα, ζώα και φυτά και όχι εικόνες προσώπων, συνεπώς τα ήδη υπάρχοντα βάρη των μοντέλων μπορεί να μην βοηθάνε ιδιαίτερα στην απόδοσή τους. Επιπρόσθετα, τα συνελικτικά δίκτυα αυτά χρησιμοποιούνται για διάφορα προβλήματα αναγνώρισης και δεν έχουν αναπτυχθεί ειδικά για το πρόβλημα αναγνώρισης συναισθήματος.

Εκπαίδευση Η εκπαίδευση των βαθιά νευρωνικών δικτύων είναι μία διαδικασία που απαιτεί πολύ χρόνο και μεγάλη υπολογιστική ισχύ. Συγκεκριμένα, για την εκπαίδευση ενός μοντέλου στο σύνολο εκπαίδευσης Aff-Wild2, το οποίο αποτελείται από πάνω 3 εκατομμύρια εικόνες, απαιτείται χρόνος μέχρι και 2 ημερών. Το γεγονός αυτό μείωσε τα πειράματα που διενεργήσαμε ώστε να επιλέξουμε τις καλύτερες υπερπαραμέτρους και μεθόδους για την επιτυχή εκπαίδευση του κάθε μοντέλου.

5.3.1 Απόδοση Μοντέλων VA

Ανατρέχοντας στον Πίνακα 5.1, όπου παρουσιάζονται τα αποτελέσματα των μοντέλων VA, παρατηρούμε ότι οι υψηλότερες τιμές για τη μετρική αξιολόγησης CCC ανήκουν στα συνδυαστικά μοντέλα. Το γεγονός αυτό δείχνει ότι η εξαγωγή χαρακτηριστικών από δύο πηγές εισόδου ταυτόχρονα, οι οποίες είναι οι εικόνες των προσώπων και τα φασματογραφήματα του ήχου, βελτιώνουν την απόδοση των μοντέλων που χρησιμοποιούν μόνο μία από τα δύο είδη πληροφορίας. Συγκεκριμένα, η απόδοση για την τιμή του σθένους βελτιώνεται κατά σχεδόν 0.03 και η τιμή της διέγερσης κατά 0.06 σε σχέση με τα οπτικά μοντέλα. Η διέγερση παρουσιάζει μεγαλύτερη

βελτίωση με την προσθήκη της ηχητικής πληροφορίας, γεγονός αναμενόμενο καθώς όταν ένα συναίσθημα έχει ενεργητικό χαρακτήρα συνδυάζεται με έντονες ηχητικές αντιδράσεις ενώ στην αντίθετη περίπτωση υπάρχει απουσία ήχων. Σε γενικές γραμμές, οι αποδόσεις των μοντέλων είναι χαμηλές αλλά είναι αναμενόμενες λόγω της δυσκολίας του προβλήματος αναγνώρισης συναίσθημάτων αλλά και των παραγόντων που αναλύσαμε παραπάνω.

5.3.2 Απόδοση Μοντέλων Basic Expressions

Αντίθετα με τη περίπτωση του προβλήματος παλινδρόμησης, υψηλότερη απόδοση στο πρόβλημα ταξινόμησης των 7 βασικών συναίσθημάτων παρουσιάζουν τα οπτικά μοντέλα. Παρατηρώντας τον πίνακα των αποτελεσμάτων (Πίνακα 5.2), φαίνεται ότι η προσθήκη των φασματογραφημάτων του ήχου ως πληροφορία εισόδου μειώνει την απόδοση του καλύτερου οπτικού μοντέλου. Το γεγονός αυτό πιθανώς να συμβαίνει για το λόγο ότι τα βασικά συναίσθηματα αυτά αναγνωρίζονται κυρίως από τις εκφράσεις του προσώπου. Ακόμη, όπως αναφέραμε και στην αρχή της Ενότητας, το σύνολο δεδομένων είναι σε μεγάλο βαθμό μη ισορροπημένο, με το πλήθος των παραδειγμάτων του συναίσθηματος της Ουδετερότητας να ξεπερνάει κατά πολύ όλες τις υπόλοιπες κλάσεις. Συνεπώς, η ικανότητα του μοντέλου να αναγνωρίζει κλάσεις με πολύ μικρό πλήθος παραδειγμάτων είναι μειωμένη και έτσι οδηγούμαστε και σε χαμηλή τιμή της μετρικής αξιολόγησης F1-Score, η οποία είναι ευαίσθητη στην απόδοση όλων των κλάσεων ανεξάρτητα από το πλήθος των παραδειγμάτων που την αντιπροσωπεύουν.

Κεφάλαιο 6

Επίλογος και Μελλοντικές Επεκτάσεις

Στο Κεφάλαιο αυτό θα συνοψίσουμε το περιεχόμενο της παρούσας διπλωματικής εργασίας χαθώς και τα συμπεράσματα που απορρέουν από το σύνολο της θεωρητικής έρευνας αλλά και της πρακτικής εφαρμογής της για την ανάπτυξη λογισμικού. Έπειτα, θα γίνουν προτάσεις για μελλοντικές επεκτάσεις της παρούσας εργασίας, οι οποίες μπορούν να ερευνηθούν με σκοπό τη βελτίωση των αποτελεσμάτων και τη περαιτέρω εξέλιξη του θέματος που πραγματεύεται.

6.1 Συμπεράσματα Διπλωματικής Εργασίας

Στόχος της παρούσας διπλωματικής εργασίας ήταν η προσέγγιση του προβλήματος της αναγνώρισης ανθρώπινων συναίσθημάτων με μεθόδους μηχανικής μάθησης από οπτικοακουστικά δεδομένα. Στο πλαίσιο της ανάπτυξης μοντέλων για την επίλυση του παραπάνω προβλήματος έγινε εκτενής μελέτη του γνωστικού αντικειμένου της μηχανικής μάθησης και των νευρωνικών δικτύων, τομείς που έχουν προσελκύσει τεράστιο ενδιαφέρον της ερευνητικής κοινότητας της επιστήμης των υπολογιστών τα τελευταία χρόνια.

Επιπρόσθετα, η μελέτη των παλαιότερων και πιο πρόσφατων ερευνητικών δημοσιεύσεων σχετικά με τις προσπάθειες του συγκεκριμένου προβλήματος αναγνώρισης μας βοήθησε στην ανάπτυξη των δικών μας μοντέλων και την διαδικασία σχεδίασης και επιλογής διάφορων παραγόντων που συνέβαλαν στο τελικό αποτέλεσμα.

Λαμβάνοντας τις παραπάνω γνώσεις, καταφέραμε να οργανώσουμε και να υλοποιήσουμε τη πειραματική διαδικασία της προεπεξεργασίας των δεδομένων, της σχεδίασης των μοντέλων και την εκπαίδευση και αξιολόγησή τους. Με την ολοκλήρωση αυτών των διαδικασιών είχαμε την ευκαιρία να δουλέψουμε πάνω σε διαφορετικά υπολογιστικά συστήματα και να αναπτύξουμε ικανότητες στη συγγραφή κώδικα σε γλώσσα Python.

Ως τελικό συμπέρασμα της παρούσας εργασίας λαμβάνουμε το γεγονός ότι η αναγνώριση ανθρώπινων συναίσθημάτων από ένα υπολογιστικό σύστημα αποτελεί μεγάλη πρόκληση για τους επιστήμονες του κλάδου και σίγουρα η ερευνητική ενασχόληση με την ανάπτυξη μοντέλων για την επίτευξή της είναι μία δύσκολη αλλά και ταυτόχρονα δημιουργική διαδικασία. Το γεγονός αυτό φαίνεται και από την χαμηλή απόδοση των πλέον σύγχρονων συνελικτικών νευρωνικών δικτύων που χρησιμοποιήσαμε για την ανάπτυξη των μοντέλων μας στο συγκεκριμένο είδος

προβλήματος. Όπως είπαμε η δυνατότητα διάχρισης μεταξύ των διάφορων ανθρώπινων συναισθημάτων είναι δύσκολη ακόμα και για τον ίδιο τον άνθρωπο!

6.2 Μελλοντικές Επεκτάσεις

Το σύνολο δεδομένων παίζει βασικό ρόλο στην επιτυχή εκπαίδευση και την απόδοση των μοντέλων μηχανικής μάθησης. Το σύνολο δεδομένων Aff-Wild2 που χρησιμοποιήθηκε για την εκπαίδευση των μοντέλων μας ήταν πολύ επιτυχημένο όσο αναφορά τη ποικιλομορφία των ανθρώπων που απεικονίζονται και την αίσθηση των πραγματικών συνθηκών. Ωστόσο, η κατανομή των δεδομένων στις κλάσεις των βασικών συναισθημάτων αλλά και το εύρος τιμών του σθένους και της διέγερσης ήταν μη ισορροπημένη δυσκολεύοντας την εκπαίδευση των μοντέλων. Συνεπώς, μελλοντική επέκταση της διπλωματικής αποτελεί ο εμπλουτισμός του συνόλου δεδομένων Aff-Wild2 με άλλο παρεμφερές σύνολο με στόχο την πιο ισορροπημένη κατανομή των κλάσεων, γεγονός που πιθανώς να οδηγήσει σε καλύτερη απόδοση των μοντέλων.

Το σύνολο δεδομένων Aff-Wild2, όπως έχουμε αναφέρει, είναι μία συλλογή από βίντεο. Στη παρούσα εργασία εξάγουμε δύο είδη πληροφορίας από τα βίντεο αυτά, τις εικόνες των προσώπων από κάθε frame και τα φασματογραφήματα του ηχητικού σήματος για διαδοχικά παράθυρα δύο δευτερολέπτων. Ένα βίντεο όμως παρέχει και άλλες πληροφορίες που μπορούν να φανούν χρήσιμες για την αναγνώριση των συναισθημάτων. Η στάση του σώματος (body posing) και η ομιλία (speech) των ανθρώπων που απεικονίζονται αποτελούν σπουδαίες πληροφορίες που μπορούν να συμβάλλουν σημαντικά στη διαδικασία αναγνώρισης του συναισθήματος. Συγκεκριμένα, η κίνηση των χεριών και η στάση του κεφαλιού και των ώμων είναι τρόποι έκφρασης συναισθημάτων. Επίσης, το περιεχόμενο της ομιλίας και η αναφορά λέξεων με συναισθηματική φόρτιση αποτελούν βασικό τρόπο αναγνώρισης συναισθημάτων για τους ίδιους του ανθρώπους. Επομένως, η εξαγωγή της στάσης του σώματος σε μορφή μάσκας από τα βίντεο με χρήση σύγχρονων αλγορίθμων αλλά και η αναγνώριση φωνής και φυσικής γλώσσας από τον ήχο των βίντεο μπορούν να αποτελέσουν μελλοντική προσθήκη στα μοντέλα αναγνώρισης συναισθημάτων που έχουμε αναπτύξει.

Μια ακόμη πρόταση για τη μελλοντική επέκταση της εργασίας είναι η ανάπτυξη μοντέλων κάνοντας χρήση όχι μόνο Συνελικτικών Νευρωνικών Δικτύων αλλά και άλλα είδη νευρωνικών δικτύων όπως είναι τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks - RNN) και τα Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (Long Short-term Memory - LSTM). Τα δύο αυτά είδη νευρωνικών δικτύων έχουν την ιδιότητα να επεξεργάζονται δεδομένα που είναι σε ακολουθίες, καθώς μπορούν να αποθηκεύσουν πληροφορία στο εσωτερικό σαν να έχουν μνήμη. Τα βίντεο δεν είναι τίποτα άλλο από μία ακολουθία εικόνων, συνεπώς είναι εύλογη η επιλογή των δικτύων αυτών για το πρόβλημα της αναγνώρισης συναισθημάτων. Επίσης, ο παράγοντας του χρόνου παίζει σημαντικό ρόλο στην αναγνώριση των συναισθημάτων καθώς υπάρχει έντονη χρονική συσχέτιση όσο αναφορά την έκφρασή τους. Ακόμη και ο συνδυασμός CNN με RNN δίκτυα μπορεί να συμβάλει σε ένα πιο επιτυχημένο μοντέλο αναγνώρισης συναισθημάτων.

Πρόταση για μελλοντική έρευνα είναι και ο διαφορετικός τρόπος ενσωμάτωσης των οπτικοακουστικών δεδομένων. Για την σχεδίαση του συνδυαστικού μοντέλου επιλέξαμε τη μέθοδο Late Fusion, η οποία αναφέρθηκε στην Ενότητα 4.5.1, ενώ υπάρχουν και οι μέθοδοι Early Fusion και

Slow Fusion, οι οποίες αναφέρονται στην ίδια ενότητα. Η διαφορετική ενσωμάτωση των δεδομένων μπορεί να οδηγεί σε καλύτερη απόδοση των συνδυαστικών μοντέλων μέχρι αποδείξεως του αντιθέτου.

Βιβλιογραφία

- [1] Yannis Avrithis, Nicolas Tsapatsoulis, and Stefanos Kollias. Broadcast news parsing using visual cues: A robust face detection approach. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, volume 3, pages 1469–1472. IEEE, 2000.
- [2] Nicolas Tsapatsoulis and Stefanos Kollias. Face detection in color images and video sequences. In *2000 10th Mediterranean Electrotechnical Conference. Information Technology and Electrotechnology for the Mediterranean Countries. Proceedings. MeleCon 2000 (Cat. No. 00CH37099)*, volume 2, pages 498–502. IEEE, 2000.
- [3] Konstantinos Rapantzikos, Nicolas Tsapatsoulis, Yannis Avrithis, and Stefanos Kollias. Bottom-up spatiotemporal visual attention model for video analysis. *IET Image Processing*, 1(2):237–248, 2007.
- [4] Konstantinos Rapantzikos, Yannis Avrithis, and Stefanos Kollias. Spatiotemporal saliency for event detection and representation in the 3d wavelet domain: potential in human action recognition. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 294–301, 2007.
- [5] Konstantinos Rapantzikos, Nicolas Tsapatsoulis, Yannis Avrithis, and Stefanos Kollias. Spatiotemporal saliency for video classification. *Signal Processing: Image Communication*, 24(7):557–571, 2009.
- [6] Phivos Mylonas, Evangelos Spyrou, Yannis Avrithis, and Stefanos Kollias. Using visual context and region semantics for high-level concept detection. *IEEE Transactions on Multimedia*, 11(2):229–243, 2009.
- [7] Ilianna Kollia, Nikolaos Simou, Andreas Stafylopatis, and Stefanos Kollias. Semantic image analysis using a symbolic neural architecture. *Image Analysis & Stereology*, 29(3):159–172, 2010.
- [8] Nicolas Tsapatsoulis, Kostas Karpouzis, George Stamou, Frederic Piat, and Stefanos Kollias. A fuzzy system for emotion classification based on the mpeg-4 facial definition parameter set. In *2000 10th European Signal Processing Conference*, pages 1–4. IEEE, 2000.
- [9] Lori Malatesta, Amaryllis Raouzaiou, Kostas Karpouzis, and S Kollias. Towards modeling embodied conversational agent character profiles using appraisal theory predictions in expression synthesis. *Applied intelligence*, 30(1):58–64, 2009.

- [10] George Caridakis. Synthesizing gesture expressivity based on real sequences.
- [11] S Kollias and Dimitris Anastassiou. Adaptive training of multilayer neural networks using a least squares estimation technique. 1988.
- [12] Manolis Wallace, Ilias Maglogiannis, Kostas Karpouzis, George Kormentzas, and Stefanos Kollias. Intelligent one-stop-shop travel recommendations using an adaptive neural network and clustering of history. *Information Technology & Tourism*, 6(3):181–193, 2003.
- [13] Manolis Wallace, Nicolas Tsapatsoulis, and Stefanos Kollias. Intelligent initialization of resource allocating rbf networks. *Neural Networks*, 18(2):117–122, 2005.
- [14] Paraskevi Tzouveli, Andreas Schmidt, Michael Schneider, Antonis Symvonis, and Stefanos Kollias. Adaptive reading assistance for the inclusion of students with dyslexia: The agent-dysl approach. In *Proceedings of the 2008 Eighth IEEE International Conference on Advanced Learning Technologies*, pages 167–171, 2008.
- [15] Dimitris Kollias, George Marandianos, Amaryllis Raouzaiou, and Andreas-Georgios Stafylopatis. Interweaving deep learning and semantic techniques for emotion analysis in human-machine interaction. In *2015 10th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pages 1–6. IEEE, 2015.
- [16] Dimitrios Kollias, Athanasios Tagaris, and Andreas Stafylopatis. On line emotion detection using retrainable deep neural networks. In *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*, pages 1–8. IEEE, 2016.
- [17] Dimitrios Kollias, Miao Yu, Athanasios Tagaris, Georgios Leontidis, Andreas Stafylopatis, and Stefanos Kollias. Adaptation and contextualization of deep neural network models. In *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on*, pages 1–8. IEEE, 2017.
- [18] 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.
- [19] Dimitrios Kollias and Stefanos Zafeiriou. A multi-task learning & generation framework: Valence-arousal, action units & primary expressions. *arXiv preprint arXiv:1811.07771*, 2018.
- [20] Dimitrios Kollias and Stefanos P Zafeiriou. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Transactions on Affective Computing*, 2020.
- [21] 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*, pages 1–30, 2020.

- [22] 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.
- [23] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. Avec 2011—the first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction*, pages 415–424. Springer, 2011.
- [24] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516, 2013.
- [25] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018.
- [26] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800.
- [27] Dimitrios Kollias, Irene Kotsia, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition, 2021.
- [28] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959.
- [29] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharrshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.
- [30] Guillaume Lample and Devendra Singh Chaplot. Playing fps games with deep reinforcement learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [31] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.
- [32] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408, 1958.
- [33] Risto Miikkulainen, Jason Liang, Elliot Meyerson, Aditya Rawal, Daniel Fink, Olivier Francon, Bala Raju, Hormoz Shahrzad, Arshak Navruzyan, Nigel Duffy, et al. Evolving deep neural networks. In *Artificial intelligence in the age of neural networks and brain computing*, pages 293–312. Elsevier, 2019.
- [34] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*,

volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

- [35] Dominik Scherer, Andreas Müller, and Sven Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In Konstantinos Diamantaras, Wlodek Duch, and Lazaros S. Iliadis, editors, *Artificial Neural Networks – ICANN 2010*, pages 92–101, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.
- [36] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.
- [37] Christopher M Bishop. Pattern recognition. *Machine learning*, 128(9), 2006.
- [38] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.
- [39] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [40] 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.
- [41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.
- [42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
- [43] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [44] Dalgleish Tim and Power Mick. *Handbook of Cognition and Emotion*. Wiley, 2000.
- [45] Roddy Cowie and Randolph Cornelius. Describing the emotional states that are expressed in speech. *Speech Communication*, 40:5–32, 04 2003.
- [46] Cynthia M Whissell. The dictionary of affect in language, 113–131, robert plutchik and henry kellerman (ed.), emotion: Theory, research, and experience, 1989.
- [47] James A Russell. Evidence of convergent validity on the dimensions of affect. *Journal of personality and social psychology*, 36(10):1152, 1978.
- [48] R Plutchik. Emotion: A psychoevolutionary synthesis harper & row new york. 1980.

- [49] 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.
- [50] 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, 2010.
- [51] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*, pages 5 pp.–, 2005.
- [52] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. Best of Automatic Face and Gesture Recognition 2008.
- [53] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and M.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, 2006.
- [54] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A high-resolution 3d dynamic facial expression database. pages 1 – 6, 10 2008.
- [55] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3:5–17, 08 2013.
- [56] 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.
- [57] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65, 02 2017.
- [58] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A. Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal ‘in-the-wild’ challenge. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1980–1987, 2017.
- [59] Joan Alabert-i Medina, Epameinondas Antonakos, James Booth, Patrick Snape, and Stefanos Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 679–682, 2014.

- [60] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S Davis. Ssh: Single stage headless face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 4875–4884, 2017.
- [61] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark, 2015.
- [62] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.
- [63] J. Sundberg. *The Human Voice*, pages 1095–1104. Springer Berlin Heidelberg, Berlin, Heidelberg, 1996.
- [64] I. Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.
- [65] 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(6-7):907–929, Feb 2019.
- [66] Min Peng, Chongyang Wang, T. Chen, Guangyuan Liu, and Xiaolan Fu. Dual temporal scale convolutional neural network for micro-expression recognition. *Frontiers in Psychology*, 8, 10 2017.
- [67] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
- [68] Dimitrios Kollias, Viktoriia Sharmancka, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021.
- [69] 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.
- [70] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.
- [71] Dimitrios Kollias, Viktoriia Sharmancka, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019.
- [72] Dimitrios Kollias, Mihalis A Nicolaou, Irene Kotsia, Guoying Zhao, and Stefanos Zafeiriou. Recognition of affect in the wild using deep neural networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1972–1979. IEEE, 2017.

- [73] Dimitrios Kollias, Shiyang Cheng, Maja Pantic, and Stefanos Zafeiriou. Photorealistic facial synthesis in the dimensional affect space. In *European Conference on Computer Vision*, pages 475–491. Springer, 2018.
- [74] Dimitrios Kollias and Stefanos Zafeiriou. A multi-component cnn-rnn approach for dimensional emotion recognition in-the-wild. *arXiv preprint arXiv:1805.01452*, 2018.
- [75] Nikolaos Simou and Stefanos Kollias. Fire: A fuzzy reasoning engine for imprecise knowledge. Citeseer.
- [76] Yannis Avrithis, Yiannis Xirouhakis, and Stefanos Kollias. Affine-invariant curve normalization for object shape representation, classification, and retrieval. *Machine Vision and Applications*, 13(2):80–94, 2001.

