Εθνικό Μετσόβιο Πολυτεχνείο
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Ευφυών Συστημάτων, Περιεχομένου και Αλληλεπίδρασης.

Ενδιάμεσα Χαρακτηριστικά Μουσικής στη Νευρωνική Μάθηση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΒΑΣΙΛΕΙΟΣ ΛΥΜΠΕΡΑΤΟΣ

Επιβλέπων: Γ. Στάμου
Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Ευφυών Συστημάτων, Περιεχομένου και Αλληλεπίδρασης

Ενδιάμεσα Χαρακτηριστικά Μουσικής στη Νευρωνική Μάθηση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΒΑΣΙΛΕΙΟΣ ΛΥΜΠΕΡΑΤΟΣ

Επιβλέπων: Γ. Στάμου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 25η Φεβρουαρίου 2021

.....	.....	.....
Γ. Στάμου	Α.-Γ. Σταφυλοπάτης	Σ. Κόλλιας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2021

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

Περίληψη

Η ανάκτηση πληροφορίας από τη μουσική έχει ευνοηθεί από τη ψηφιοποίηση της μουσικής. Μεγάλος όγκος δεδομένων μουσικής είναι εύκολα διαθέσιμος. Η εφαρμογή της Μηχανικής Μάθησης σε αυτόν έχει απαντήσει σε πολλές προκλήσεις. Κάποιες από αυτές είναι η δημιουργία Συστημάτων Προτάσεων Μουσικής, η ταξινόμηση της μουσικής σε είδη και σε συναισθήματα ακόμα και η δημιουργία μουσικής. Στη διπλωματική αυτή ασχοληθήκαμε με το πρόβλημα της ταξινόμησης της μουσικής με βάση τα συναισθήματα. Η προσέγγιση μας βασίστηκε στη ταξινόμηση της σε πρώτο στάδιο σε ενδιάμεσα χαρακτηριστικά, τα οποία είναι εύκολα κατανοητά από τον άνθρωπο, και σε δεύτερο στάδιο στη σύνδεσή τους με τα συναισθήματα. Με αυτό το τρόπο δόθηκε ερμηνεία της ταξινόμησης των τραγουδιών με βάση τα συναισθήματα.

Στα πειράματα μας δοκιμάσαμε νέες αναπαραστάσεις της μουσικής με τη χρήση του Spotify API. Κάναμε πειράματα με πληθώρα αρχιτεκτονικών νευρωνικών δικτύων παρατηρώντας την επίδραση που έχει σε αυτά η ρύθμιση των υπερ-παραμέτρων τους. Ορισμένες από τις πολλές τεχνικές ομαλοποίησης που δοκιμάσαμε βασίζονται στην Εκμάθηση Αναπαραστάσεων (Representation Learning) και στη προ-εκπαίδευση (pre-training). Δοκιμάσαμε διάφορα πειράματα ταξινόμησης της μουσικής με τη αναπαράσταση χαρακτηριστικών του Spotify API.

Από τα πειράματα βγάλαμε συμπεράσματα για την ακρίβεια της αναπαράστασης της μουσικής με MFCCs και με χαρακτηριστικά του Spotify API. Συγκρίναμε τα διάφορα νευρωνικά δίκτυα που χρησιμοποιήσαμε με αντίστοιχα προηγούμενων ερευνών. Μελετήσαμε τη σημασία τεχνικών ομαλοποίησης για την εκπαίδευση των μοντέλων μας και δώσαμε παραδείγματα για τη σύνδεση των ενδιάμεσων χαρακτηριστικών με τα συναισθήματα.

Στο τέλος δώσαμε ορισμένες πιθανές κατευθύνσεις για μελλοντική εργασία ως αποτέλεσμα των ευρημάτων της διπλωματικής αυτής.

Λέξεις κλειδιά

Ανάκτηση Μουσικής Πληροφορίας, Αναγνώριση Συναισθήματος, Ενδιάμεσα Χαρακτηριστικά, Βαθιά Μάθηση, Συνελικτικά Νευρωνικά Δίκτυα, Τεχνητά Νευρωνικά Δίκτυα, Τεχνικές Ομαλοποίησης, Προ-Εκπαίδευση

Abstract

Music Information Retrieval has been favored by the digitization of music. A large amount of music data is readily available. The application of Machine Learning to music data has brought up a lot of MIR tasks. Some of them are the creation of Recommender Systems, the classification of music to genre and emotion as well as the creation of music. In this diploma thesis we dealt with the task of music classification on the basis of emotions. Our approach was based on the linear regression of music to mid-level features, features that are easily conceived by humans, and then the linear regression of mid-level features to emotions. In this way we obtained an interpretation of the classification of songs to emotions.

In our experiments we tested new representations of music given from Spotify API. We tested many neural networks and observed the impact of hyperparameters on them. We tried many regularization methods, some of them based on Representation Learning and on Pre-training. We ran a lot of experiments on classification tasks, using the representation of music given from Spotify API.

We reached conclusions for the quality of the different ways of the representation of music. We compared the different neural networks that we used. We studied the importance of the regularization methods on the training of our models and gave examples for the connection of the mid-level features with emotions.

In the end we suggested some possible directions that future investigations could follow based on our this thesis work.

Key words

Music Information Retrieval, Music Emotion Recognition, Mid-level Features, Deep Learning, Convolutional Neural Networks, Artificial Neural Networks, Regularization methods, Pre-training

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή μου Γ.Στάμου για τη δυνατότητα και την έμπνευση που μου έδωσε για την ενασχόληση μου με την Νευρωνική Μάθηση και τη Μουσική, τον Ε.Δερβάκο για την καθοδήγηση και βοήθεια που μου παρείχε και την οικογένεια μου και την Ε.Γ. για τη στήριξη που μου δώσανε.

Περιεχόμενα

Περίληψη 5

Abstract 7

Ευχαριστίες 11

Περιεχόμενα 13

Κατάλογος πινάκων 15

Κατάλογος σχημάτων 17

1. Εισαγωγή 19
   1.1. Τα Ενδιάμεσα Χαρακτηριστικά στη Μουσική 19
   1.2. Κίνητρο 20
   1.3. Συνεισφορά 20
   1.4. Δομή 20

2. Θεωρία 23
   2.1. Μουσική και Πληροφορία 23
      2.1.1. Music Information Retrieval 23
      2.1.2. Music Emotion Recognition 23
      2.1.3. Ερμηνεία Music Emotion Recognition 27
   2.2. Βαθιά Μάθηση 29
      2.2.1. Τεχνητή Νοημοσύνη 29
      2.2.2. Μηχανική Μάθηση 29
      2.2.3. Βαθιά Μάθηση 30
      2.2.4. Τεχνητά Νευρωνικά Δίκτυα 30
      2.2.5. Συνελικτικά Νευρωνικά Δίκτυα 34
      2.2.6. Αυτοκωδικοποιητές 35

3. Σχεδιασμός και Υλοποίηση Πειραμάτων 39
   3.1. Περιβάλλον Υλοποίησης 39
   3.2. Δεδομένα 39
      3.2.1. Σύνολα Δεδομένων 40
         3.2.1.1. Aljanaki’s Mid-level Perceptual Features 40

3.2.1.2. Σύνολα Δεδομένων για Προ-Εκπαίδευση 41
3.2.2. Αναπαράσταση Δεδομένων 42
3.2.2.1. Librosa 42
3.2.2.2. Spotify 43
3.2.3. Προ-επεξεργασία δεδομένων 47
3.3. Μετρικές Εκπαίδευσης. 49
3.4. Υπερ-παράμετροι Εκπαίδευσης 50
3.5. Νευρωνικά Δίκτυα 54
3.5.1. Τεχνικά Νευρωνικά Δίκτυα 55
3.5.2. Συνελικτικά Νευρωνικά Δίκτυα 56
3.5.2.1. Librosa 56
3.5.2.2. Spotify 58
3.5.3. Αναδρομικά Νευρωνικά Δίκτυα 60
3.5.4. Αυτοκωδικοποιητές 61
3.6. Τεχνικές Ομαλοποίησης 62
3.6.1. Dropout 62
3.6.2. Όροι Ποινής 63
3.6.3. Early Stopping 64
3.6.4. Data Augmentation 65
3.6.5. Unsupervised Pre-training 66
3.6.6. Supervised Pre-training 70
3.6.7. Multitask Learning 73

4. Αξιολόγηση Πειραμάτων και Συμπεράσματα 77
4.1. Τρόποι Αναπαράστασης της Μουσικής. 77
4.2. Αρχιτεκτονικές Νευρωνικών Δικτύων 79
4.3. Τεχνικές Ομαλοποίησης. 80
4.4. Σύγκριση με Παλαιότερες Εργασίες 83
4.5. Συναισθήματα και Ενδιάμεσα Χαρακτηριστικά 85

5. Επίλογος 87
5.1. Σύνοψη 87
5.2. Μελλοντικές εργασίες 88

Βιβλιογραφία 89

Κατάλογος Πινάκων

2.1 Οι 5 ομάδες συναισθήματος που όρισε το MIREX το 2007 25
2.2 Confusion matrix για την επικάλυψη των MIREX συναισθήματων. Οι τιμές που είναι σημειωμένες με αστερίσκο αντιστοιχούν στις πιο όμοιες ομάδες 25

3.1 Τα ενδιάμεσα χαρακτηριστικά και οι ερωτήσεις που έγιναν στους βαθμολογητές προκειμένου να καταλάβουνε τις έννοιες 41
3.2 Περιγραφή του segment αντικειμένου 43
3.3 Περιγραφή των audio features του Spotify API 45
3.4 Το Pearsons Correlation των τιμών πρόβλεψης με των πραγματικών τιμών για διαφορετικές μεθόδους κανονικοποίησης των δεδομένων . . 48
3.5 Περιέχει τους κυριότερους αλγορίθμους αυτοματοποιημένης ρύθμισης των υπερ-παραμέτρων 52
3.6 Παράδειγμα χρήσης του αλγορίθμου grid search για την εύρεση κατάλληλης τιμής για το batch_size 53
3.7 Παράδειγμα χρήσης του αλγορίθμου ASHA για την εύρεση κατάλληλης τιμής για το learning_rate 53
3.8 Πλήρως συνδεδεμένα δίκτυα 55
3.9 Οι τιμές Pearson Correlation για τα ενδιάμεσα χαρακτηριστικά . 57
3.1 Οι τιμές Pearson Correlation για τα συναισθήματα 58
3.11 Συνελιτικά Δίκτυα 58
3.12 Αναδρομικά Δίκτυα 60
3.13 Ο κατάλληλος αριθμός “παγωμένων” επιπέδων για κάθε μοντέλο 68
3.14 Οι τιμές Pearson Correlation για τα ενδιάμεσα χαρακτηριστικά . 74
3.15 Οι τιμές Pearson Correlation για τα συναισθήματα 74

4.1 Σύγκριση επιδόσεων CNN και ANN μοντέλων 79
4.2 Οι τιμές Pearson Correlation για τα συναισθήματα 83
4.3 Οι τιμές Pearson Correlation για τα ενδιάμεσα χαρακτηριστικά . 84

Κατάλογος Σχημάτων

2.1 Russell’s Circumplex Model of Affect 24
2.2 Τα διαφορετικά επίπεδα ανάλυσης του ήχου 27
2.3 Περιγραφή της λειτουργίας του AudioLime 28
2.4 Το Perceptron του Rosenbalt 31
2.5 Η γραφική παράσταση της Sigmoid 32
2.6 Η γραφική παράσταση της Tanh 33
2.7 Η γραφική παράσταση της ReLU 33
2.8 Δύο τρόποι εφαρμογής CNN δικτύου αναδεικνύοντας τα δεκτικά τους πεδία 35
2.9 Η αρχιτεκτονική του δικτύου “A Universal Music Translation Network” 36

3.1 Τα είδη τραγουδιών σε WordCloud 42
3.2 Παράδειγμα αρχείου MFCC 43
3.3 Απεικόνιση της pitch αναπαράστασης 44
3.4 Οι 12 βασικές συναρτήσεις για την αναπαράσταση της “χροιάς” 44
3.5 Η κατανομή της κλίμακας και του κλειδιού των δεδομένων της Aljanaki’s dataset 46
3.6 Η κατανομή των ενδιάμεσων χαρακτηριστικών στην Aljanki’s dataset 46
3.7 Τα audio features που χρησιμοποιούνται εμβόλιμα στην αρχιτεκτονική τωνδικτύων 47
3.8 Σύγκριση των κανονικοποιήσεων MinMax και Standard στην αρχιτεκτονική A2Mid_Small 48
3.9 Απεικόνιση του ομαλού και αιχμηρού ελαχίστου 51
3.10 Σύγκριση αλγορίθμων Random και Grid Search 52
3.11 3 πειράματα για κάθε συνδυασμό υπερ-παραμέτρων στο A2Mid μοντέλο 54
3.12 Σύγκριση πλήρως συνδεδεμένων δικτύων 55
3.13 Η αρχιτεκτονική του Dense 56

3.14 Η αρχιτεκτονική του A2Mid2E 57

3.15 Σύγκριση Συνελικτικών Δικτύων 59

3.16 Η αρχιτεκτονική του A2Mid_Medium 59

3.17 Σύγκριση Αναδρομικών Δικτύων 60

3.18 Αυτοκωδικοποιητής με πλήρως συνδεδεμένα δίκτυα (Autoencoder_Dense_Small) 61

3.19 Αυτοκωδικοποιητής με συνελικτικα δίκτυα (Autoencoder_Small) 61

3.20 Πριν και μετά τη χρήση dropout επιπέδων 62

3.21 Παράδειγμα εφαρμογής L2 regularization στο μοντέλο A2Mid . . . 63

3.22 Οι διακεκομμένες γραμμές αντιστοιχούν στο L2 regularization ενώ οι συνεχής στο Early Stopping 64

3.23 Αριστέρα εκπαίδευση με Early Stopping. Δεξιά εκπαίδευση χωρίς Early Stopping 64

3.24 Data augmentation στην εκπαίδευση του μοντέλου Dense_Small και του A2Mid 65

3.25 Απεικόνιση φίλτρων από ένα DBN εκπαιδευμένο στην InfiniteMNIST. Οι πάνω εικόνες απεικονίζουν τα βάρη των φίλτρων χωρίς να έχει γίνει pre-training, ενώ οι κάτω συμβολίζουν τα βάρη των φίλτρων με το να έχει γίνει pre-training. Από αριστερά προς τα δεξιάς έχουμε τους νευρώνες του 1^{ου}, 2^{ου} και 3^{ου} επιπέδου 66

3.26 Εφαρμογή Unsupervised Pre-training σε όλα τα μοντέλα στην Pre-training dataset 67

3.27 Σύγκριση μοντέλων A2Mid_Small με διαφορετικό αριθμό παγωμένων επιπέδων (Pre-training) 68

3.28 Εφαρμογή Unsupervised Pre-training σε διάφορα μοντέλα στην Large-Pre-training dataset 69

3.29 Σύγκριση μοντέλων A2Mid_Small με διαφορετικό αριθμό παγωμένων επιπέδων (Large-Pre-training) 69

3.30 Εφαρμογή Supervised Pre-training σε δίαφορα μοντέλα στην Pre-training dataset 70

3.31 Σύγκριση μοντέλων A2Mid_Medium διαφορετικό αριθμό παγωμένων επιπέδων (Large-Pre-training Supervised) 71

3.32 Γραφικές παραστάσεις που αναδεικνύουν την επίδραση του Supervised Pre-training στη μεγάλη dataset 71

3.33 Εφαρμογή Supervised Pre-training σε δίαφορα μοντέλα στην Pre-training dataset με χρήση του convnet feature 72

3.34 Παράδειγμα Multi-task Learning. Το x είναι η είσοδος, το $h^{(shared)}$ είναι οι κοινές κρυμμένες μονάδες, τα $h^{(1)}$, $h^{(2)}$, $h^{(3)}$ είναι κρυμμένες μονάδες διαφορετικών προβλημάτων, τα $y^{(1)}$, $y^{(2)}$ είναι έξοδοι διαφορετικών προβλημάτων 73

3.35 Η αρχιτεκτονική του A2Mid2E-Joint 74

4.1 Το accuracy του CNN μοντέλου πρόβλεψης των 5 ομάδων του MIREX 78

4.2 Το accuracy του CNN μοντέλου πρόβλεψης των 4 ομάδων του Moodylyrics 78

4.3 A2Mid_Medium που έχει προ-εκπαιδευτεί με Supervised Pre-training με τη χρήση και μη του convnet χαρακτηριστικού 80

4.4 A2Mid_Medium που έχει προ-εκπαιδευτεί με Supervised Pre-training και με Unsupervised Pre-training 80

4.5 Μοντέλα τα οποία έχουν προ-εκπαιδευτεί στη Pre-training dataset και στη Large-Pretraing dataset 81

4.6 Το μοντέλο A2Mid_Small προ-εκπαιδευμένο με τη Pre-training dataset (αριστερά) και με τη Large-Pre-training dataset (δεξιά) με “παγωμένο” διαφορετικό αριθμό επιπέδων κάθε φορά 82

4.7 Το μοντέλο A2Mid_Medium προ-εκπαιδευμένο με τη Pre-training dataset (αριστερά) και με τη Large-Pre-training dataset (δεξιά) με “παγωμένο” διαφορετικό αριθμό επιπέδων κάθε φορά 82

4.8 Το μοντέλο A2Mid προ-εκπαιδευμένο με τη Pre-training dataset (αριστερά) και με τη Large-Pre-training dataset (δεξιά) με “παγωμένο” διαφορετικό αριθμό επιπέδων κάθε φορά 83

4.9 Correlation matrix των πραγματικών τιμών των ενδιάμεσων χαρακτηριστικών με τα συναισθήματα 84

4.10 Correlation matrix των βαρών του μοντέλου A2Mid2E-Joint 85

4.11 Boxplot του effect του μοντέλου A2Mid2E-Joint 85

Κεφάλαιο 1

ΕΙΣΑΓΩΓΗ

Τα τελευταία χρόνια πολύ μεγάλο μέρος της ανθρώπινης δραστηριότητας του ανθρώπου έχει μεταφερθεί σε ψηφιακή μορφή. Υπάρχουν πολλά παραδείγματα στις καθημερινές σχέσεις, στην εργασία, στην εκπαίδευση, στη διασκέδαση ακόμα και στην τέχνη. Ειδικά στη μουσική, έχει παίξει σημαντικό ρόλο η αλλαγή της αναλογικής της μορφής στην ψηφιακή. Οι λόγοι είναι ότι με τη νέα μορφή της είναι εύκολη η αποθήκευση, η προστασία, η μεταφορά, η ανταλλαγή και η μελέτη της. Ο μεγάλος όγκος δεδομένων μουσικής που υπάρχει ευνοεί την εφαρμογή συστημάτων Μηχανικής Μάθησης για την επεξεργασία της. Η μελέτη της έχει συνδράμει στην ταξινόμηση της σε είδη και σε συναισθήματα, στην ενίσχυση Συστημάτων Προτάσεων Μουσικής, ακόμα και στη δημιουργία νέων μουσικών τραγουδιών, σε σημείο που δεν είναι ευδιάκριτο το αν οι δημιουργοί τους είναι άνθρωποι ή μηχανές. Στην παρούσα διπλωματική θα ασχοληθούμε με το ζήτημα της ερμηνείας της ταξινόμησης των τραγουδιών σε συναισθήματα. Η ερμηνεία δίνεται μέσω των ενδιάμεσων χαρακτηριστικών της μουσικής.

1.1 Τα Ενδιάμεσα Χαρακτηριστικά στη Μουσική

Πολλά συστήματα μηχανικής μάθησης έχουν προσπαθήσει να ταξινομήσουν τα τραγούδια σε συναισθήματα, πολλές φορές με επιτυχή τρόπο. Ένα από τα προβλήματα βρίσκεται όμως στο κομμάτι της ερμηνείας των αποτελεσμάτων, που δεν είναι εύκολη ειδικά όσον αφορά τη μουσική. Τη λύση σε αυτό το πρόβλημα προσπαθούν να δώσουν τα ενδιάμεσα χαρακτηριστικά. Τα ενδιάμεσα χαρακτηριστικά είναι μουσικά χαρακτηριστικά που μπορεί να καταλάβει ένας άνθρωπος χωρίς ιδιαίτερες μουσικές γνώσεις. Είναι χαρακτηριστικά, που έχουν μουσικό νόημα, όπως η τονικότητα, η σταθερότητα στο ρυθμό ή η κλίμακα. Αν χρησιμοποιηθούν ως ενδιάμεσο στάδιο για την ταξινόμηση των τραγουδιών σε συναίσθημα δίνουν κατανοητές για τον άνθρωπο εξηγήσεις για το πως λειτουργεί το μοντέλο και για ποιο λόγο επέλεξε να κάνει τις συγκεκριμένες ταξινομήσεις. Είναι μια πρόκληση αρκετά χρήσιμη και απαντά στη τάση που υπάρχει να χρησιμοποιούνται μοντέλα Μηχανικής Μάθησης στη μουσική σαν μαύρα κουτιά.

1.2 Κίνητρο

Βασικό κίνητρο για τη διπλωματική είναι το ενδιαφέρον για τη Μουσική και την Τεχνητή Νοημοσύνη. Τα τελευταία χρόνια υπάρχει έντονη και ενδιαφέρουσα ερευνητική δραστηριότητα στον τομέα αυτόν. Παρ' όλα αυτά η χρήση τη Βαθιάς Μάθησης στη μουσική είναι σε αρχικά στάδια, σε σχέση τουλάχιστον με τη χρήση της στην εικόνα και υπάρχει αναγκαιότητα ενίσχυσής της. Ο λόγος βρίσκεται στο ότι η μουσική αναπαρίσταται σε δυσνόητη μορφή. Έμπνευση και βάση για τη συγκεκριμένη διπλωματική είναι το έργο που έχουν κάνει οι Verena Haunschmid και Gerhard Widmer γύρω από την ερμηνεία προβλέψεων συναισθήματος τραγουδιών.

1.3 Συνεισφορά

Στη διπλωματική αυτή μελετάμε τη σύνδεση του συναισθήματος με τα ενδιάμεσα χαρακτηριστικά της μουσικής, δίνοντας ερμηνεία σε προβλέψεις συναισθήματος. Χρησιμοποιούμε έναν νέο τρόπο αναπαράστασης των τραγουδιών μέσω του Spotify API, με στόχο την καλύτερη γενίκευση μέσω Representation Learning. Δοκιμάζουμε και συγκρίνουμε διάφορες τεχνικές ομαλοποίησης, πετυχαίνοντας ενδιαφέροντα αποτελέσματα προς συζήτηση. Μελετάμε πλήθος μοντέλων βαθιάς μάθησης για την πρόβλεψη των ενδιάμεσων χαρακτηριστικών, παρατηρώντας τη σύνδεση τους με τις παραμέτρους και τις υπερ-παραμέτρους της εκπαίδευσης.

1.4 Δομή

Στο Κεφάλαιο 2 δίνεται η απαραίτητη θεωρία για τη κατανόηση της διπλωματικής αυτής. Αρχικά σχολιάζονται επιστημονικά αντικείμενα της επεξεργασίας της μουσικής. Δίνονται ορισμοί και μικρές ιστορικές αναφορές στο Music Emotion Recognition. Επιπλέον δίνονται κάποια παραδείγματα παρεμφερών εργασιών στην ερμηνεία του MER. Στο δεύτερο μέρος του κεφαλαίου δίνονται ορισμοί για τα βασικά επιστημονικά αντικείμενα που χρησιμοποιήσαμε όσον αφορά την Τεχνητή Νοημοσύνη.

Στο Κεφάλαιο 3 δίνονται όλα τα απαραίτητα στοιχεία για το σχεδιασμό και την υλοποίηση των πειραμάτων μας. Περιγράφονται τα εργαλεία που χρησιμοποιήσαμε, τα δεδομένα που διαλέξαμε και τους διαφορετικούς τρόπους αναπαράστασης τους, τις αρχιτεκτονικές των δικτύων που επιλέξαμε και τα αποτελέσματα των επιδόσεων τους και τέλος οι τεχνικές ομαλοποίησης που εφαρμόσαμε.

Στο Κεφάλαιο 4 γίνεται αποτίμηση των αποτελεσμάτων των πειραμάτων μας. Βγαίνουν συμπεράσματα για τους τρόπους αναπαράστασης της μουσικής, για τις αρχιτεκτονικές των δικτύων και για τις τεχνικές ομαλοποίησης. Τέλος γίνεται σύγκριση με παλαιότερες παρόμοιες δουλειές.

Στο Κεφάλαιο 5 γίνεται σύνοψη της διπλωματικής και περιγράφονται μελλοντικές πιθανές προεκτάσεις της.

Κεφάλαιο 2

Θεωρία

Στο Κεφάλαιο 2 θα δώσουμε τους απαραίτητους ορισμούς και το επιστημονικό υπόβαθρο για την κατανόηση της διπλωματικής εργασίας. Το Κεφάλαιο χωρίζεται σε δύο βασικά μέρη. Το πρώτο ασχολείται με τη Μουσική και την Πληροφορία και το δεύτερο με την Τεχνητή Νοημοσύνη και τους τομείς της που χρησιμοποιήσαμε.

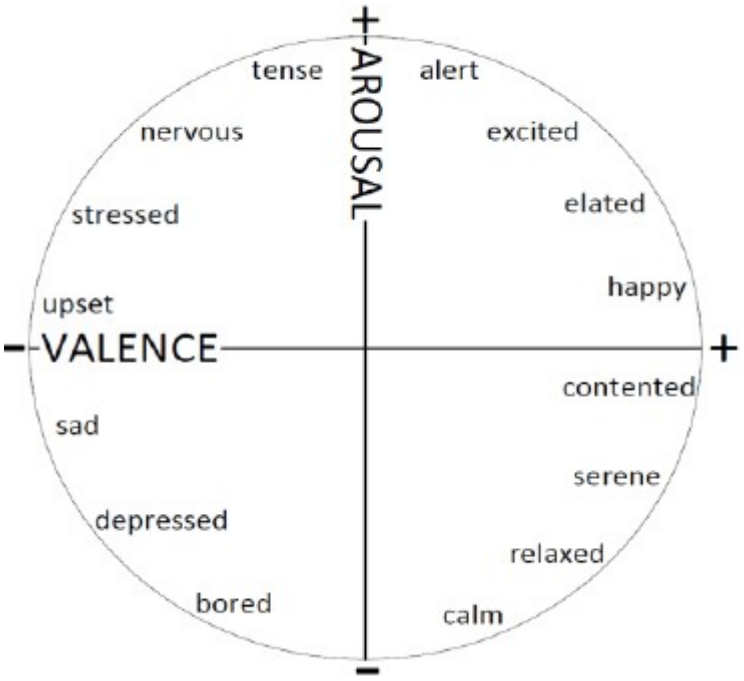
2.1 Μουσική και Πληροφορία

2.1.1 Music Information Retrieval

Το Music information retrieval (MIR) είναι ένα διεπιστημονικό πεδίο που ασχολείται με τη ανάκτηση πληροφορίας από τη μουσική. Είναι περιορισμένο το πεδίο της έρευνας, αλλά έχει πολλές χρήσιμες εφαρμογές στην πραγματικότητα. Βασικά γνωστικά αντικείμενα που είναι προαπαιτούμενα είναι η μουσικολογία, η ψυχολογία, η ψηφιακή επεξεργασία σήματος, η μηχανική μάθηση, η πληροφορική και πολλά άλλα. Μερικές χρήσιμες εφαρμογές είναι τα Συστήματα Προτάσεων, Track separation and instrument recognition, Automatic music transcription, Automatic categorization (της οποίας υποκατηγορία είναι το Music Emotion Recognition), Music generation και Music Style Transfer.

2.1.2 Music Emotion Recognition

Το Music Emotion Recognition είναι ένα κομμάτι του MIR που σχετίζεται με την αναγνώριση των ανθρώπινων συναισθημάτων που εκφράζει η μουσική. Είναι μια πρόκληση δύσκολη γιατί η σύνδεση της μουσικής και των συναισθημάτων είναι υποκειμενική και ανεξήγητη. Εδώ είναι καλό να διευκρινιστεί ότι η αναγνώριση δεν έχει να κάνει με τα συναισθήματα που μπορεί να προκαλέσει η μουσική, καθώς αυτό καθορίζεται από ασταθείς παράγοντες όπως το κοινωνικό πλαίσιο της μουσικής εμπειρίας ή το προσωπικό κίνητρο του ακροατή (συγκέντρωση για διάβασμα). Διαφορετικά, ο στόχος της είναι η αναγνώριση των συναισθημάτων που εκφράζονται με τη μουσική.



Σχήμα 2.1 : Russell’s Circumplex Model of Affect.

Ένα ερώτημα είναι ο τρόπος κατηγοριοποίησης του συναισθήματος. Ο επικρατέστερος είναι μέσω του κύκλου του Russel, όπου τοποθετείται το συναίσθημα σε έναν δισδιάστατο συνεχή χώρο, όπου ο οριζόντιος άξονας αντιστοιχεί στο σθένος του τραγουδιού και ο κάθετος στη διέγερση του, αναπαριστώντας με αυτό το τρόπο οποιοδήποτε συναίσθημα. Ένας διαφορετικός τρόπος είναι η κατηγοριοποίηση των τραγουδιών σε ομάδες. Η πρώτη προσέγγιση έγινε από τον Henver το 1936, ο οποίος χώρισε τη μουσική σε 8 βασικές ομάδες. Η πιο σύγχρονη προσέγγιση είναι από το MIREX το 2007, που χώρισε τη μουσική σε 5 ομάδες σύμφωνα με κριτήρια μουσικών και ψυχολόγων. Ο κύκλος του Russel είναι προτιμητέος γιατί δίνει λύση στο δύσκολο πρόβλημα του σαφή προσδιορισμού των συναισθημάτων από τη στιγμή που παίρνει συνεχείς τιμές. Εκτός αυτού, οι ομάδες συναισθημάτων είναι πιθανό να επικαλύπτονται. Ένα παράδειγμα δίνεται από τον πίνακα 2.2 όπου εμφανίζεται το ποσοστό επικάλυψης μεταξύ των ομάδων του συνόλου δεδομένων της MIREX (Cyril Laurier).

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Passionate	Rollicking	Literate	Humorous	Aggresive
Rousing	Cheerful	Poignant	Silly	Fiery
Confident	Fun	Wistful	Campy	Tense/Anxious
Boisterous	Sweet	Bittersweet	Quirky	Intense
Rowdy	Amiable/Good Natured	Autumnal	Whimsical	Volatile
		Brooding	Witty	Visceral
			Wry	

Πίνακας 2.1 : Οι 5 ομάδες συναισθήματος που όρισε το MIREX το 2007.

	C1	C2	C3	C4	C5
C1	0	0.74	0.128*	0.204	0.108*
C2	0.74	0	0.859	0.816	**0.876**
C3	0.128*	0.859	0	0.319	0.265
C4	0.204	0.816	0.319	0	0.526
C5	0.108*	**0.876**	0.265	0.526	0

Πίνακας 2.2 : Confusion matrix για την επικάλυψη των MIREX ομάδων συναισθήματος. Οι τιμές που είναι σημειωμένες με αστερίσκο αντιστοιχούν στις πιο όμοιες ομάδες.

Ένα άλλο κρίσιμο ερώτημα είναι το ποια θα είναι η πηγή για να βαθμολογηθεί το συναίσθημα. Επιγραμματικά οι κύριες πηγές σύμφωνα με τους Youngmoo E. Kim et al. είναι οι εξής: Ερωτηματολόγια σε κοινό ή σε επαγγελματίες μουσικούς και ψυχολόγους. Στοχευμένα παιχνίδια σε κοινό. Ετικέτες που διαλέγει το κοινό σε κοινωνικά blog. Εξαγωγή πληροφοριών από αρχεία στο διαδίκτυο σχετικά με το όνομα του τραγουδιού. Ανάλυση του συναισθηματικού φορτίου των στίχων των τραγουδιών. Ανάλυση του ήχου του τραγουδιού. Ο ακριβέστερος προσδιορισμός του συναισθήματος επιτυγχάνεται με την αξιοποίηση πολλαπλών πηγών προσδιορισμού του. Αυτό οφείλεται στο ότι το συναίσθημα δεν βρίσκεται μόνο στον ήχο αλλά και σε άλλα χαρακτηριστικά της μουσικής.

Παρόλο που οι πηγές προσδιορισμού του συναισθήματος είναι πολλές, η μεγάλη αύξηση των δεδομένων λόγω της ψηφιοποίησης των τραγουδιών και το διαδίκτυο οδήγησε στην ανάγκη χρησιμοποίησης αυτοματοποιημένων μηχανισμών ταξινόμησης του συναισθήματος.

Παρακάτω δίνονται επιγραμματικά ορισμένα ιστορικά παραδείγματα μηχανικής μάθησης που κινήθηκαν σε αυτήν την κατεύθυνση.

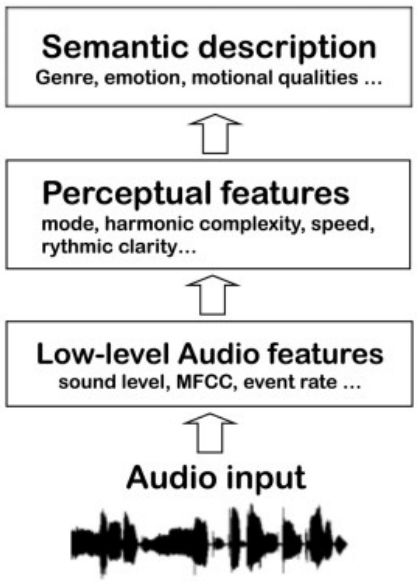
Το MIREX συνέδριο το 2007 ξεκινάει μια μεθοδική προσέγγιση του προβλήματος. Όρισε διαγωνισμό για το “Audio Mood Classification”, όπου έγινε κατηγοριοποίηση των τραγουδιών σε 5 ομάδες διάθεσης (Xiao Hu et al.). Έγιναν δοκιμές διάφορων τεχνικών μηχανικής μάθησης (KNN, SVM). Τα καλύτερα αποτελέσματα ήρθαν από τον Τζανετάκη ο οποίος είχε 61.50% ακρίβεια με χρήση Support Vector Machines (SVM) και αναπαράσταση MFCCs (Perfecto Herrera et al.). Εκείνη τη περίοδο υπάρχουν οι πρώτες προσπάθειες χρήσης πολλαπλής αναπαράστασης των τραγουδιών και χρήση των στίχων για προβλήματα ταξινόμησης (Robert Neumayer et al., Cyril Laurier). Η τάση αυτή οδήγησε το MIREX το 2013 στην δημοσίευση ενός πολύτροπου συνόλου δεδομένων όπου περιέχει τραγούδια σε αναπαράσταση MIDI, σε AUDIO και σε lyrics (R. Panda et al.). Σταδιακά άρχισαν να χρησιμοποιούνται μοντέλα βαθιάς μάθησης στο MER (Alexander Schindler et al., Tong Liu et al.) τα οποία βοήθησαν στην επίδοση των συστημάτων. Σε τελευταίες δημοσιεύσεις, φαίνεται ότι η καλύτερη προσέγγιση του MER είναι μέσω πολλαπλής αναπαράστασης των τραγουδιών και χρήσης μοντέλων βαθιάς μάθησης (Remi Delbouys et al., Κωσταντίνος Πυροβολάκης).

Παρόλο που υπάρχουν τρόποι να αποδοθεί το συναίσθημα σε τραγούδια μέσω μηχανικής μάθησης, είναι δύσκολο να καταλάβει κανείς πως οδηγήθηκε το μοντέλο στις προβλέψεις του. Στην επεξεργασία εικόνας, όταν τα μοντέλα κάνουν κάποια πρόβλεψη, καταλαβαίνει κανείς για ποιο λόγο την έκαναν μελετώντας τα βάρη των ενδιάμεσων επιπέδων. Αντιθέτως στη μουσική δεν είναι εύκολα ερμηνεύσιμα, καθώς η αναπαράσταση της μουσικής δεν είναι εύληπτη.

2.1.3 Ερμηνεία Music Emotion Recognition

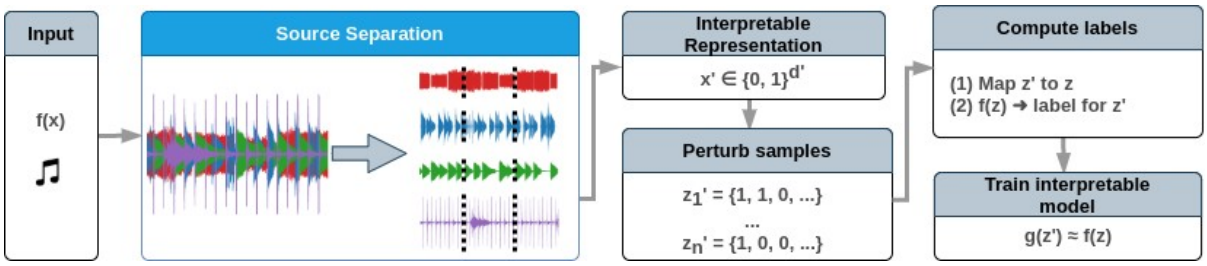
Η προσέγγιση των ενδιάμεσων χαρακτηριστικών προσπάθησε να δώσει τις κατάλληλες εξηγήσεις για τις προβλέψεις των MER μοντέλων. Τα ενδιάμεσα χαρακτηριστικά ενός τραγουδιού είναι χαρακτηριστικά που μπορούν να είναι κατανοητά από έναν άνθρωπο που δεν έχει απαραίτητα μουσική παιδεία. Με αυτό τον τρόπο, αν κάποιος σε ένα τραγούδι κάνει τη πρόβλεψη των ενδιάμεσων χαρακτηριστικών του και ύστερα τα αντιστοιχήσει γραμμικά με κάποια πιθανά συναισθήματα, μπορεί να καταλάβει τη σχέση των ενδιάμεσων χαρακτηριστικών με τα συναισθήματα. Για παράδειγμα σε ένα τραγούδι που προβλέπεται ότι είναι σε κλίμακα ματζόρε και στη συνέχεια αντιστοιχίζεται στο συναίσθημα της χαράς, μπορεί να καταλάβει κανείς για ποιο λόγο έκανε αυτή τη πρόβλεψη.

Οι πρώτες αναφορές στα ενδιάμεσα χαρακτηριστικά γίνονται από τον Lace Wedin, ο οποίος προσπαθεί να συνδυάσει τα ενδιάμεσα χαρακτηριστικά με τα συναισθήματα μουσικών. Ύστερα οι Aalf Gabrielsson et al. θα γράψουν για τη σχέση της δομής της μουσικής με το συναίσθημα. Μετά θα ακολουθήσουν οι Anders Friberg et al. που θα προσπαθήσουν να εξάγουν ενδιάμεσα χαρακτηριστικά από MIDI και Audio αρχεία. Από τις πιο σύγχρονες προσεγγίσεις είναι αυτή των Aljanaki et al. όπου συγκροτούν σύνολο δεδομένων (dataset) 5000 τραγουδιών, χαρακτηρισμένων με ενδιάμεσα χαρακτηριστικά από μουσικούς. Αυτό το σύνολο δεδομένων θα χρησιμοποιήσουμε στα πειράματά μας.



Σχήμα 2.2 : Τα διαφορετικά επίπεδα ανάλυσης του ήχου.

Μια διαφορετική προσέγγιση για το πρόβλημα της ερμηνείας είναι μέσω της χρήσης της βιβλιοθήκης “LIME” (Ribeiro et al.). Το “LIME” αποδίδει την πρόβλεψη ενός μοντέλου σε ένα μέρος του αρχικού δεδομένου. Η διαδικασία αυτή έχει τα εξής βήματα: την πολυδιάσπαση του δεδομένου σε κομμάτια, την απόδοση κάθε κομματιού της πιθανότητας να πετύχει την πραγματική πρόβλεψη και την είσοδο των κομματιών σε ένα γραμμικό επίπεδο για την αποτύπωση στα βάρη, της σημασίας τους για την τελική πρόβλεψη. Παρόμοιες προσπάθειες έχουν γίνει και στον χώρο της μουσικής. Ένα παράδειγμα είναι μέσω του πολυκατακερματισμού του spectogram και του διαχωρισμού των θετικών και αρνητικών βαρών για τη πρόβλεψη (Haunscmid et al.). Στην ίδια λογική έχει αναπτυχθεί και το AudioLIME μόνο που στην περίπτωση αυτή ο πολυκατακερματισμός δεν γίνεται πάνω σε spectogram αλλά πάνω σε Audio, δίνοντας στη διαδικασία περισσότερη μουσικότητα (Haunschmid et al.).



Σχήμα 2.3 : Περιγραφή της λειτουργίας του AudioLime.

2.2 Βαθιά Μάθηση

2.2.1 Τεχνητή Νοημοσύνη

Η Τεχνητή Νοημοσύνη είναι ο τομέας της Επιστήμης των Υπολογιστών που ασχολείται με τη σχεδίαση και την υλοποίηση προγραμμάτων, τα οποία είναι ικανά να μιμηθούν τις ανθρώπινες γνωστικές ικανότητες, εμφανίζοντας έτσι χαρακτηριστικά που αποδίδουμε συνήθως σε ανθρώπινη συμπεριφορά, όπως για παράδειγμα η επίλυση προβλημάτων, η αντίληψη μέσω της όρασης, η μάθηση, η εξαγωγή συμπερασμάτων, η κατανόηση φυσικής γλώσσας (Stuart Russel et al.). Η επιστήμη αυτή συνδυάζει και χρησιμοποιεί πλήθος διαφορετικών επιστημών όπως την φιλοσοφία, τα μαθηματικά, την ψυχολογία, την νευρολογία και την επιστήμη των οικονομικών. Παρά τις “σκοτεινές” περιόδους που έχει περάσει, τα τελευταία χρόνια μπορεί να χαρακτηριστεί ως αναδυόμενη επιστήμη. Σημαντικό κομμάτι της είναι η Μηχανική Μάθηση.

2.2.2 Μηχανική Μάθηση

Η Μηχανική Μάθηση δίνει το ερώτημα για το πως να φτιάξεις υπολογιστές που βελτιώνονται αυτόματα μέσα από την εμπειρία. Είναι ένα από τα πιο ενεργά επιστημονικά πεδία, που συνδυάζει τη στατιστική, την επιστήμη υπολογιστών, την τεχνητή νοημοσύνη και την επιστήμη των δεδομένων (M.I.Jordan). Το 1959, ο Άρθουρ Σάμουελ όρισε τη μηχανική μάθηση ως "Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί". Η μηχανική μάθηση επηρεάζει μεγάλη γκάμα πεδίων της καθημερινής ζωής και των επιστημών. Οι κύριοι τρόποι μάθησης είναι τρεις:

- **Επιβλεπόμενη Μάθηση (Supervised Learning)**, σύμφωνα με την οποία ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους (σύνολο εκπαίδευσης) σε γνωστές επιθυμητές εξόδους, με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο. Χρησιμοποιείται σε προβλήματα:
  - Ταξινόμησης
  - Πρόγνωσης
  - Διερμηνείας

- **Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)**, σύμφωνα με την οποία ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπό μορφή παρατηρήσεων, χωρίς να γνωρίζει τις επιθυμητές εξόδους. Χρησιμοποιείται σε προβλήματα:
  - Ανάλυσης Συσχετισμών
  - Ομαδοποίησης
- **Ενισχυτική Μάθηση (Reinforcement Learning)**, σύμφωνα με την οποία ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον. Χρησιμοποιείται κυρίως σε προβλήματα Σχεδιασμού, όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους.

2.2.3 Βαθιά Μάθηση

Η Βαθιά Μάθηση είναι ένας νέος χώρος της έρευνας στη Μηχανική Μάθηση. Έχει εισαχθεί με στόχο της να μετακινήσει τη Μηχανική Μάθηση πιο κοντά σε έναν από τους αρχικούς στόχους – την Τεχνητή Νοημοσύνη. Η Βαθιά Μάθηση αναφέρεται στην εκμάθηση των πολλαπλών επιπέδων της αναπαράστασης και αφαίρεσης και βοηθά να γίνουν κατανοητά δεδομένα, όπως εικόνες, ήχος και κείμενο. Απαντάει στο πρόβλημα της αναπαράστασης των δεδομένων μέσα από την περιγραφή τους με απλούστερες αναπαραστάσεις και βοηθάει τους υπολογιστές να βγάζουν πιο σύνθετα συμπεράσματα μέσα από απλές έννοιες (Goodfellow et al.).

2.2.4 Τεχνητά Νευρωνικά Δίκτυα

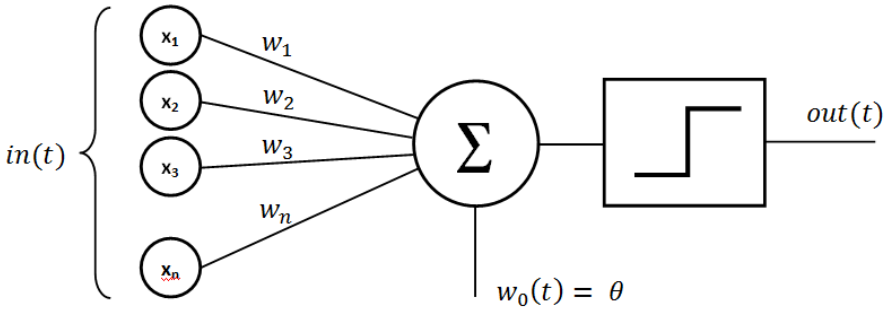
Ένα Νευρωνικό Δίκτυο είναι ένας τεράστιος παράλληλος επεξεργαστής με κατανεμημένη αρχιτεκτονική, ο οποίος αποτελείται από απλές μονάδες επεξεργασίας και έχει από τη φύση του τη δυνατότητα να αποθηκεύει εμπειρική γνώση και να την καθιστά διαθέσιμη για χρήση. Μοιάζει με τον ανθρώπινο εγκέφαλο σε δύο σημεία:

1. Το δίκτυο προσλαμβάνει τη γνώση από το περιβάλλον του, μέσω μιας διαδικασίας μάθησης.
2. Η ισχύς των συνδέσεων μεταξύ των νευρώνων, που αποκαλείται συναπτικό βάρος, χρησιμοποιείται για την αποθήκευση της γνώσης που αποκτιέται.

(Haykin). Η επένδυση στην εξέλιξη των νευρωνικών δικτύων ήταν ένας από τους λόγους που βοήθησαν την Τεχνητή Νοημοσύνη να βγει από τον “χειμώνα” που είχε μπει μια περίοδο. Δόθηκε περισσότερη

έμφαση σε αυτά λόγω του διαδικτύου, μέσω του οποίου έγινε πιο εύκολη η πρόσβαση σε μεγάλο όγκο δεδομένων, λόγω νέων αλγορίθμων βελτιστοποίησης και λόγω της αξιοποίησης των μονάδων επεξεργαστών GPU (Graphics Processing Unit), καθιστώντας δυνατή τη μαζική επεξεργασία δεδομένων. Οι μονάδες αυτές αρχικά σχεδιάστηκαν για άλλη λειτουργία και συγκεκριμένα για την απεικόνιση γραφικών. Ωστόσο η αρχιτεκτονική τους έχει πολλά χαρακτηριστικά που τις κάνει καταλληλότερες για την εκπαίδευση βαθιών νευρωνικών δικτύων απ' ό,τι η CPU, με κυριότερο την ύπαρξη ενός πολύ μεγάλου αριθμού πυρήνων που μπορούν να επιτελέσουν παράλληλο υπολογισμό που συνίσταται κυρίως σε πολλαπλασιασμό πολυεπίπεδων πινάκων. Τα Τεχνητά Νευρωνικά δίκτυα είναι σημαντικότατα λόγω της αποτελεσματικότητας που έχουν επιφέρει στην επεξεργασία πληροφορίας. Βρίσκουν εφαρμογές στην επεξεργασία εικόνες, ανίχνευση ομιλίας, στον έλεγχο και σε πολλά άλλα πεδία. Ο πυρήνας της δομής τους είναι ο νευρώνας (perceptron).

Το Perceptron του Rosenbalt βασίζεται σε ένα μη γραμμικό νευρώνα, συγκεκριμένα στο μοντέλο ενός νευρώνα των McCulloch-Pitts. Το μοντέλο αποτελείται από ένα γραμμικό συνδυαστή ο οποίος ακολουθείται από μια συνάρτηση ενεργοποίησης, ο οποίος εκτελεί τη συνάρτηση προσήμου. Ο κόμβος άθροισης του νευρωνικού μοντέλου υπολογίζει ένα γραμμικό συνδυασμό των εισόδων του εφαρμόζονται στις συνάψεις του, και ενσωματώνει επίσης, μια εξωτερικά εφαρμοζόμενη “προδιάθεση” ή “πόλωση”. Το προκύπτον άθροισμα, δηλαδή, το παραγόμενο τοπικό πεδίο, εφαρμόζεται σε μια συνάρτηση ενεργοποίησης. Ως απόκριση, ο νευρώνας παράγει έξοδο ίση με +1 εάν η είσοδος του απότομου περιοριστή είναι θετική και -1 εάν είναι αρνητική. (Haykin)



Σχήμα 2.4 : Το Perceptron του Rosenbalt.

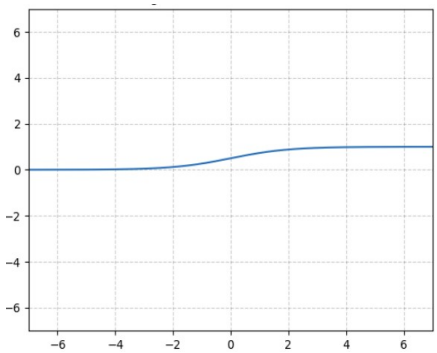
Η αναγκαιότητα να απαντηθούν πιο σύνθετα προβλήματα με μη γραμμικά δεδομένα, οδήγησε τη δημιουργία perceptron πολλών επιπέδων. Τα Perceptron πολλών επιπέδων χαρακτηρίζεται από τα εξής στοιχεία:

- Το μοντέλο κάθε νευρώνα στο δίκτυο περιλαμβάνει μια μη γραμμική συνάρτηση ενεργοποίησης , η οποία είναι διαφορίσιμη.
- Το δίκτυο περιέχει ένα ή περισσότερα επίπεδα τα οποία παραμένουν κρυφά για τους κόμβους των επιπέδων εισόδου και εξόδου.
- Το δίκτυο επιδεικνύει μεγάλη διασυνδεσιμότητα, ο βαθμός της οποίους καθορίζεται από τα συναπτικά βάρη.

Μεγάλη σημασία για την αντιμετώπιση των μη γραμμικών προβλημάτων είναι και η επιλογή κατάλληλων συναρτήσεων ενεργοποίησης. Η συνάρτηση ενεργοποίησης καθορίζει σε τελική ανάλυση την έξοδο των perceptrons. Παρακάτω δίνονται τρεις χαρακτηριστικές συναρτήσεις ενεργοποίησης:

1. **Sigmoid** $$Sigmoid(x) = \frac{1}{1 + \exp(-x)}$$

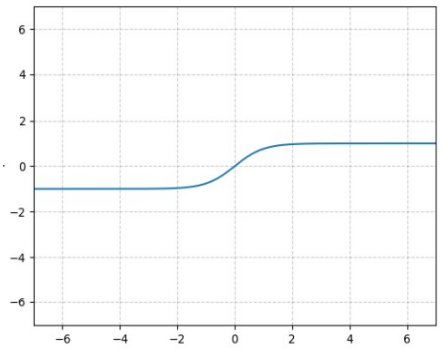
Οι τιμές που παίρνει είναι από [0,1]. Χρησιμοποιείται κυρίως σε προβλήματα που θέλουμε να δοθεί κάποια πιθανότητα. Το όνομα της προκύπτει γιατί η γραφική της παράσταση μοιάζει με S. Στο πρόβλημα της ταξινόμησης τη συναντάμε συχνά.



Σχήμα 2.5 : Η γραφική παράσταση της Sigmoid.

2. **Tanh** $$Tanh(x) = \frac{\exp(x) + \exp(-x)}{\exp(x) + \exp(-x)}$$

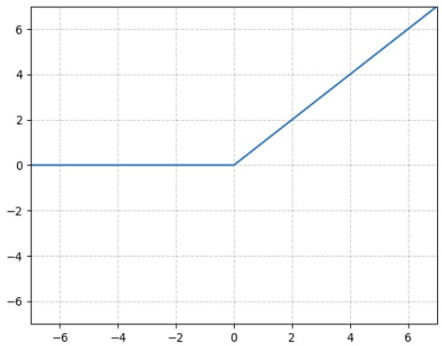
Το σύνολο τιμών της Tanh είναι το [-1,1] και χρησιμοποιείται παρόμοια με τη Sigmoid. Το θετικό με αυτή τη συνάρτηση είναι ότι έχει και αρνητικές τιμές.



Σχήμα 2.6 : Η γραφική παράσταση της Tanh.

3. **ReLU** $$ReLU(x) = max(0, x)$$

Η πιο συνηθισμένη συνάρτηση ενεργοποίησης. Χρησιμοποιείται συχνά σε προβλήματα βαθιάς μάθησης και είναι προτιμητέα για την αποτελεσματικότητα και τον εύκολο υπολογισμό της. Παίρνει τιμές στο [0,+∞).



Σχήμα 2.7 : Η γραφική παράσταση της ReLU.

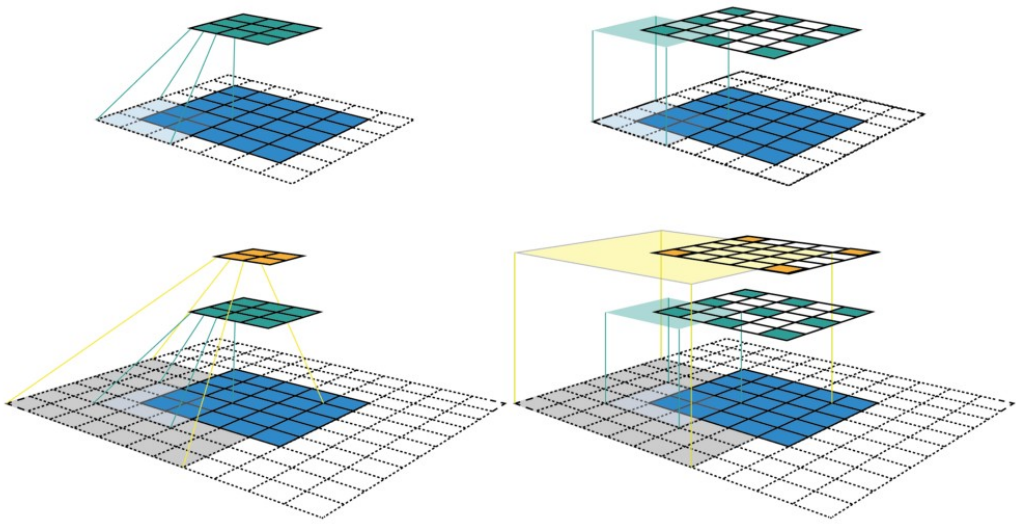
2.2.5 Συνελικτικά Νευρωνικά Δίκτυα

Τα συνελικτικά νευρωνικά δίκτυα (CNN) αποτέλεσαν τομή και ενθάρρυναν την επιστημονική κοινότητα για την αποτελεσματικότητα των Νευρωνικών Δικτύων. Οι βασικοί λόγοι που το κατάφεραν είναι ότι είναι υπολογιστικά πιο ελαφρά και ότι είναι ικανά να μελετήσουν τοπικά χαρακτηριστικά.

Ένα συνελικτικό νευρωνικό δίκτυο είναι ένα perceptron πολλών επιπέδων ειδικά σχεδιασμένο ώστε να αναγνωρίζει δισδιάστατα σχήματα με υψηλό βαθμό μη-ευαισθησίας στη μετατόπιση, την κλιμάκωση, την στρέβλωση και άλλες μορφές παραμόρφωσης. Αυτή η δύσκολη εργασία διδάσκεται με επιβλεπόμενο τρόπο, μέσω ενός δικτύου του οποίου η δομή περιλαμβάνει τις ακόλουθες μορφές περιορισμών. (LeCun και Bengio):

1. **Εξαγωγή χαρακτηριστικών.** Κάθε νευρώνας λαμβάνει τις συναπτικές εισόδους του από ένα τοπικό δεκτικό πεδίο του προηγούμενου επιπέδου υποχρεώνοντας το να εξάγει τοπικά χαρακτηριστικά. Αφού εξαχθεί ένα χαρακτηριστικό, η ακριβής θέση του γίνεται λιγότερο σημαντική, εφόσον διατηρείται η σχετική του θέση ως προς άλλα χαρακτηριστικά.
2. **Αντιστοίχιση χαρακτηριστικών.** Κάθε υπολογιστικό επίπεδο του δικτύου απαρτίζεται από πολλαπλούς χάρτες χαρακτηριστικών, με κάθε χάρτη χαρακτηριστικών να είναι στη μορφή ενός επιπέδου μέσα στο οποίο οι μεμονωμένοι νευρώνες ελέγχονται ώστε να μοιράζονται το ίδιο σύνολο συναπτικών βαρών. Αυτή η δεύτερη μορφή δομικού περιορισμού έχει τα ακόλουθα ευεργετικά επακόλουθα:
    - Μη-ευαισθησία ως προς τη μετατόπιση, η οποία επιβάλλεται στη λειτουργία ενός χάρτη χαρακτηριστικών μέσω της χρήσης μια συνέλιξης με έναν πυρήνα μικρού μεγέθους, η οποία ακολουθείται από την εφαρμογή μιας σιγμοειδούς συνάρτησης.
    - Μείωση του αριθμού των ελεύθερων παραμέτρων, η οποία επιτυγχάνεται μέσω του διαμοιρασμού των βαρών.
3. **Υποδειγματοληψία.** Κάθε συνελικτικό επίπεδο ακολουθείται από ένα υπολογιστικό επίπεδο το οποίο εκτελεί τοπικό υπολογισμό μέσων όρων και υποδειγματοληψία, διά των οποίων η ανάλυση του χάρτη χαρακτηριστικών μειώνεται. Αυτή η λειτουργία έχει ως αποτέλεσμα τη μείωσης ευαισθησίας της εξόδου του χάρτη χαρακτηριστικών στις μετατοπίσεις και άλλες μορφές παραμόρφωσης.

Μια έννοια στην οποία πρέπει να δοθεί έμφαση για την κατανόηση της λειτουργία του CNN μοντέλου είναι το δεκτικό πεδίο που αναφέρθηκε και παραπάνω. Το δεκτικό πεδίο ορίζεται ως η περιοχή στην αρχική είσοδο όπου επεξεργάζεται ένας πυρήνας του CNN. Μπορεί να περιγραφεί από το κέντρο του και από το μέγεθός του. Όσο πιο κοντά είναι ένα στοιχείο στο κέντρο του δεκτικό πεδίο τόσο πιο σημαντικό είναι.



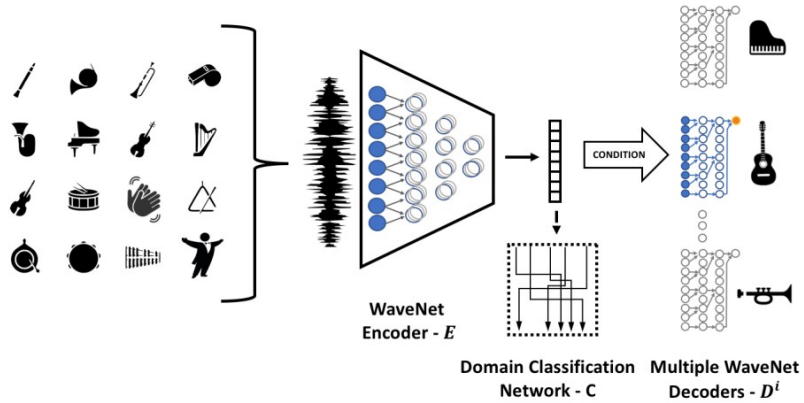
Σχήμα 2.8 : *Δύο τρόποι εφαρμογής CNN δικτύου αναδεικνύοντας τα δεκτικά τους πεδία.*

2.2.6 Αυτοκωδικοποιητές

Ένας Αυτοκωδικοποιητής (Autoencoder) είναι ένα νευρωνικό δίκτυο που μαθαίνει να αντιγράφει την είσοδο του στην έξοδο. Χρησιμοποιείται στην μη-επιβλεπόμενη μάθηση και αποτελείται από δύο μέρη: την κωδικοποίηση της εισόδου και την αποκωδικοποίησή της. Ο στόχος της διαδικασίας δεν είναι απλά να μάθει να αντιγράφει την είσοδο αλλά να μάθει να κωδικοποιεί την πιο χρήσιμη και συμπυκνωμένη πληροφορία της εισόδου. Αυτό επιτυγχάνεται μέσα από διάφορους περιορισμούς που μπαίνουν στην εκπαίδευση του.

Παραδοσιακά χρησιμοποιούνται για συμπύκνωση δεδομένων και για χαρτογράφηση των χαρακτηριστικών της εισόδου. Τελευταία όμως χρησιμοποιούνται και για άλλους λόγους, όπως είναι η παραγωγή νέων δεδομένων. Υπάρχουν παραδείγματα εφαρμογών που σχετίζονται και με το Music Style transfer. Ένα παράδειγμα βρίσκεται στη δημοσίευση “A-Universal-Music-Translation-Network” (Noam Mor et al.). Το μοντέλο που έχουν φτιάξει μπορεί να κωδικοποιήσει ένα τραγούδι σε εκτέλεση κιθάρας και να το αποκωδικοποιήσει σε εκτέλεση πιάνου. Το κλειδί για το αποτέλεσμα αυτό είναι η τεχνική disentanglement representation. Η

τεχνική αυτή έχει να κάνει με το πως μπορείς να κρατήσεις τα πιο χρήσιμα στοιχεία, για το σκοπό σου, ενός δεδομένου σου. Στο παράδειγμα που δώσαμε παραπάνω, το μοντέλο μαθαίνει να κρατάει την αναπαράσταση του τραγουδιού, κρατώντας μόνο τη τονικότητα του και μαθαίνει από την τονικότητα αυτή να το αποκωδικοποιεί σε άλλο μουσικό στυλ.



Σχήμα 2.9 : Η αρχιτεκτονική του δικτύου “A Universal Music Translation Network”.

Τα είδη των Αυτοκωδικοποιητών είναι αρκετά, ανάλογα τις αρχιτεκτονικής τους και των ποινών που εφαρμόζουμε. Κάποια παραδείγματα ενδεικτικά παρουσιάζονται παρακάτω:

- **Undercomplete Autoencoder.**

  Στους Υποπλήρης Αυτοκωδικοποιητές (Undercomplete Autoencoders) επιβάλλεται περιορισμός στη διάσταση της κωδικοποίησης, έτσι ώστε να είναι μικρότερη της διάστασης της εισόδου, κρατώντας τα πιο σημαντικά στοιχεία. Συνήθως οι υλοποιήσεις του δεν προσφέρουν πολύ στην εκμάθηση χαρακτηριστικών αλλά αποτελούν απλή αντιγραφή της εισόδου στην έξοδο.

- **Deep Autoencoder.**

  Στους Βαθείς Αυτοκωδικοποιητές (Deep Autoencoders) οι συναρτήσεις κωδικοποίησης δεν αποτελούνται από ένα μόνο επίπεδο αλλά από περισσότερα. Χρησιμοποιούνται κυρίως στην εξαγωγή χαρακτηριστικών του δεδομένου. Αποτελούνται από πολλά επίπεδα κωδικοποίησης μέχρι τον λανθάνων χώρο (latent space) και από πολλά επίπεδα αποκωδικοποίησης. Η απώλεια του

είναι το σφάλμα κατασκευής που προκύπτει από τη σύγκριση της εξόδου με την είσοδο.

- **Variational Autoencoder.**

  Οι VAE διαφέρουν από τις άλλες υλοποιήσεις που έχουν εξεταστεί μέχρι τώρα στο ότι χρησιμοποιούν την τυχαιότητα. Τα αποτελέσματα του κωδικοποιητή δειγματοληπτούνται από μία κανονική κατανομή. Έτσι μετατρέπονται σε ένα σύνολο από σημεία που μπορούν να διαχωριστούν και να δειγματοληφθούν εύκολα. Η συνάρτηση απώλειας του αποτελείται από δύο μέρη: Το πρώτο μέρος είναι η απώλεια ανακατασκευής ενώ το δεύτερο είναι η λανθάνουσα απώλεια, έτσι ώστε η πιθανότητα που διαλέγονται τα δείγματα να προέρχεται από κανονική κατανομή. Χρησιμοποιείται η KL απόκλιση μεταξύ της κανονικής και της υπάρχουσας κατανομής. Η κύρια λειτουργία των VAE είναι στη παραγωγή νέων δεδομένων παρά στην εκμάθηση χαρακτηριστικών. Είναι ξεχωριστοί γιατί κατάφεραν να έχουν καλύτερη γενίκευση σε σχέση με τις πρώτες υλοποιήσεις των αυτοκωδικοποιητών.

Κεφάλαιο 3

Σχεδιασμός και Υλοποίηση Πειραμάτων

Στο Κεφάλαιο αυτό περιγράφεται ο σχεδιασμός και η υλοποίηση των πειραμάτων. Γίνεται αναφορά στο περιβάλλον υλοποίησης και στα εργαλεία που χρειαστήκαμε. Περιγράφονται τα σύνολα δεδομένων και οι μετρικές που είχαμε στην εκπαίδευση. Αναλύονται οι αρχιτεκτονικές των δικτύων που δοκιμάσαμε και τα αποτελέσματα των επιδόσεων τους. Τέλος αναλύονται οι τεχνικές ομαλοποίησης που χρησιμοποιήσαμε.

3.1 Περιβάλλον Υλοποίησης

Τα πειράματα τα υλοποιήσαμε στη γλώσσα προγραμματισμού Python. Για βιβλιοθήκη βαθιάς μάθησης προτιμήσαμε τη Pytorch, καθώς είναι εύκολα κατανοητά και σαφή τα στάδια εκπαίδευσης που διαθέτει. Χρησιμοποιήσαμε πλήθος βιβλιοθηκών για τα διαγράμματα, για την επεξεργασία δεδομένων και για διάφορες αναγκαιότητες που προέκυψαν.

Το περιβάλλον υλοποίησης μας αρχικά ήταν το Google Colab. Το Colab είναι ένα φιλικό προς το χρήστη περιβάλλον που προσφέρει η Google. Με αυτό μπορείς να οργανώσεις με ωραίο τρόπο τα πειράματα, μπορείς να χρησιμοποιήσεις το Google Drive για απευθείας σύνδεση των δεδομένων και του πειράματος και μπορείς να κάνεις ταχύτερη την εκπαίδευση των μοντέλων χρησιμοποιώντας επεξεργαστές GPU. Λόγω μείωσης του επιτρεπτού χρόνου δέσμευσης GPU της Google και λόγω της απαιτητικότητας σε μνήμη, ορισμένων πειραμάτων μας, οδηγηθήκαμε στο να χρησιμοποιήσουμε GPU που προσφέρονται από τον server pinkfloyd του AILS εργαστηρίου του Εθνικού Μετσόβιου Πολυτεχνείου. Τους GPU τους χρησιμοποιήσαμε μέσω σύνδεσης ssh.

3.2 Δεδομένα

Παρακάτω αναλύονται τα σύνολα δεδομένων που χρησιμοποιήσαμε καθώς και δύο τρόποι αναπαράστασης τους. Επιπλέον περιγράφονται ορισμένες προ-επεξεργασίες που κάναμε, οι οποίες ήταν απαραίτητες για την εκπαίδευση.

3.2.1 Σύνολα Δεδομένων

Τα σύνολα δεδομένων που χρησιμοποιήσαμε ήταν για δύο σκοπούς. Ο πρώτος σχετίζεται με τη σύνδεση των ενδιάμεσων χαρακτηριστικών και του συναισθήματος (το σύνολο δεδομένων των Aljanaki et al). Ο δεύτερος σχετίζεται με τα πειράματα που κάναμε για την προ-εκπαίδευση (το σύνολο δεδομένων που συγκροτήσαμε μέσω του Spotify API).

3.2.1.1 Aljanaki's Mid-level Perceptual Features

Το σύνολο δεδομένων της Aljanaki et al. είναι το σύνολο δεδομένων που χρησιμοποιήσαμε στα περισσότερα πειράματα μας. Αποτελείται από 5000 τραγούδια τα οποία έχουν χαρακτηριστεί με 7 ενδιάμεσα χαρακτηριστικά (Melodiousness, Articulation, R. Stability, R. Complexity, Dissonance, Tonal Stability, Minorness) από μουσικούς. Η επιλογή των μουσικών έγινε μέσα από τον έλεγχο τους στη κατανόηση βασικών μουσικών εννοιών (αρμονία, χρωματισμός, τονικότητα). Επιλέχθηκαν 155 από τους 2236 οι οποίοι βαθμολόγησαν τα ενδιάμεσα χαρακτηριστικά των τραγουδιών με κλίμακα 0.1 έως 1 ανάλογα με την ένταση της παρουσίας των χαρακτηριστικών στα τραγούδια. Ως σημείο αναφοράς είχαν 100 τραγούδια από διάφορα είδη που εμφάνιζαν ακραίες περιπτώσεις των χαρακτηριστικών. (έντονη παρουσία του leggato , συνεχής αλλαγή τονικότητας κτλ.)

Τα τραγούδια του συνόλου δεδομένων επιλέχθηκαν από άλλα σύνολα δεδομένων (Magnatune 829 , Jamendo 2746, Soundtracks 360, mirex-bimodal 162, mirex-multimodal 903). Η Soundtracks και η mirex-multimodal έχουν και ταμπέλες συναισθήματος. Συγκεκριμένα η Soundtracks έχει 360 τραγούδια από ταινίες με 8 ταμπέλες συναισθήματος (valence, energy, tension, anger, feat, happy, sad, tender) βαθμολογημένες σε κλίμακα 0.1 έως 1 (Eerola et al.). Το σύνολο δεδομένων mirex-multimodal, που αναφέρεται και παραπάνω, έχει 903 τραγούδια ταξινομημένα στις 5 Mirex ομάδες. Αυτά τα σύνολα δεδομένων θα τα χρησιμοποιήσουμε αργότερα για να μελετήσουμε τη σύνδεση των ενδιάμεσων χαρακτηριστικών με το συναίσθημα.

Perceptual Feature	Question asked to human raters
Melodiousness	To which excerpt do you feel like singing along?
Articulation	Which has more sounds with staccato articulation?
Rhythmic Stability	Imagine marching along with the music. Which is easier to march along with?
Rhythmic Complexity	Is it difficult to repeat by tapping? Is it difficult to find the meter? Does the rhythm have many layers?
Dissonance	Which excerpt has noisier timbre? Has more dissonant intervals (tritones, seconds, etc.)?
Tonal Stability	Where is it easier to determine the tonic and key? In which excerpt are there more modulations?
Modality ('Minorness')	Imagine accompanying this song with chords. Which song would have more minor chords?

Πίνακας 3.1: Τα ενδιάμεσα χαρακτηριστικά και οι ερωτήσεις που έγιναν στους βαθμολογητές προκειμένου να καταλάβουνε τις έννοιες.

3.2.1.2 Σύνολα Δεδομένων για τη Προ-Εκπαίδευση

Το Spotify παρέχει σε προγραμματιστές τη δυνατότητα άντλησης πληροφοριών για τα τραγούδια που κάνει stream. Συγκεκριμένα μέσω του Spotify API και με τη βοήθεια βιβλιοθηκών όπως το Spotipy μπορεί κάποιος να κάνει αιτήσεις (requests) σχετικά με πληροφορίες τραγουδιών. Οι πληροφορίες που προσφέρει είναι από πολύ αφαιρετικές (ένταση τραγουδιού) μέχρι χαμηλού επιπέδου (τόνος τραγουδιού). Περισσότερες λεπτομέρειες για τα χαρακτηριστικά που παρέχει θα δοθούν στην υποενότητα 3.2.2.

Ο λόγος που χρησιμοποιήσαμε το Spotify API είναι γιατί μπορούμε να συγκροτήσουμε εύκολα μεγάλα σύνολα δεδομένων, τα οποία τα χρειαζόμαστε για την προ-εκπαίδευση των μοντέλων βαθιάς μάθησης. Επειδή η προ-εκπαίδευση έχει σαν στόχο να μάθει να αναπαριστά βασικά χαρακτηριστικά των τραγουδιών ανεξάρτητα από το σε ποιο είδος ανήκουν, σαν κριτήριο για την επιλογή των τραγουδιών είχαμε την ποικιλία τους σε είδη.

Επιλέξαμε για αυτόν το λόγο να χρησιμοποιήσουμε το "Every Noise at Once". Το "Every Noise at Once" είναι μια αλγοριθμική προσπάθεια καταγραφής του χάρτη όλων των ειδών των μουσικών τραγουδιών, βασισμένη σε αναλύσεις και δεδομένα που παράγονται από το Spotify API. Τα είδη που καταγράφει είναι γύρω στα 5000. Γράψαμε κώδικα σε Python που κάνει request κάθε φορά ένα τυχαίο τραγούδι από κάθε είδος, και με αυτό τον τρόπο συγκροτήσαμε δύο σύνολα δεδομένων. Το πρώτο αποτελείται από 5 τραγούδια από κάθε είδος, άρα από 25000 τραγούδια, και το ονομάσαμε Pre-training. Το δεύτερο αποτελείται από 25 τραγούδια από κάθε είδος, άρα από 200.000, και το ονομάσαμε Large-Pre-training.

Σχήμα 3.1: Τα είδη τραγουδιών σε WordCloud.

3.2.2 Αναπαράσταση Δεδομένων

Η επιλογή της αναπαράστασης των τραγουδιών είναι σημαντική καθώς κάθε αναπαράσταση παρέχει διαφορετικά στοιχεία για τα δεδομένα μας. Οι τρόποι αναπαράστασης που χρησιμοποιήσαμε ήταν δύο. Ο πρώτος μέσω των MFCC και ο δεύτερος μέσω των χαρακτηριστικών του Spotify API.

3.2.2.1 Librosa

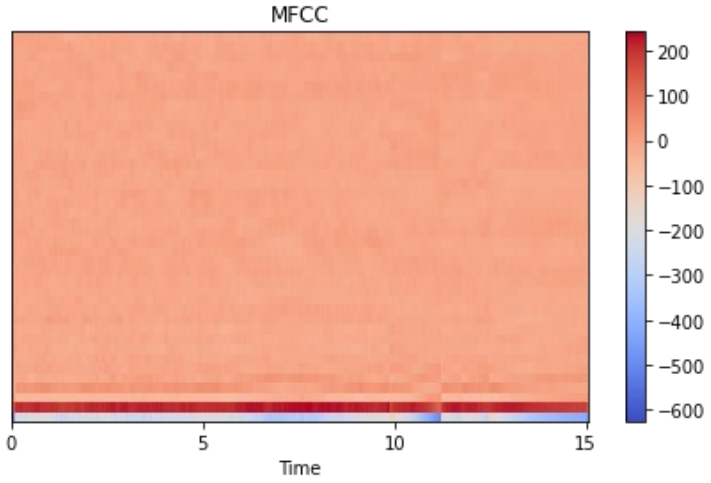
Η Librosa είναι βιβλιοθήκη της Python όπου χρησιμοποιείται σε MIR εφαρμογές. Εμείς την χρησιμοποιήσαμε προκειμένου να μετατρέψουμε τα αρχεία ήχου του συνόλου δεδομένου της Aljanaki σε MFCCs (Mel Frequency Cepstral Coefficients).

Τα MFCCs είναι χαρακτηριστικά που παρέχουν φασματική πληροφορία σημάτων ήχου. Χρησιμοποιούνται ευρέως σε MIR εφαρμογές. Αυτό που καθιστά τα MFCCs τόσο αποτελεσματικά είναι ότι οι ζώνες συχνοτήτων τοποθετούνται λογαριθμικά πάνω στη κλίμακα Mel, η οποία προσεγγίζει την απόκριση της ανθρώπινης ακοής πολύ καλύτερα από τη γραμμική κλίμακα συχνοτήτων.

Σαν πρώτο βήμα για την εξαγωγή των συντελεστών MFCC υπολογίζεται ο μετασχηματισμός Fourier του σήματος. Στη συνέχεια, η απόλυτη τιμή του φάσματος που προκύπτει φιλτράρεται σε ένα τριγωνικό filterbank που αντιστοιχεί στην κλίμακα Mel και το αποτέλεσμα μετασχηματίζεται με DCT (Discrete Cosine Transform), με αυτό τον τρόπο

συμπυκνώνεται το μεγαλύτερο ποσοστό της ενέργειας σε λίγους συντελεστές.

Η εξαγωγή των MFFCs έγινε με τη βοήθεια της συνάρτησης *librosa.feature.mfcc().* Μετατρέψαμε τα 5000 αρχεία ήχου του συνόλου δεδομένων της Aljanki σε Mel-frequency cepstrum με sample rate ίσο με 22050 και με αριθμό των MFFCs 40.



Σχήμα 3.2: Παράδειγμα αρχείου MFCC.

3.2.2.2 Spotify

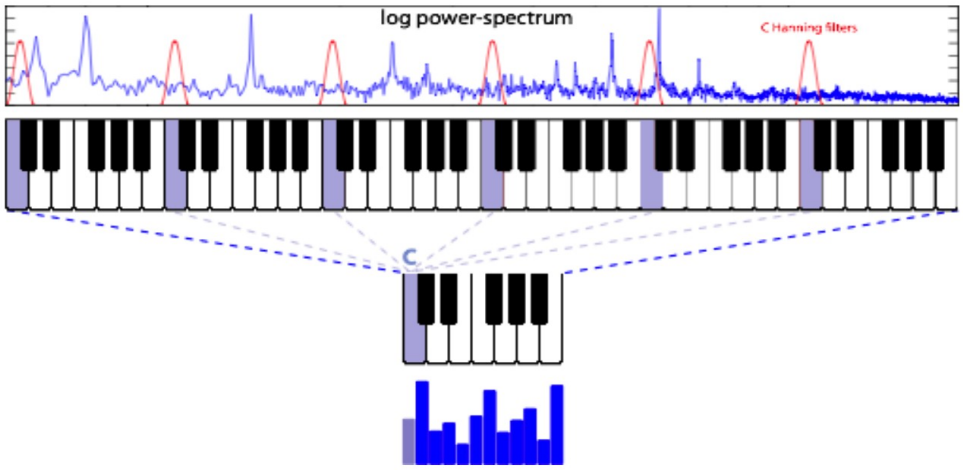
Τα βασικά χαρακτηριστικά των τραγουδιών που προσφέρει το Spotify API, που χρησιμοποιήσαμε, είναι τα segments και τα audio features. Ένα τραγούδι χωρίζεται χρονικά σε segments που το κάθε segment έχει κάποια χαρακτηριστικά. Τα κύρια είναι το Loudness, το Timbre και το Pitch.

ΚΛΕΙΔΙ	ΤΥΠΟΣ	ΠΕΡΙΓΡΑΦΗ
start	Float	Σημείο έναρξης του segment.
duration	Float	Διάρκεια του segment.
confidence	Float	Εγκυρότητα του segmentation.
loudness_start	Float	Εναρξη του υπολογισμού της έντασης του segment.
loudness_max	Float	Max της έντασης του segment.
loudness_max_time	Float	Διάρκεια του max της έντασης του segment.
loudness_end	Float	Λήξη του υπολογισμού της έντασης του segment.
pitches	Array of floats	Ένα “chroma” διάνυσμα που αναπαριστά το τονικό περιεχόμενο του segment, ανταποκρίνεται σε 12 τονικές κλάσεις όπου η κάθε μια παίρνει τιμή από 0.0 σε 1.0 ανάλογα με το πόσο κυριαρχεί στο τόνο.
timbre	Array of floats	Το Timbre είναι η ποιότητα της μουσικής νότας ή του ήχου όπου διαχωρίζει τους διαφορετικούς τύπους μουσικών οργάνων ή φωνών.

Πίνακας 3.2: Περιγραφή του segment αντικειμένου.

Pitch

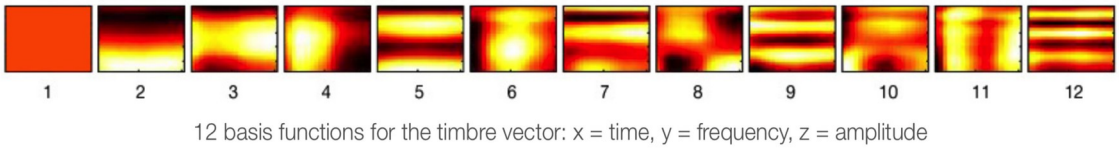
Το τονικό περιεχόμενο δίνεται από ένα “chroma” διάνυσμα, που αποτελείται από 12 τονικές κλάσεις, με τιμές από 0.0 σε 1.0 που αντιπροσωπεύουν το μέγεθος της παρουσίας του τόνου (νότας) στη χρωματική κλίμακα. Για παράδειγμα η C μείζονα συγχορδία θα έχει μεγάλες τιμές για τους τόνους C, E και G (Jehan et al.).



Σχήμα 3.3: Απεικόνιση της pitch αναπαράστασης.

Timbre

Η “χροιά” είναι η ποιότητα μιας μουσικής νότας ή ήχου που διακρίνει διαφορετικού τύπου μουσικά όργανα ή φωνές. Είναι μια περίπλοκη έννοια που αναφέρεται ως ηχόχρωμα, υφή ή τονική ποιότητα και προέρχεται από ένα segment. Το στοιχείο της “χροιάς” είναι ένα διάνυσμα όπου περιέχει 12 τιμές περίπου κεντραρισμένες στο 0. Η πρώτη διάσταση αναπαριστά τη μέση ένταση του segment, η δεύτερη δίνει έμφαση στη φωτεινότητα, η τρίτη στο πόσο μονότονος είναι ο ήχος κτλ. Η πραγματική “χροιά” του segment είναι ένας γραμμικός συνδυασμός των 12 διαστάσεων πολλαπλασιασμένων με τα κατάλληλα βάρη: timbre = c1 x b1 + c2 x b2 + ... + c12 x b12 (Jehan et al.).



Σχήμα 3.4: Οι 12 βασικές συναρτήσεις για την αναπαράσταση της “χροιάς”.

Τα audio features είναι κάποια πιο αφαιρετικά χαρακτηριστικά των τραγουδιών. Τα χαρακτηριστικά αυτά τα χρησιμοποιήσαμε εμβόλιμα στα τελευταία συνδετικά επίπεδα των αρχιτεκτονικών των δικτύων μας ή για Supervised Pre-training. Συγκεκριμένα τα χαρακτηριστικά αυτά είναι το κλειδί, η κλίμακα, το μέτρο, η χορευτικότητα, η ακουστικότητα, η ενέργεια, η ορχηστρικότητα, η παρουσία ή απουσία κοινού, η ένταση, η ποσότητα στίχου, η θετικότητα και ο ρυθμός.

ΚΛΕΙΔΙ	ΤΥΠΟΣ	ΠΕΡΙΓΡΑΦΗ ΤΙΜΩΝ
key	int	Το κλειδί του τραγουδιού.
mode	int	Το είδος της κλίμακας (μείζονα / ελάσσονα).
time_signature	int	Τι μέτρο, το πόσα beat είναι σε κάθε bar.
acousticness	float	Η ακουστική, όταν η τιμή είναι 1 σημαίνει ότι το τραγούδι έχει ακουστική.
danceability	float	Πόσο κατάλληλο είναι το τραγούδι για χορό.
energy	float	Η ενέργεια, το πόσο γρήγορα, θορυβώδη και δυνατά σε ένταση είναι τα τραγούδια.
instrumentalness	float	Ύπαρξη η μη φωνητικών. Η Rap για παράδειγμα έχει κυρίως vocals.
liveness	float	Εντοπίζει αν είναι live. Την ύπαρξη ή μη του κοινού.
loudness	float	Ενταση σε db. Η έκταση της τιμής του είναι από -60 μέχρι 0 db.
speechiness	float	Αν έχει λόγια το τραγούδι. Τραγούδια με τιμές κοντά στο 0, δεν έχουν στοίχους
valence	float	Δραστικότητα και σθένος του τραγουδιού. Το πόσο προκαλεί αισιοδοξία και θετική διάθεση.
tempo	float	Ο ρυθμός. Ο μέσος όρος των beat αν λεπτό.

Πίνακας 3.3: Περιγραφή των audio features του Spotify API.

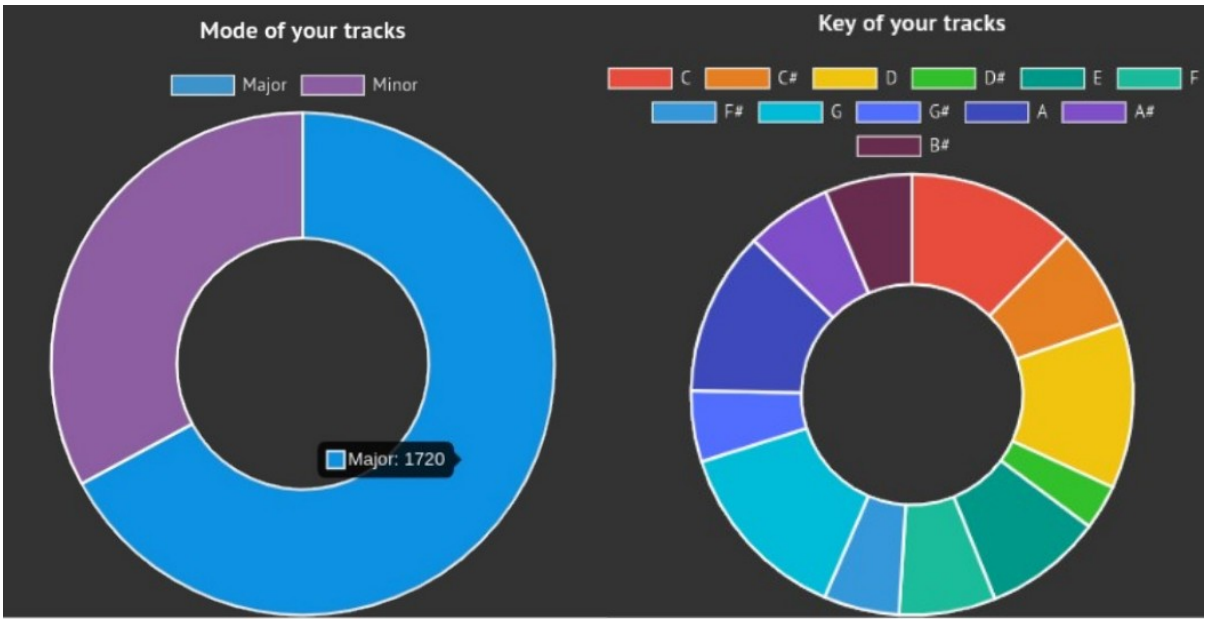
Με το Spotify API δίνονται και άλλα χαρακτηριστικά του τραγουδιού, όπως το bar το beat και το tatums με τα οποία μπορείς να προσδιορίσεις το ρυθμό του τραγουδιού. Επίσης μπορεί να χωριστεί το κομμάτι σε sections και να δοθούν χαρακτηριστικά όπως η κλίμακα, το κλειδί, ο ρυθμός και η ένταση του κάθε section του τραγουδιού. Τα παραπάνω χαρακτηριστικά δεν τα χρησιμοποιήσαμε.

Όπως αναφέρεται στην υποενότητα 3.2.1.2, χρησιμοποιήσαμε το Spotify API για να συγκροτήσουμε μεγάλα σύνολα δεδομένων για προ-εκπαίδευση. Επιπλέον παίρνοντας τα μετα-δεδομένα του συνόλου δεδομένων της Aljanaki προσπαθήσαμε να το συγκροτήσουμε με αναπαράσταση των χαρακτηριστικών που προσφέρει το Spotify API. Η διαδικασία συγκρότησης που ακολουθήσαμε είναι η εξής.

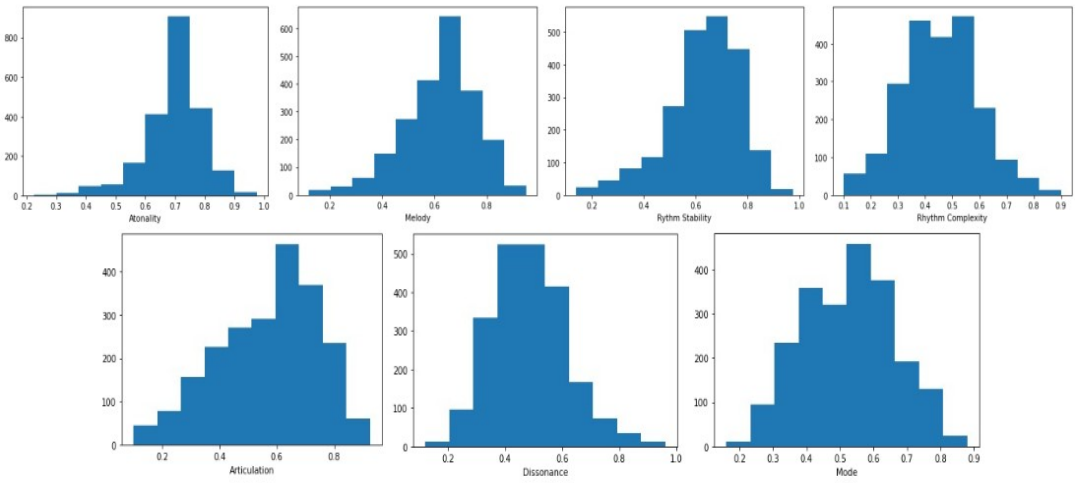
Κάναμε αναζητήσεις στο Spotify API με βάση τα μετα-δεδομένα του συνόλου δεδομένων της Aljanaki με διαφορετικούς συνδυασμούς. Δηλαδή με βάση το όνομα του Καλλιτέχνη και του Τραγουδιού, το όνομα του

Άλμπουμ και του Τραγουδιού και με το ένα ή το άλλο. Αποδείχτηκε ότι η αναζήτηση με βάση το όνομα του Άλμπουμ και του Τραγουδιού προκαλεί “μπερδέματα” στο σύνολο δεδομένων. Συνεπώς καταλήξαμε στο να χρησιμοποιήσουμε το σύνολο δεδομένων που βασίζεται στην αναζήτηση στο Spotify API με βάση το όνομα του Καλλιτέχνη και του Τραγουδιού. Βρήκαμε εν τέλει 2196 τραγούδια από τα 5000 του συνόλου δεδομένων της Aljanaki σε μορφή spotify χαρακτηριστικών.

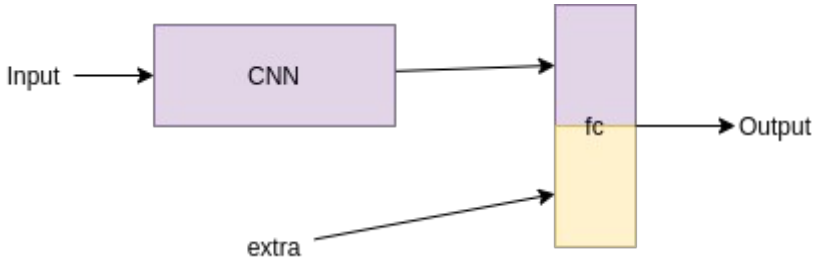
Η αναπαράσταση των τραγουδιών που χρησιμοποιήσαμε ήταν με τα 78 πρώτα segments των τραγουδιών που είχαν τιμές για το Pitch, το Timbre και το Loudness. Επίσης χρησιμοποιήσαμε και τα 12 αφαιρετικά audio features, εμβόλιμα στις αρχιτεκτονικές των δικτύων μας.



Σχήμα 3.5: Η κατανομή της κλίμακας και του κλειδιού των δεδομένων της Aljanaki’s dataset.



Σχήμα 3.6: Η κατανομή των ενδιάμεσων χαρακτηριστικών στην Aljanki’s dataset.



Σχήμα 3.7: Τα audio features που χρησιμοποιούνται εμβόλιμα στην αρχιτεκτονική των δικτύων.

3.2.3 Προ-επεξεργασία Δεδομένων

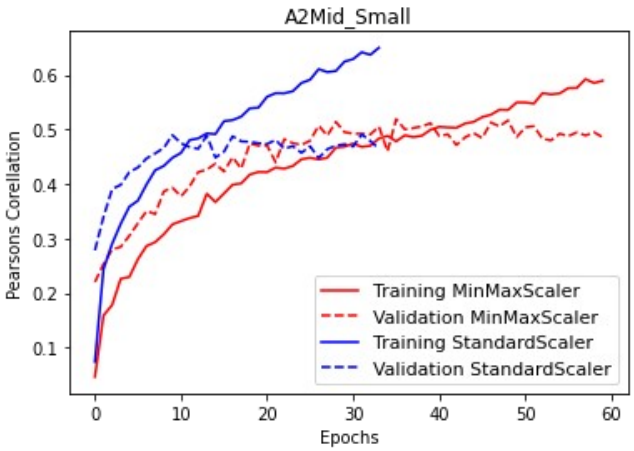
Προκειμένου να ξεκινήσουμε την εκπαίδευση των αρχιτεκτονικών μας δικτύων απαραίτητη είναι η προ-επεξεργασία των δεδομένων μας. Τα σύνολα δεδομένων μας που προέρχονται από το σύνολο δεδομένων της Aljanaki, τα χωρίσαμε σε 3 μέρη. Το πρώτο για το σύνολο εκπαίδευσης (training set) σε ποσοστό 60%, με το οποίο έγινε η εκπαίδευση τον μοντέλων. Το δεύτερο για το σύνολο επαλήθευσης (validation set) σε ποσοστό 20%, με το οποίο έγινε η αξιολόγηση των μοντέλων κατά τη διάρκεια της εκπαίδευσης. Το τρίτο για το σύνολο δοκιμής (testing set) σε ποσοστό 20%, το οποίο χρησιμοποιήθηκε για την αντικειμενική αξιολόγηση της επίδοσης του μοντέλου. Τα σύνολα των δεδομένων για τη προ-εκπαίδευση τα χωρίσαμε σε training set 92% και σε validation set 8%.

Πολύ σημαντικό για την εκπαίδευση των μοντέλων είναι τα δεδομένα που χρησιμοποιούμε να είναι κανονικοποιημένα. Με αυτό τον τρόπο αποφεύγονται φαινόμενα άνισης εκπαίδευσης των βαρών του δικτύου. Δοκιμάσαμε πολλούς τρόπους κανονικοποίησης, καταλήγοντας όμως στο συμπέρασμα ότι όσον αφορά τις μετρικές δεν επηρεάζει το ποιο θα διαλέξουμε. Οι συναρτήσεις που χρησιμοποιήσαμε προέρχονται από τη βιβλιοθήκη της sklearn, η οποία παρέχει πολλά εργαλεία χρήσιμα για τη μηχανική μάθηση.

Name	Sklearn_Class	Pearsons Correlation
StandardScaler	StandardScaler	0.578
MinMaxScaler	MinMaxScaler	0.585
MaxAbsScaler	MaxAbsScaler	0.543
RobustScaler	RobustScaler	0.588
QuantileTransformer-Normal	QuantileTransformer(output_distribution='normal')	0.578
QuantileTransformer-Uniform	QuantileTransformer(output_distribution='uniform')	0.599
PowerTransformer-Yeo-Johnson	PowerTransformer(method='yeo-johnson')	0.555
Normalizer	Normalizer	0.585

Πίνακας 3.4: Το Pearsons Correlation των τιμών πρόβλεψης με των πραγματικών τιμών για διαφορετικές μεθόδους κανονικοποίησης των δεδομένων.

Τα αποτελέσματα όσον αφορά την επίδοση των μοντέλων δεν επηρεάζονται από την επιλογή της μεθόδου κανονικοποίησης, επηρεάζεται όμως η ταχύτητα εκπαίδευσης και σύγκλισης του μοντέλου. Η ταχύτητα σύγκλισης σχετίζεται με το αν οι τιμές των δεδομένων είναι κοντά στο 0 και στο αν έχουν παρόμοια τυπική απόκλιση (Yann LeCun et al.). Αυτό το πετυχαίνει καλύτερα το StandardScaler, το οποίο όπως φαίνεται παρακάτω συγκλίνει πιο γρήγορα και για αυτόν τον λόγο το επιλέξαμε στα πειράματα μας.



Σχήμα 3.8: Σύγκριση των κανονικοποιήσεων MinMax και Standard στην αρχιτεκτονική A2Mid_Small.

3.3 Μετρικές Εκπαίδευσης

Οι μετρικές στην εκπαίδευση βοηθάνε στο να αξιολογήσεις την επίδοση των μοντέλων σου και να τη συγκρίνεις με διαφορετικά μοντέλα. Η επιλογή τους διαφέρει από πρόβλημα σε πρόβλημα. Για παράδειγμα σε προβλήματα όπου γίνεται πρόβλεψη τιμών σε συνεχή χώρο χρησιμοποιούνται διαφορετικές μετρικές από ότι σε προβλήματα ταξινόμησης. Παρακάτω δίνονται ορισμένες βασικές μετρικές που χρησιμοποιήσαμε στην εκπαίδευση των μοντέλων μας.

1. Το **MSE (mean squared error)** υπολογίζεται με τη μέση τιμή του τετραγώνου της απόκλισης των προβλέψεων από τις πραγματικές τιμές. Η μετρική αυτή χρησιμοποιείται σε προβλήματα που έχουν να κάνουν με προβλέψεις τιμών σε συνεχή χώρο.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

   Όπου n ο αριθμός των δεδομένων, Y_i οι πραγματικές τιμές και $\widehat{Y}_i$.

2. Η **Categorical Crossentropy** λαμβάνει υπόψιν την αβεβαιότητα της πρόβλεψης. Χρησιμοποιείται όταν η έξοδος του μοντέλου είναι πιθανοτικές προβλέψεις που αντιστοιχούν σε πολλές κλάσεις.

$$CategoricalCrossEntropy = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * \log(p_{ij})$$

   Όπου το y_{ij} είναι 1 αν το δείγμα i ανήκει στην κλάση j, αλλιώς είναι 0 και p_{ij} είναι η πιθανότητα το μοντέλο να προβλέψει ότι το δείγμα i ανήκει στην κλάση j.

3. Η **Ακρίβεια (Accuracy)** υπολογίζει τον λόγο των σωστά ταξινομημένων δειγμάτων προς το σύνολο όλων των δειγμάτων. Χρησιμοποιείται ευρέως και αποτελεί αντικειμενική μετρική για τη σύγκριση αρχιτεκτονικών μοντέλων. Χρησιμοποιείται κυρίως σε προβλήματα ταξινόμησης.

4. Ο συντελεστής **Pearson correlation** είναι ένα στατιστικό μέτρο που υπολογίζει τη γραμμική σύνδεση μεταξύ δύο μεταβλητών X και

Y. Έχει τιμή μεταξύ -1 και 1. Η τιμή 1 σημαίνει ότι έχουν θετική γραμμική σχέση, η τιμή 0 καθόλου και η τιμή -1 ότι έχουν αρνητική γραμμική σχέση. Το χρησιμοποιήσαμε ως μετρική στα προβλήματα όπου χρειάστηκε να κάνουμε linear regression.

$$PearsonCorrelation = r_{X,Y} = \frac{cov(X,Y)}{s_X s_Y}$$

Όπου $cov(X,Y)$ η συνδιακύμανση τον μεταβλητών X, Y και s_X, s_Y η τυπική τους απόκλιση.

Για την αξιολόγηση της εκπαίδευση των μοντέλων μας, σχεδιάζαμε δύο γραφικές παραστάσεις. Η πρώτη περιέγραφε τη μεταβολή της απώλειας στον αριθμό εποχών και η δεύτερη τη μεταβολή του Pearsons Correlations ή του Accuracy στον αριθμό εποχών. Τη μετρική Pearsons Correlation τη χρησιμοποιήσαμε για την πρόβλεψη των ενδιάμεσων χαρακτηριστικών (σύνολο δεδομένων της Aljanaki) και του συναισθήματος (σύνολο δεδομένων Soundtracks). Τη μετρική Accuracy τη χρησιμοποιήσαμε για την ταξινόμηση του πολυτροπικού συνόλου δεδομένου στις 5 ομάδες του MIREX.

3.4 Υπερ-παράμετροι Εκπαίδευσης

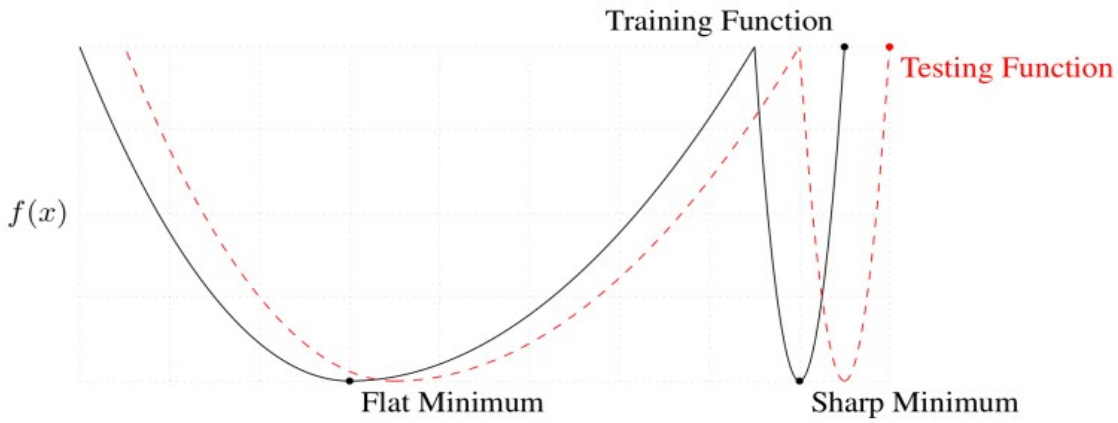
Οι περισσότεροι αλγόριθμοι εκπαίδευσης έχουν ρυθμίσεις οι οποίες ελέγχουν την εκπαίδευση που ονομάζονται υπερ-παράμετροι. Οι ρυθμίσεις αυτές αφορούν διάφορες μεταβλητές της εκπαίδευσης οι οποίες αξιολογούνται από το validation set. Κάποια παραδείγματα υπερ-παραμέτρων είναι ο αριθμός κρυμμένων επιπέδων ενός δικτύου, η τιμή ενός dropout layer, ο ρυθμός μάθησης της εκπαίδευσης ή το μέγεθος του batch. Η επιλογή των κατάλληλων υπερ-παραμέτρων επηρεάζει άμεσα την επίδοση των μοντέλων μας. Εδώ είναι σημαντικό να αναφέρουμε ότι για αλγόριθμο βελτιστοποίησης της συνάρτησης απώλειας χρησιμοποιήσαμε τον ADAM (Kinga et al.). Οι σημαντικότεροι υπερ-παράμετροι είναι ο ρυθμός μάθησης της εκπαίδευσης και το μέγεθος του batch (ο αριθμός των δεδομένων που τροφοδοτείται το δίκτυο). Η επιλογή λάθος τιμών για τις παραπάνω μεταβλητές μπορεί να οδηγήσει σε πολύ κακά αποτελέσματα.

Οι τρόποι για να επιλέξει κάποιος τις τιμές των υπερ-παραμέτρων είναι δύο. Ο πρώτος είναι ο εμπειρικός και ο δεύτερος μέσω της αυτοματοποιημένης επιλογής τους. Εμείς προσπαθήσαμε να δοκιμάσουμε και τα δύο. Αρχικά επιλέξαμε τις υπερ-παραμέτρους εμπειρικά με

ορισμένα κριτήρια που θέτουν οι Sepp Hochreiter et al. και Dinh, L et al. για την ομαλότητα του τοπικού ελαχίστου και τη σύνδεση του με το batch_size και το learning_rate.

Χρειαζόμαστε την ομαλότητα στο τοπικό ελάχιστο προκειμένου να έχουμε καλύτερη γενίκευση. Αν έχουμε ομαλότητα, τότε σε μια μικρή μετατόπιση στη συνάρτηση απώλειας, η οποία δικαιολογείται λόγω διαφορών του training set και του test set, δεν θα υπάρξει αισθητή αύξηση της απώλειας. Σε αντίθεση αν προσέγγιζε η συνάρτηση απώλειας ένα τοπικό ελάχιστο όπου ήταν απότομη η καμπύλη, τότε η διαφορά στη συνάρτηση απώλειας από το training set στο test set θα ήταν αισθητή. (Σχήμα 3.9) Ποιοι είναι όμως οι παράγοντες που οδηγούν τη συνάρτηση απώλειας σε ομαλό τοπικό ελάχιστο;

Έχει αποδειχθεί ότι το μέγεθος του batch επηρεάζει την ομαλότητα του τοπικού ελαχίστου. Συγκεκριμένα όσο πιο μικρό μέγεθος του batch έχουμε, τόσο περισσότερο θορυβώδη συνάρτηση απώλειας έχουμε, άρα τόσο πιο πιθανό να αποφευχθούν τα αιχμηρά ελάχιστα. Το ίδιο ισχύει αντίστοιχα και για το ρυθμό μάθησης (learning rate). Για το λόγο αυτό επιλέξαμε σχετικά μικρές τιμές, για το batch size 8 και για το learning rate 0.0005, προκειμένου να έχουμε ομαλά τοπικά ελάχιστα.

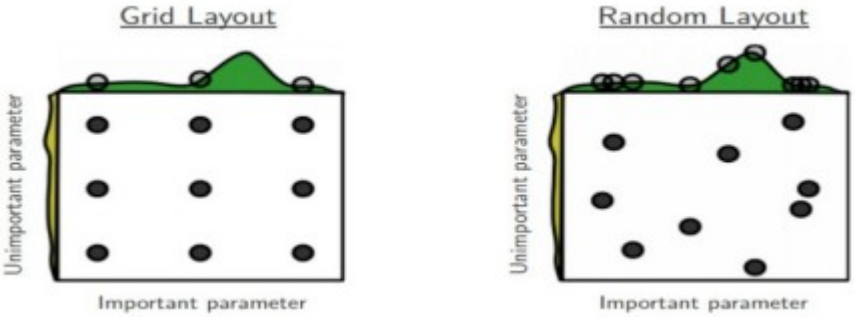


Σχήμα 3.9: Απεικόνιση του ομαλού και αιχμηρού ελαχίστου.

Πέρα από τις εμπειρικές προσεγγίσεις της ρύθμισης των υπερ-παραμέτρων, εκμεταλλευόμενοι τη δυνατότητα των υπολογιστών και των επεξεργαστών GPU για παράλληλη επεξεργασία, η ρύθμιση μπορεί να γίνει αυτόματα. Παρακάτω περιγράφονται ορισμένοι αλγόριθμοι αυτόματης ρύθμισης των υπερ-παραμέτρων.

• **Grid search**— Με τον grid αλγόριθμο δοκιμάζουμε όλους τους πιθανούς συνδυασμούς των υπερ-παραμέτρων. Παρόλο που είναι εύκολη διαδικασία προκειμένου να πραγματοποιηθεί παράλληλα η πολυπλοκότητα της αυξάνεται εκθετικά με τον αριθμό των υπερ-παραμέτρων. (the curse of dimensionality)

• **Random search**— Η τυχαία αναζήτηση είναι μια παραλλαγή του grid search όπου επιλέγονται τυχαία οι συνδυασμοί των υπερ-παρμέτρων. Σε γενικές γραμμές είναι πιο αποτελεσματικό για υψηλών διαστάσεων χώρους αναζήτησης ενώ υπολειτουργεί για μικρών διαστάσεων.



Σχήμα 3.10: Σύγκριση αλγορίθμων Random και Grid Search.

• **ASHA**— Ο ASHA Scheduler είναι κλιμακωτός αλγόριθμος με early stopping. Ο ASHA τερματίζει τις δοκιμές που εκτιμάει ότι δε θα έχουν καλό αποτέλεσμα και δίνει χρόνο σε άλλες με πιθανώς καλύτερο αποτέλεσμα. Επιπλέον επιτρέπει την παράλληλη επεξεργασία δοκιμών με χρήση πολλαπλών επεξεργαστών GPU, CPU. (Liam Li et al.)

	Random	Adaptive	Evolutionary
Sequential	Grid search / Random search	Bayesian optimization	Genetic algorithm
Parallel	Asynchronous Successive Halving Algorithm (ASHA)	Bayesian optimization with Hyperband (BOHB)	Population Based Training

Πίνακας 3.5: Περιέχει τους κυριότερους αλγορίθμους αυτοματοποιημένης ρύθμισης των υπερ-παραμέτρων.

Για τη χρήση των αλγορίθμων χρησιμοποιήσαμε τη βιβλιοθήκη tune στο μοντέλο A2Mid.

Trial name	status	loc	batch_size	loss	Pearsons_Correlation	training_iteration
DEFAULT_5c660_00000	TERMINATED		2	0.0143655	0.660678	20
DEFAULT_5c660_00001	TERMINATED		4	0.0142721	0.663793	20
DEFAULT_5c660_00002	TERMINATED		8	0.014004	0.670547	20
DEFAULT_5c660_00003	TERMINATED		16	0.0134458	0.693899	20
DEFAULT_5c660_00004	TERMINATED		2	0.0149055	0.642885	20
DEFAULT_5c660_00005	TERMINATED		4	0.0141448	0.672515	20
DEFAULT_5c660_00006	TERMINATED		8	0.0134424	0.686591	20
DEFAULT_5c660_00007	TERMINATED		16	0.0138063	0.675868	20
DEFAULT_5c660_00008	TERMINATED		2	0.0167815	0.615554	20
DEFAULT_5c660_00009	TERMINATED		4	0.0139709	0.672444	20
DEFAULT_5c660_00010	TERMINATED		8	0.0141376	0.671609	20
DEFAULT_5c660_00011	TERMINATED		16	0.013147	0.694764	20
DEFAULT_5c660_00012	TERMINATED		2	0.0144829	0.659293	20
DEFAULT_5c660_00013	TERMINATED		4	0.0137832	0.677614	20
DEFAULT_5c660_00014	TERMINATED		8	0.0144323	0.674205	20
DEFAULT_5c660_00015	TERMINATED		16	0.0135316	0.685168	20

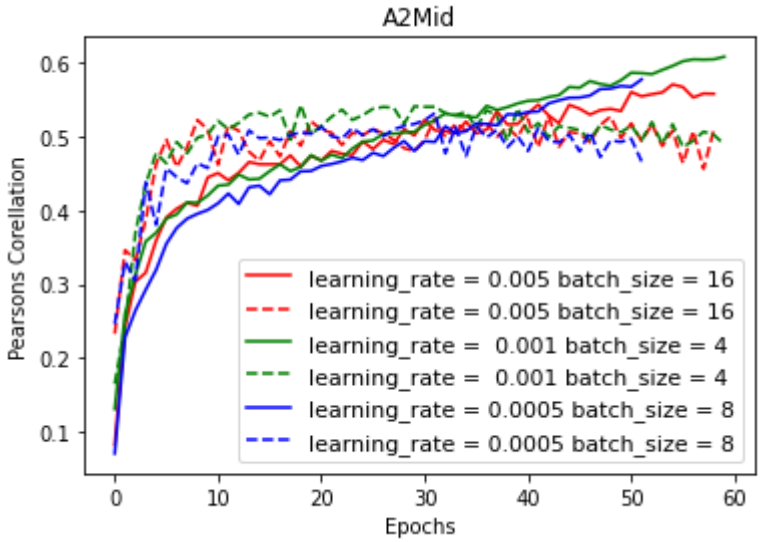
Πίνακας 3.6: Παράδειγμα χρήσης του αλγορίθμου grid search για την εύρεση κατάλληλης τιμής για το batch_size.

Trial name	status	loc	lr	loss	Pearsons_Correlation	training_iteration
DEFAULT_135b0_00000	TERMINATED		5e-05	0.0142534	0.66494	30
DEFAULT_135b0_00001	TERMINATED		0.01	0.0136485	0.684239	30
DEFAULT_135b0_00002	TERMINATED		0.01	0.0141951	0.669082	30
DEFAULT_135b0_00003	TERMINATED		0.05	0.0187527	0.540373	2
DEFAULT_135b0_00004	TERMINATED		0.05	0.0183825	0.55067	2
DEFAULT_135b0_00005	TERMINATED		5e-05	0.0737282	0.356819	1
DEFAULT_135b0_00006	TERMINATED		0.1	0.0191324	0.497706	2
DEFAULT_135b0_00007	TERMINATED		0.005	0.013864	0.676339	30
DEFAULT_135b0_00008	TERMINATED		0.001	0.0189781	0.586013	4
DEFAULT_135b0_00009	TERMINATED		5e-05	0.0482963	0.292327	1
DEFAULT_135b0_00010	TERMINATED		0.0001	0.0181406	0.548363	2
DEFAULT_135b0_00011	TERMINATED		0.1	0.0329495	0.436479	1
DEFAULT_135b0_00012	TERMINATED		0.0001	0.0251354	0.404397	1
DEFAULT_135b0_00013	TERMINATED		0.001	0.0128773	0.703091	30
DEFAULT_135b0_00014	TERMINATED		5e-05	0.0551192	0.32103	1
DEFAULT_135b0_00015	TERMINATED		0.001	0.0185007	0.55272	2
DEFAULT_135b0_00016	TERMINATED		5e-05	0.0467333	0.282117	1
DEFAULT_135b0_00017	TERMINATED		0.1	0.0194972	0.482536	2
DEFAULT_135b0_00018	TERMINATED		5e-05	0.0391049	0.328173	1
DEFAULT_135b0_00019	TERMINATED		0.001	0.0172882	0.609628	4
DEFAULT_135b0_00020	TERMINATED		0.0001	0.0186704	0.52672	2
DEFAULT_135b0_00021	TERMINATED		0.05	0.0184502	0.532347	2
DEFAULT_135b0_00022	TERMINATED		5e-05	0.0438413	0.335705	1
DEFAULT_135b0_00023	TERMINATED		0.01	0.0144914	0.656567	16
DEFAULT_135b0_00024	TERMINATED		0.05	0.0197373	0.487014	2
DEFAULT_135b0_00025	TERMINATED		0.0005	0.0194787	0.544133	4
DEFAULT_135b0_00026	TERMINATED		0.0005	0.0172	0.602433	4
DEFAULT_135b0_00027	TERMINATED		0.001	0.0167294	0.591892	4
DEFAULT_135b0_00028	TERMINATED		0.005	0.0134125	0.688757	30
DEFAULT_135b0_00029	TERMINATED		0.005	0.0184012	0.597482	2
DEFAULT_135b0_00030	TERMINATED		0.0001	0.0215316	0.472893	1
DEFAULT_135b0_00031	TERMINATED		0.001	0.0190617	0.591689	2

```
Best trial config: {'lr': 0.001}
Best trial final validation loss: 0.012877253540368243
Best trial final validation accuracy: 0.7030909386591357
```

Πίνακας 3.7: Παράδειγμα χρήσης του αλγορίθμου ASHA για την εύρεση κατάλληλης τιμής για το learning_rate.

Οι καλύτερες τιμές στις οποίες κατέληξαν οι αλγόριθμοι για το learning rate και το batch_size είνα learning rate 0.0005 και batch_size 8, learning rate 0.001 και batch_size 4, learning rate 0.005 και batch_size 16. Τελικά επιλέξαμε να κρατήσουμε το learning_rate 0.0005 και batch_size 8 γιατί ήταν η επιλογή μας και στην εμπειρική προσέγγιση. Επιπλέον οι αποδόσεις των τριών συνδυασμών όσον αφορά τις μετρικές και την ταχύτητα εκπαίδευσης ήταν παρόμοιες (Σχήμα 3.11).



Σχήμα 3.11: 3 πειράματα για κάθε συνδυασμό υπερ-παραμέτρων στο A2Mid μοντέλο.

3.5 Νευρωνικά Δίκτυα

Η επιλογή της κατάλληλης αρχιτεκτονικής του νευρωνικού δικτύου για το πρόβλημα που επιλύουμε είναι σημαντική. Ένα από τα κριτήρια για την επιλογή κατάλληλου δικτύου είναι το μέγεθος της χωρητικότητας του. Η χωρητικότητα σχετίζεται με το πόσα χαρακτηριστικά μπορεί να μάθει το μοντέλο από τα δεδομένα. Σε περίπτωση που είναι μικρή, δεν μπορεί να μάθει τα δεδομένα σαν σύνολο. Σε περίπτωση που είναι μεγάλη, μαθαίνει λεπτομέρειες των δεδομένων και όχι την ουσία τους, κάτι το οποίο χειροτερεύει τη γενίκευση του.

Οι αρχιτεκτονικές που επιλέξαμε βασίζονται σε πλήρως συνδεδεμένα δίκτυα, σε συνελικτικά δίκτυα και σε αναδρομικά δίκτυα. Στο τέλος χρησιμοποιήσαμε αυτοκωδικοποιητές για να υλοποιήσουμε την προ-εκπαίδευση των αρχιτεκτονικών μας δικτύων. Το μεγαλύτερο μέρος των πειραμάτων μας έγινε στο σύνολο δεδομένων της Aljanaki με αναπαράσταση spotify χαρακτηριστικών. Έγιναν πειράματα και στο σύνολο δεδομένων της Aljanaki με αναπαράσταση MFCC και χρήση συνελικτικών δικτύων.

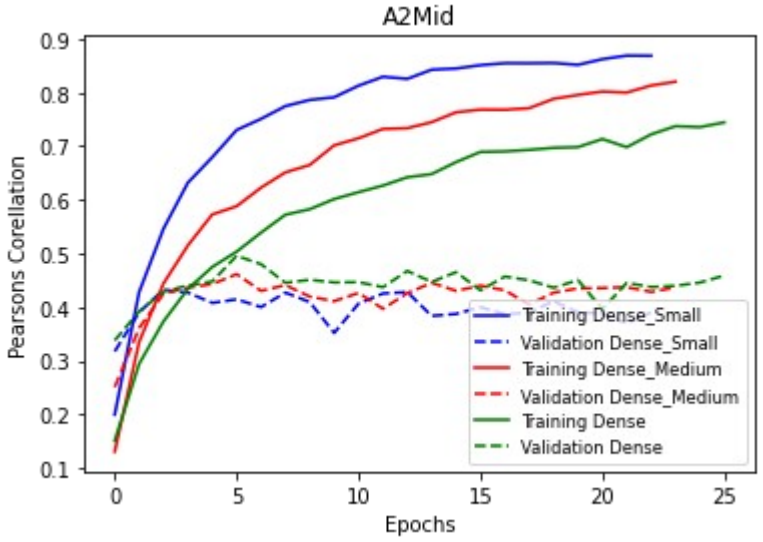
3.5.1 Τεχνητά Νευρωνικά Δίκτυα

Τα τεχνητά νευρωνικά μας δίκτυα είναι πλήρως συνδεδεμένα δίκτυα. Το αρνητικό με αυτά είναι ότι χρειάζονται πολλές παραμέτρους και δεν μπορούν να μαθαίνουν τοπικά χαρακτηριστικά των δεδομένων. Τα παραπάνω δύο χαρακτηριστικά οδηγούν σε φαινόμενα υπερπροσαρμογής (overfitting), όπου η καμπύλη του training set έχει μεγάλη απόκλιση από την καμπύλη του validation set. Παρόλα αυτά τα αποτελέσματα των δικτύων που δοκιμάσαμε δεν ήταν άσχημα.

Δοκιμάσαμε αρχιτεκτονικές πλήρως συνδεδεμένων δικτύων με διαφορετικό βάθος και αριθμό παραμέτρων και τις συγκρίναμε. Η εκπαίδευση έγινε με το σύνολο δεδομένων της Aljanaki στα ενδιάμεσα χαρακτηριστικά με αναπαράσταση spotify χαρακτηριστικών.

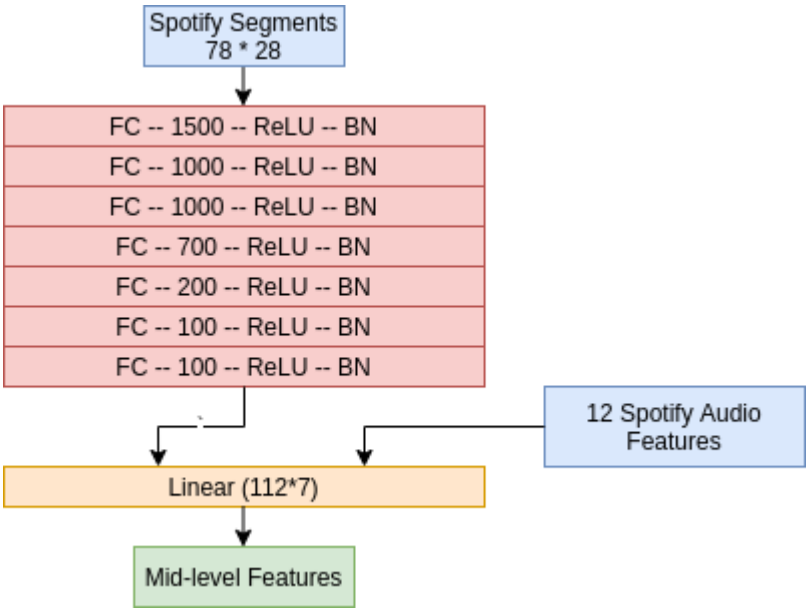
Μοντέλο	Αριθμός Νευρώνων	Αριθμός Επιπέδων	Αριθμός παραμέτρων	Pearson Correlation
Dense	4600	8	6.700.000	0.480
Dense_Medium	1100	5	1.700.000	0.433
Dense_Small	400	3	500.000	0.410

Πίνακας 3.8: Πλήρως συνδεδεμένα δίκτυα.



Σχήμα 3.12: Σύγκριση πλήρως συνδεδεμένων δικτύων.

Παρατηρήσαμε ότι το καλύτερο δίκτυο είναι το Dense καθώς έχει λιγότερο overfitting και καλύτερη επίδοση στις μετρικές.



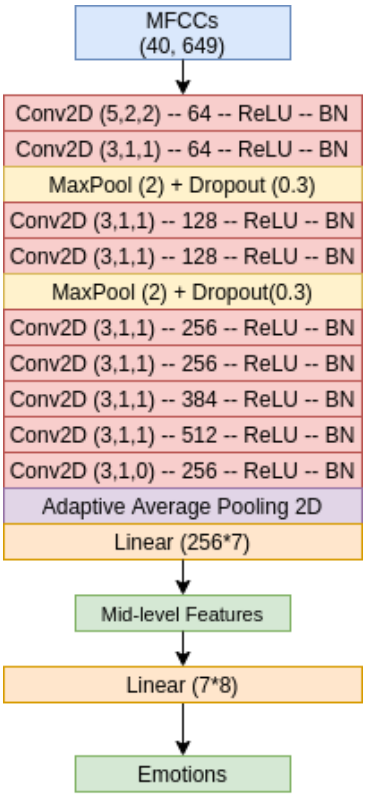
Σχήμα 3.13: Η αρχιτεκτονική του Dense.

3.5.2 Συνελικτικά Νευρωνικά Δίκτυα

Τα συνελικτικά νευρωνικά δίκτυα σε αντίθεση με τα τεχνητά χρειάζονται λιγότερες παραμέτρους και μπορούν να μάθουν τοπικά χαρακτηριστικά. Δοκιμάσαμε συνελικτικά νευρωνικά δίκτυα στο σύνολο δεδομένων της Aljanaki και για τις δύο αναπαραστάσεις των συνόλων δεδομένων. Στην συνέχεια αναλύεται το κάθε πείραμα ξεχωριστά.

3.5.2.1 Librosa

Πήραμε την αναπαράσταση MFCC του συνόλου δεδομένων της Aljanaki και εκπαιδεύσαμε CNN μοντέλα έτσι ώστε να προβλέπουν ενδιάμεσα χαρακτηριστικά και ύστερα να συνδέονται γραμμικά με το συναίσθημα. Η υλοποίηση έγινε με CNN δίκτυο που αποτελείται από Conv2D, dropout, BatchNormalization και ReLU επίπεδα, όπου η έξοδος τους συνδέεται με γραμμικό επίπεδο με τα ενδιάμεσα χαρακτηριστικά τα οποία συνδέονται με γραμμικό επίπεδο με τα συναισθήματα (Σχήμα 3.14).



Σχήμα 3.14: Η αρχιτεκτονική του A2Mid2E.

Η εκπαίδευση έγινε σε δύο στάδια. Στο πρώτο εκπαιδεύσαμε το μοντέλο A2MidE, χωρίς το τελευταίο γραμμικό επίπεδο, για να προβλέπει τα ενδιάμεσα χαρακτηριστικά των 5000 τραγουδιών. Στο δεύτερο στάδιο παγώσαμε τα επίπεδα που προβλέπουν τα ενδιάμεσα χαρακτηριστικά και προσθέσαμε το γραμμικό επίπεδο που συνδέει τα ενδιάμεσα χαρακτηριστικά με το συναίσθημα. Τα αποτελέσματα που πήραμε σχετικά με τις μετρικές ήταν πολύ καλά.

Mid-level Features	**Pearson Correlation**
Melody	0.721
Articulation	0.861
Rythm_Complexity	0.472
Rythm_Stability	0.708
Dissonance	0.745
Atonality	0.529
Mode	0.463
Average	0.630

Πίνακας 3.9: Οι τιμές του Pearson Correlation για τα ενδιάμεσα χαρακτηριστικά.

Emotions	Pearson Correlation
Valence	0.673
Energy	0.705
Tension	0.677
Anger	0.620
Fear	0.626
Happy	0.442
Sad	0.569
Tender	0.590
Average	0.600

Πίνακας 3.10: Οι τιμές του Pearson Correlation για τα συναισθήματα.

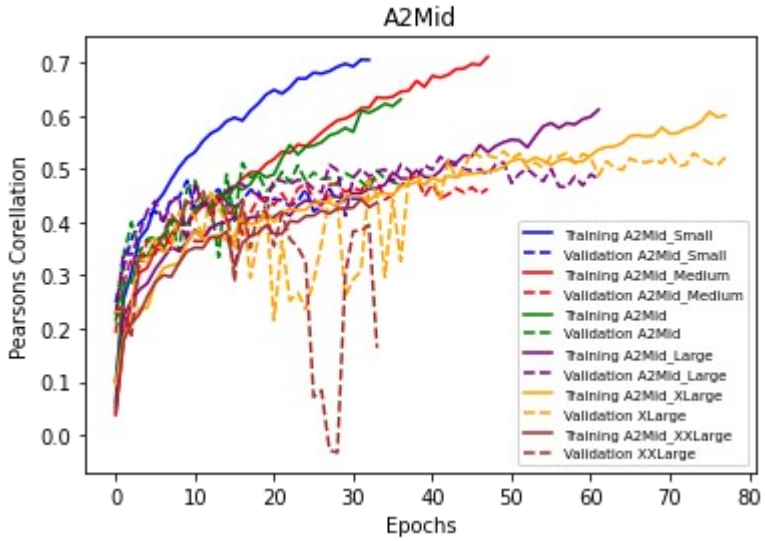
3.5.2.2 Spotify

Όσον αφορά την εκπαίδευση στο σύνολο δεδομένων της Aljanaki, σε αναπαράσταση spotify χαρακτηριστικών, χρησιμοποιήσαμε διάφορες αρχιτεκτονικές συνελικτικών δικτύων. Το βασικό τους περιεχόμενο ήταν Conv1d, dropout, maxpool, ReLU και Fc επίπεδα. Σε αυτή την αναπαράσταση σε σύγκριση με την MFCC χρησιμοποιήσαμε μιας διάστασης συνελικτικά δίκτυα, έτσι ώστε η πράξη της συνέλιξης να γίνεται στο πεδίο του χρόνου και όχι στα πεδία του χρόνου και της συχνότητας.

Παρακάτω περιγράφονται τα αποτελέσματα των διάφορων δικτύων που δοκιμάσαμε. Η βασική διαχωριστική τους είναι το μέγεθος τους και ο αριθμός των παραμέτρων που περιέχουν. Η εκπαίδευση τους έγινε για την πρόβλεψη ενδιάμεσων χαρακτηριστικών στα 2196 τραγούδια του συνόλου δεδομένων της Aljanaki σε αναπαράσταση spotify χαρακτηριστικών.

Μοντέλα	Αριθμός παραμέτρων	Αριθμός επιπέδων	Pearson Correlation
A2Mid_Small	150.000	7	0.432
A2Mid_Medium	300.000	10	0.452
A2Mid	1.300.000	12	0.479
A2Mid_Large	4.000.000	15	0.501
A2Mid_XLarge	8.000.000	18	0.510
A2Mid_XXLarge	15.000.000	21	0.401

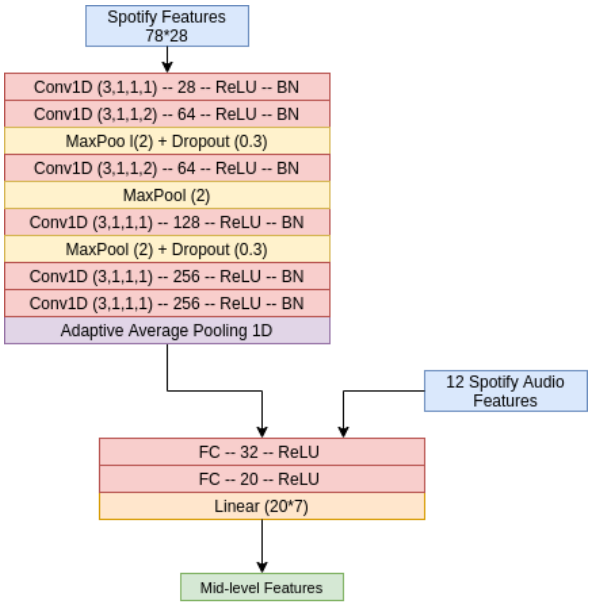
Πίνακας 3.11: Συνελιτικά Δίκτυα.



Σχήμα 3.15: Σύγκριση Συνελικτικών Δικτύων.

Παρατηρώντας τα συνελικτικά δίκτυα διαπιστώνει κανείς ότι έχουμε βελτίωση των μετρικών καθώς αυξάνουμε το μέγεθος του δικτύου μέχρι ένα σημείο όπου έχουμε απότομη πτώση του. Αυτό ερμηνεύεται με το γεγονός ότι από ένα σημείο και πέρα το δίκτυο μας κάνει overfitting, λόγω μεγάλης χωρητικότητας του δικτύου. Για τον λόγο αυτόν, το μοντέλο A2Mid_XXLarge έχει επίδοση 0.401, ενώ το μοντέλο A2Mid_XLarge έχει επίδοση 0.510.

Παρόλο που το A2Mid_XLarge έχει την καλύτερη επίδοση στις μετρικές, παρατηρήσαμε ότι εμφανίζει συχνά overfitting. Το ίδιο συμβαίνει και με το A2Mid_Large. Για τους παραπάνω λόγους επιλέξαμε τα μοντέλα A2Mid και A2Mid_Medium με την προοπτική να βελτιώσουμε την επίδοση τους μέσω τεχνικών ομαλοποίησης.



Σχήμα 3.16: Η αρχιτεκτονική του A2Mid_Medium.

3.5.3 Αναδρομικά Νευρωνικά Δίκτυα

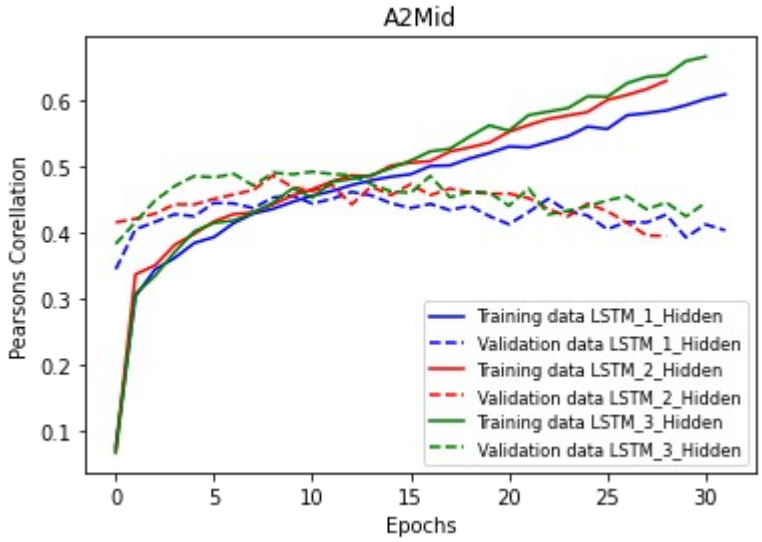
Μια από τις διαφορές που έχουν τα Αναδρομικά Νευρωνικά Δίκτυα σε σχέση με τα υπόλοιπα που χρησιμοποιήσαμε είναι ότι η έξοδος τους εξαρτάται από προηγούμενους υπολογισμούς. Στα πειράματα μας χρησιμοποιήσαμε το LSTM (Long Short Term Memory), το οποίο έχει την ικανότητα να θυμάται παλιούς υπολογισμούς.

Δοκιμάσαμε τρία μοντέλα LSTM τα οποία διαφέρουν στον αριθμό των κρυμμένων επιπέδων. Τα εκπαιδεύσαμε στο σύνολο δεδομένων της Aljanaki σε αναπαράσταση spotify χαρακτηριστικών για την πρόβλεψη τον ενδιάμεσων χαρακτηριστικών.

Μοντέλο	Αριθμός κρυμμένων επιπέδων	Pearson Correlation
LSTM_1_Hidden	1	0.440
LSTM_2_Hidden	2	0.469
LSTM_3_Hidden	3	0.465

Πίνακας 3.12: Αναδρομικά Δίκτυα.

Από ότι φαίνεται ανεξάρτητα του αριθμού των κρυμμένων επιπέδων τα αναδρομικά νευρωνικά δίκτυα έχουν παρόμοια επίδοση.

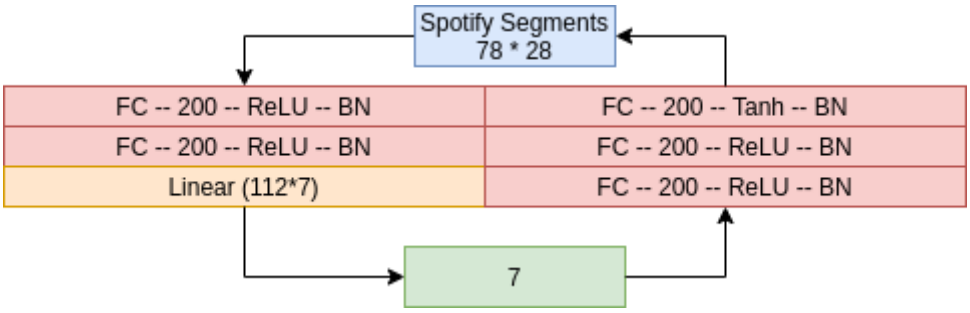


Σχήμα 3.17: Σύγκριση Αναδρομικών Δικτύων.

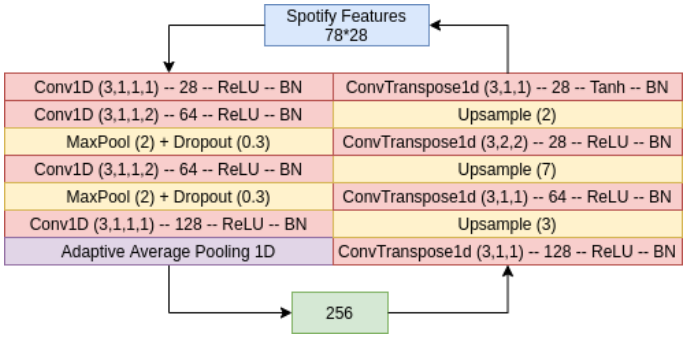
3.5.4 Αυτοκωδικοποιητές

Χρησιμοποιήσαμε Deep Autoencoders για το στάδιο της προ-εκπαίδευσης των μοντέλων μας. Η δομή των αυτοκωδικοποιητών ήταν τέτοια ώστε να μπορούμε να αντιγράψουμε τα βάρη στα τεχνητά και συνελικτικά δίκτυα που συγκροτήσαμε στην υποενότητα 3.5.3. Συγκροτήσαμε 4 αυτοκωδικοποιητές, οι οποίοι προορίζονται για τα συνελικτικά δίκτυα και 3 αυτοκωδικοποιητές που προορίζονται για τα τεχνητά.

Η δομή τους ήταν τέτοια ώστε ο κωδικοποιητής (encoder) να μειώνει το μέγεθος της εισόδου και ο αποκωδικοποιητής (decoder) να το αυξάνει. Για τους αυτοκωδικοποιητές συνελικτικών δικτύων, αντιγράψαμε τα επίπεδα που χρησιμοποιούμε στα συνελικτικά δίκτυα για τον κωδικοποιητή και για τον αποκωδικοποιητή χρησιμοποιήσαμε ConvTranspose και Upsample επίπεδα. Για τους αυτοκωδικοποιητές τεχνητών δικτύων χρησιμοποιήσαμε αντίστοιχα τα επίπεδα των τεχνητών δικτύων για τον κωδικοποιητή και για τον αποκωδικοποιητή χρησιμοποιήσαμε πλήρως συνδεδεμένα δίκτυα. Η εκπαίδευση τους έγινε με τα σύνολα δεδομένων Pre-training και Large-Pre-training που συγκροτήσαμε στην υποενότητα 3.2.1.2. Παρακάτω δίνονται δύο ενδεικτικές αρχιτεκτονικές των δικτύων των αυτοκωδικοποιητών που χρησιμοποιήσαμε.



Σχήμα 3.18: Αυτοκωδικοποιητής με πλήρως συνδεδεμένα δίκτυα (Autoencoder_Dense_Small).



Σχήμα 3.19: Αυτοκωδικοποιητής με συνελικτικα δίκτυα (Autoencoder_Small).

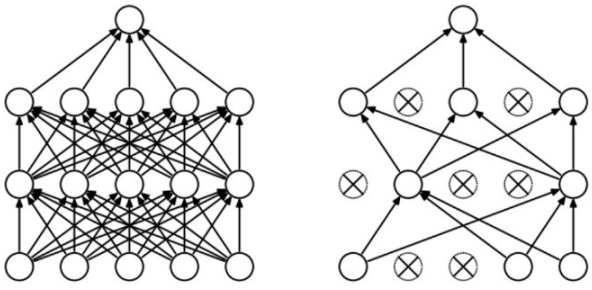
3.6 Τεχνικές Ομαλοποίησης

Οι τεχνικές ομαλοποίησης ενός νευρωνικού δικτύου απαντούν σε ένα συχνό πρόβλημα της εκπαίδευσης, το overfitting. Το overfitting του δικτύου στα δεδομένα εκπαίδευσης οδηγεί στο να χάνουν την ικανότητα γενίκευσης τους. Αυτό παρατηρείται όταν η καμπύλη εκπαίδευσης του training set σε σχέση με του validation set έχει μεγάλη απόκλιση.

Οι τεχνικές που θα περιγράψουμε ποικίλουν. Συγκεκριμένα θα αναφερθούμε στο dropout, στους όρους ποινής, στο Early Stopping, στο Data Augmentation, στο Unsupervised Pre-training, στο Supervised Pre-training και στο Multitask Learning.

3.6.1 Dropout

Η προσθήκη επιπέδων Dropout στο μοντέλο είναι ένας από τους τρόπους για την αντιμετώπιση του overfitting. Αυτό που κάνει είναι να ”απενεργοποιεί” κόμβους του νευρωνικού δικτύου κατά την εκπαίδευση, με αποτέλεσμα σε κάθε εποχή να χρησιμοποιούνται διαφορετικοί συνδυασμοί κόμβων, έτσι ώστε να μην προσαρμόζεται το σύνολο των βαρών στα δεδομένα εκπαίδευσης. Έτσι γίνεται εκπαίδευση πολλών διαφορετικών μοντέλων, μόνο που γίνεται σε λιγότερο χρόνο καθώς αυτά γίνονται λιγότερο πυκνά (Nitish Srivastava et al.).

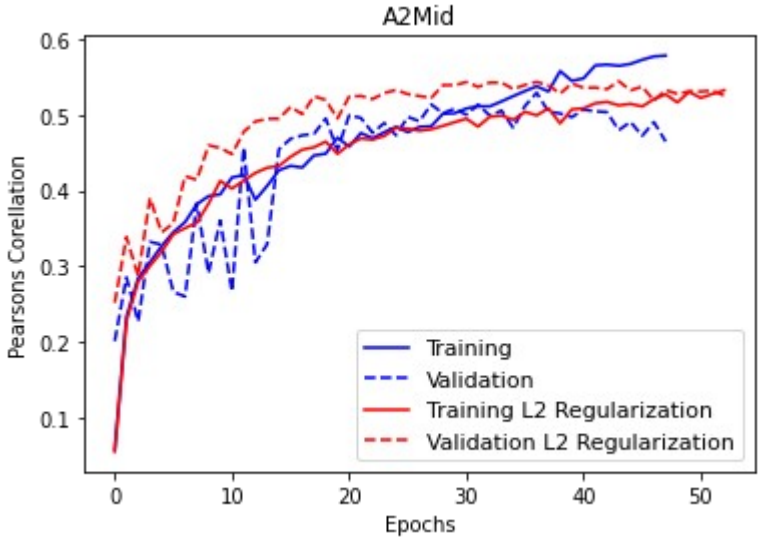


Σχήμα 3.20: Πριν και μετά τη χρήση dropout επιπέδων.

3.6.2 Όροι Ποινής

Ένας διαφορετικός τρόπος να αποφευχθεί το overfitting είναι μέσω της εφαρμογής όρων ποινής στη συνάρτηση απώλειας. Έτσι αποφεύγονται οι μεγάλες τιμές στα βάρη του μοντέλου και επιτυγχάνεται ομοιομορφία στην κατανομή των βαρών. Παρακάτω δίνονται τρία παραδείγματα όρων ποινής.

- L1 regularization $\tilde{J}(w; \mathbf{X}, \mathbf{y}) = J(w; \mathbf{X}, \mathbf{y}) + \alpha \frac{1}{2} \boldsymbol{w}^T \boldsymbol{w}$
- L2 regularization $\tilde{J}(w; \mathbf{X}, \mathbf{y}) = J(w; \mathbf{X}, \mathbf{y}) + \alpha \|w\|_1$
- Sparse Representations $\tilde{J}(\theta; \mathbf{X}, \mathbf{y}) = J(\theta; \mathbf{X}, \mathbf{y}) + \alpha \Omega(\boldsymbol{h})$

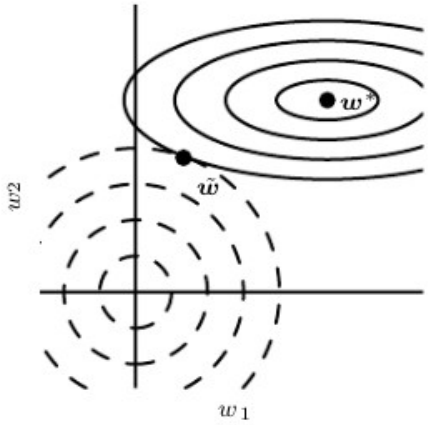


Σχήμα 3.21: Παράδειγμα εφαρμογής L2 regularization στο μοντέλο A2Mid.

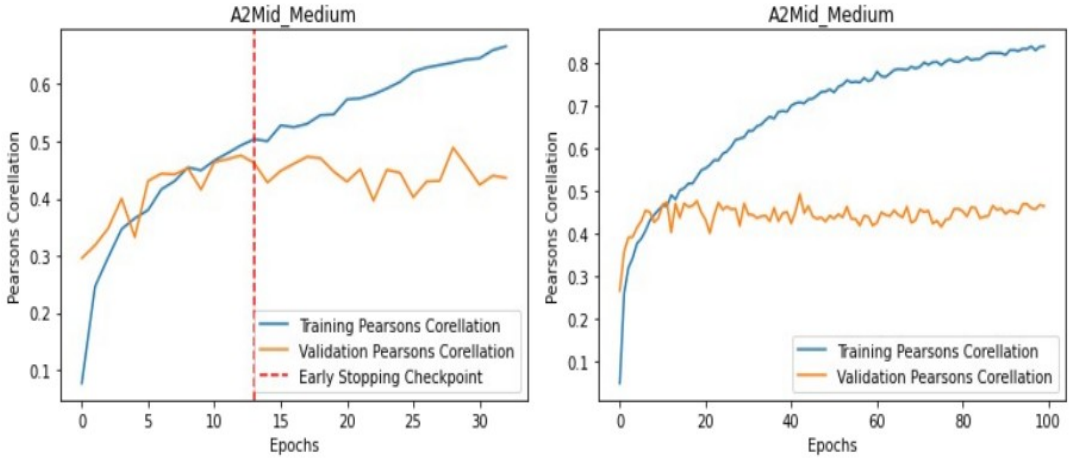
Όπως φαίνεται και στο Σχήμα 3.21 με την L2 κανονικοποίηση (L2 regularization) πετυχαίνουμε καλύτερη γενίκευση και καλύτερο αποτέλεσμα στις μετρικές. Η L2 κανονικοποίηση έχει μια υπερπαράμετρο α η οποία καθορίζει τη βαρύτητα του όρου ποινής. Όσο πιο περίπλοκο είναι το δίκτυο στο οποίο την εφαρμόζουμε τόσο πιο μεγάλη πρέπει να είναι η τιμή του.

3.6.3 Early Stopping

Μια διαφορετική τεχνική για την αποφυγή του overfitting είναι το EarlyStopping. Με την τεχνική αυτή αποθηκεύονται οι παράμετροι του μοντέλου όταν έχουμε μείωση του validation loss. Σε αντίθετη περίπτωση συνεχίζουμε την εκπαίδευση μέχρι ενός ορίου εποχών που ορίζεται σαν παράμετρος. Στο τέλος της εκπαίδευσης κρατάμε τις παραμέτρους του μοντέλου για τις οποίες είχε το μικρότερο validation loss. Η επιδόσεις του Early Stopping παρατηρείται ότι είναι παρόμοιες με της L2 κανονικοποίησης.



Σχήμα 3.22: Οι διακεκομμένες γραμμές αντιστοιχούν στο L2 regularization ενώ οι συνεχής στο Early Stopping.



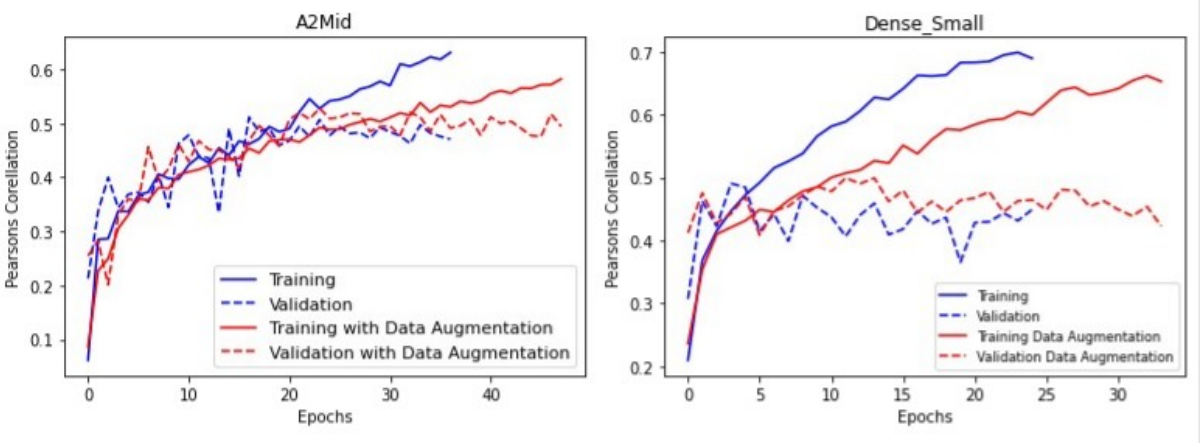
Σχήμα 3.23: Αριστέρα εκπαίδευση με Early Stopping. Δεξιά εκπαίδευση χωρίς Early Stopping.

Όπως φαίνεται και στο Σχήμα 3.23 στην περίπτωση που δεν έχουμε Early Stopping το μοντέλο εμφανίζει το φαινόμενο του overfitting.

3.6.4 Data Augmentation

Ένα μοντέλο βαθιάς μάθησης μπορεί να γενικευθεί καλύτερα αν του δοθούν περισσότερα δεδομένα. Σε ορισμένες περιπτώσεις μπορούμε να δημιουργήσουμε «τεχνητά» δεδομένα και να τα προσθέσουμε στο σύνολο εκπαίδευσης. Σε κάθε εποχή ο dataloader εφαρμόζει τυχαία μετατροπές στα δεδομένα. Άρα σε κάθε εποχή έχουμε διαφορετικά αντικείμενα. Οι μετατροπές μπορούν να είναι διάφορες. Συνηθισμένες είναι η μετατόπιση μεταβλητών, η προσθήκη θορύβου και για την επεξεργασία εικόνας, η περιστροφή εικόνας, η αποκοπή εικόνας κτλ.

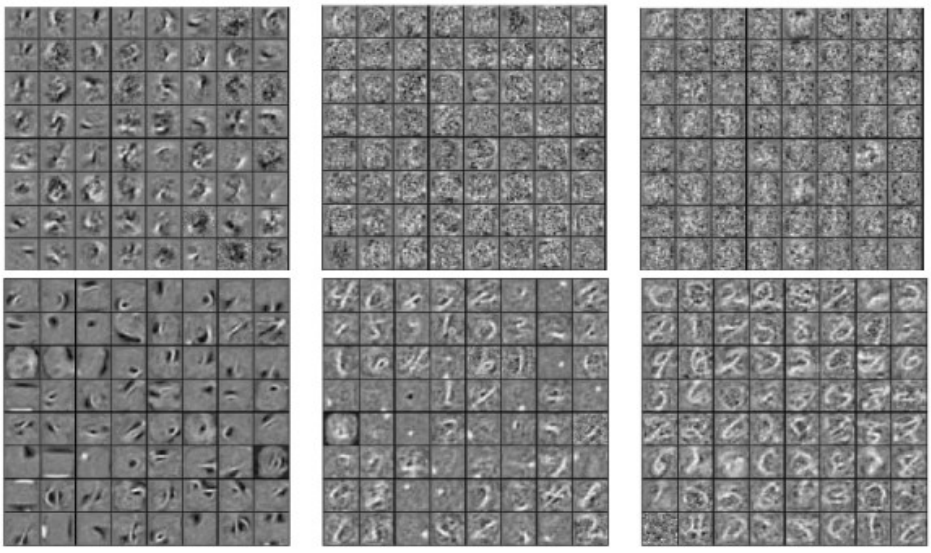
Επειδή τα δεδομένα μας αναπαριστούν μουσική διαλέξαμε να μην εφαρμόσουμε τις κλασσικές μεθόδους επαύξησης δεδομένων (data augmentation). Συγκεκριμένα επιλέξαμε να αλλάξουμε την τονικότητα και το κλειδί της εκτέλεσης του τραγουδιού. Αυτό το πετύχαμε κάνοντας σε κάθε δεδομένο μια τυχαία κυκλική μετατόπιση των θέσεων του διανύσματος του pitch. Να θυμίσουμε ότι το διάνυσμα pitch αναπαρίσταται από 12 τιμές που αντιστοιχούν σε 12 τονικότητες βαθμολογημένες σε κλίμακα από 0 μέχρι 1 ανάλογα της παρουσίας της στο segment του τραγουδιού.



Σχήμα 3.24: Data augmentation στην εκπαίδευση του μοντέλου Dense_Small και του A2Mid.

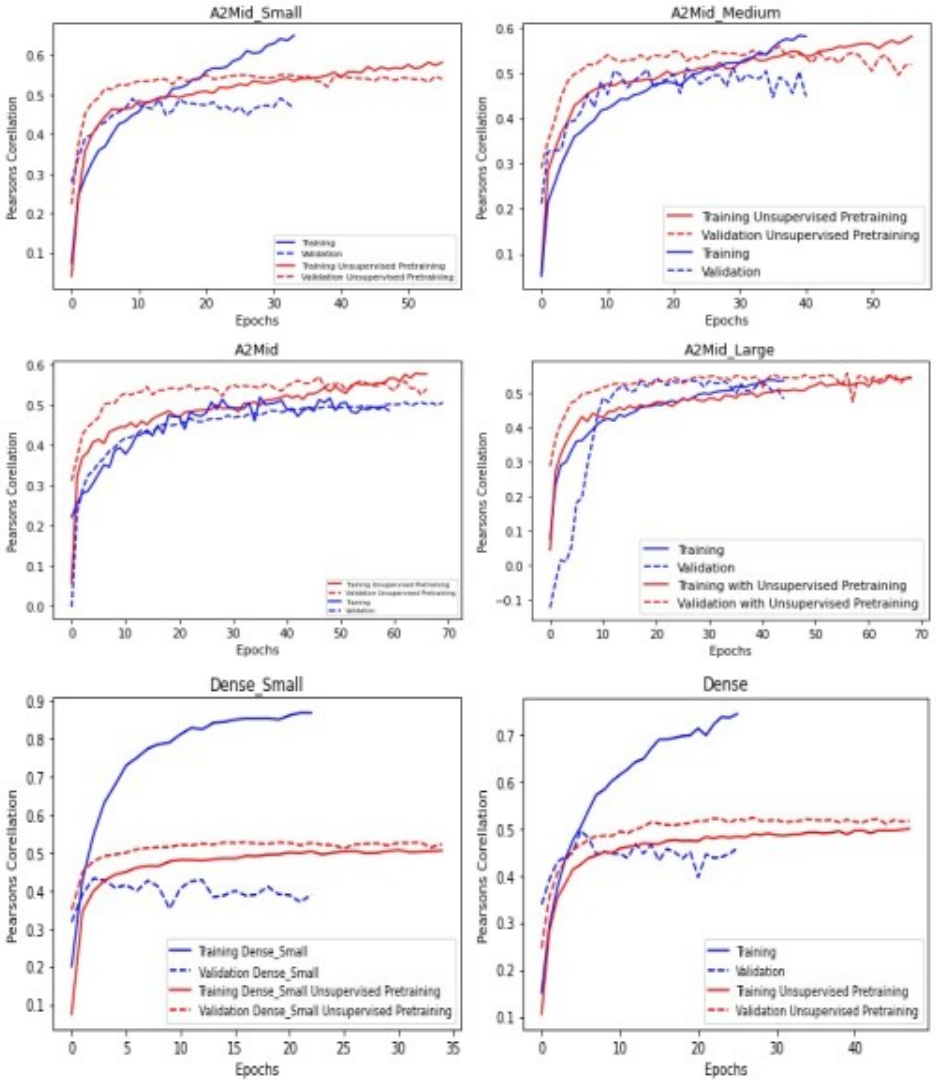
3.6.5 Unsupervised Pre-training

Η μη-επιβλεπόμενη προ-εκπαίδευση (Unsupervised Pretraining) πραγματοποιείται πριν την εκπαίδευση ενός μοντέλου επιβλεπόμενης μάθησης. Ονομάζεται “Pretraining” γιατί πραγματοποιείται πριν την εκπαίδευση και “Unsupervised” γιατί πραγματοποιείται σε δεδομένα χωρίς ταμπέλα. Στόχος της είναι η αρχικοποίηση των βαρών ενός δικτύου με βάση την εκπαίδευση ανακατασκευής μεγάλου όγκου δεδομένων. Με αυτό τον τρόπο μαθαίνει χαρακτηριστικά των δεδομένων, βοηθώντας στην εκπαίδευση και δίνει τιμές στα βάρη που διαφορετικά θα ήταν απρόσιτες, πετυχαίνοντας βελτιστοποίηση και καλύτερη γενίκευση. Τα αποτελέσματα του Unsupervised Pre-training τα αποδεικνύουν με πολλά πειράματα οι Duimtru Erhan et al.. Ο πιο συνηθισμένος τρόπος υλοποίησης του Unsupervised Pretraining είναι μέσω των αυτοκωδικοποιητών. Χρησιμοποιείται κυρίαρχα στην βαθιά μάθηση κειμένων όπου η αναπαράσταση των δεδομένων δεν είναι ικανοποιητική, αλλά όπως φαίνεται και στο Σχήμα 3.25 βοηθάει και στην βαθιά μάθηση εικόνων. Συνηθισμένη πρακτική είναι το “πάγωμα” των επιπέδων που χρησιμοποιούνται για την εξαγωγή χαρακτηριστικών και ύστερα την τοποθέτηση ενός απλού γραμμικού ταξινομητή πάνω στα προηγούμενα επίπεδα προκειμένου να γίνει η επιβλεπόμενη μάθηση.



Σχήμα 3.25: Απεικόνιση φίλτρων από ένα DBN εκπαιδευμένο στην InfiniteMNIST. Οι πάνω εικόνες απεικονίζουν τα βάρη των φίλτρων χωρίς να έχει γίνει pre-training, ενώ οι κάτω συμβολίζουν τα βάρη των φίλτρων με το να έχει γίνει pre-training. Από αριστερά προς τα δεξιάς έχουμε τους νευρώνες του 1ου, 2ου και 3ου επιπέδου.

Στη δικιά μας περίπτωση χρησιμοποιήσαμε αυτοκωδικοποιητές για να τους εκπαιδεύσουμε στα Pre-training σύνολα δεδομένων και ύστερα αντιγράψαμε τα βάρη των επιπέδων των κωδικοποιητών στις αρχιτεκτονικές μας. Σε όλες τις περιπτώσεις παρατηρείται ότι έχουμε βελτίωση στην επίδοση και στο overfitting. Αρχικά δοκιμάσαμε την προ-εκπαίδευση στο μικρό σύνολο δεδομένων, Pre-training.

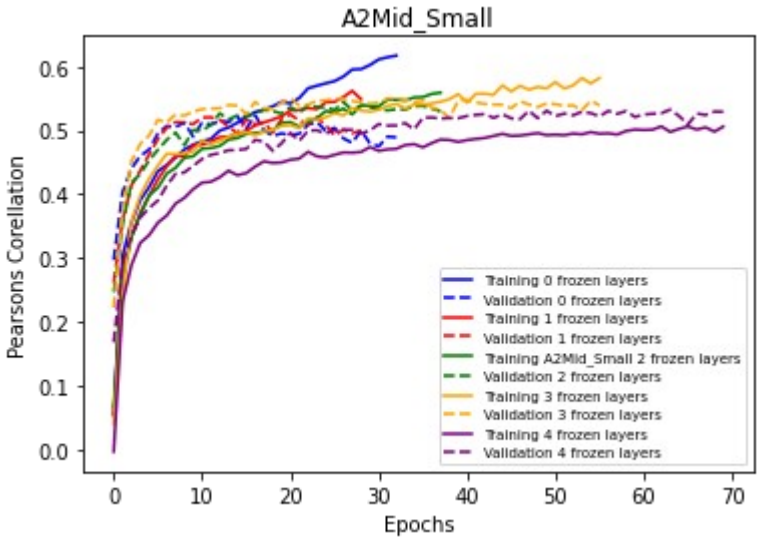


Σχήμα 3.26: Εφαρμογή Unsupervised Pre-training σε όλα τα μοντέλα στην Pre-training dataset.

Το Unsupervised Pre-training μαθαίνει στο δίκτυο να αναπαριστά και να ξεχωρίζει λεπτά χαρακτηριστικά των δεδομένων. Για τον λόγο αυτόν συνηθισμένη πρακτική είναι το “πάγωμα” των πρώτων επιπέδων του δικτύου τα οποία δε χρειάζονται περαιτέρω εκπαίδευση. Με τον όρο “πάγωμα” εννοούμε ότι δεν επιτρέπουμε την ανανέωση των βαρών των επιπέδων. Για την επιλογή του κατάλληλου αριθμού “παγωμένων” επιπέδων κάναμε πειράματα για όλες τις περιπτώσεις.

Type	DENSE			CNN			
Model	Dense_Small	Dense_Medium	Dense	A2Mid_Small	A2Mid_Medium	A2Mid	A2Mid_Large
Frozen layers	2	2	5	3	3	6	9
Average Pearsons Correlation	0,503	0,510	0,513	0,500	0,508	0,512	0,489

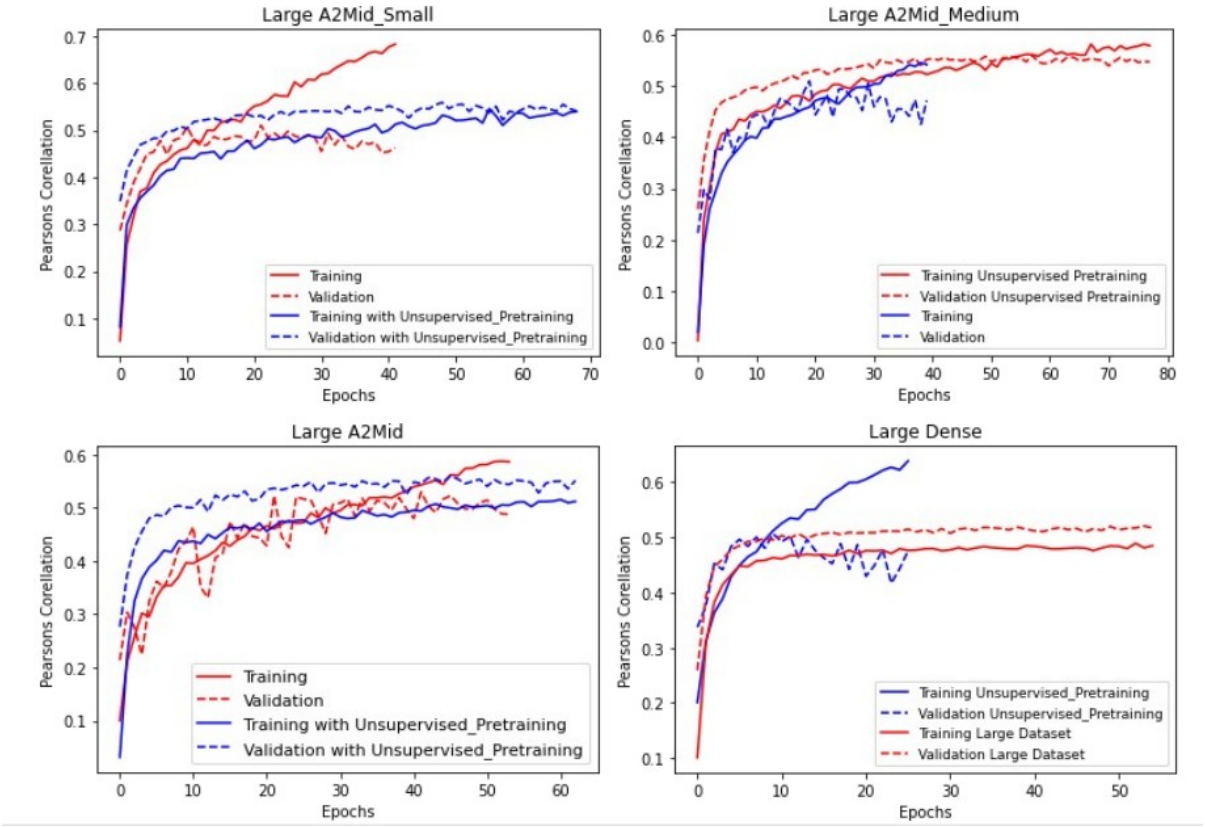
Πίνακας 3.13: Ο κατάλληλος αριθμός “παγωμένων” επιπέδων για κάθε μοντέλο.



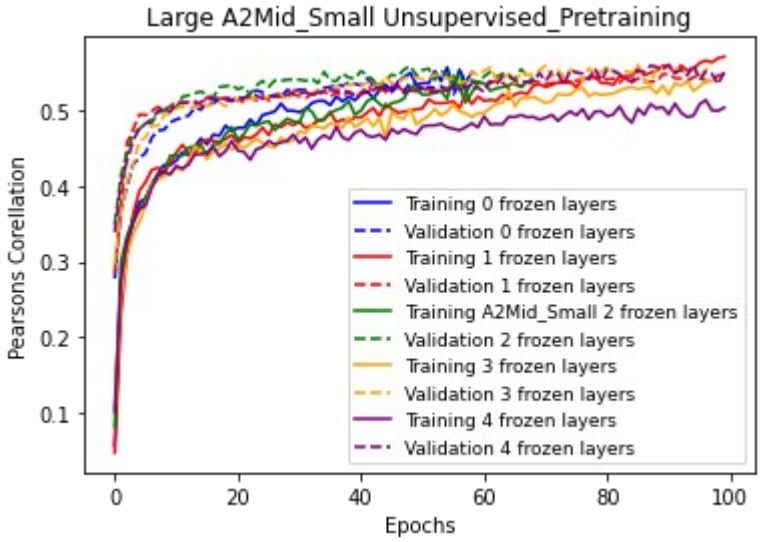
Σχήμα 3.27: Σύγκριση μοντέλων A2Mid_Small με διαφορετικό αριθμό παγωμένων επιπέδων (Pre-training).

Στο παράδειγμα του μοντέλου A2Mid_Small (Σχήμα 3.27) παρατηρείται ότι το “πάγωμα” των τριών πρώτων επιπέδων επιφέρει τα βέλτιστα αποτελέσματα όσον αφορά τις μετρικές και το overfitting. Επιπλέον φαίνεται πως στην περίπτωση “παγώματος” όλων των επιπέδων, το μοντέλο εμφανίζει το φαινόμενο της υποπροσαρμογής (underfitting).

Στη συνέχεια δοκιμάσαμε το ίδιο πείραμα μόνο που η προ-εκπαίδευση έγινε σε οκταπλάσιο σε μέγεθος σύνολο δεδομένων (Large-Pre-training). Τα αποτελέσματα ήταν καλύτερα ως προς τη μείωση του overfitting.



Σχήμα 3.28: Εφαρμογή Unsupervised Pre-training σε διάφορα μοντέλα στην Large-Pre-training dataset.



Σχήμα 3.29: Σύγκριση μοντέλων A2Mid_Small με διαφορετικό αριθμό παγωμένων επιπέδων (Large-Pre-training).

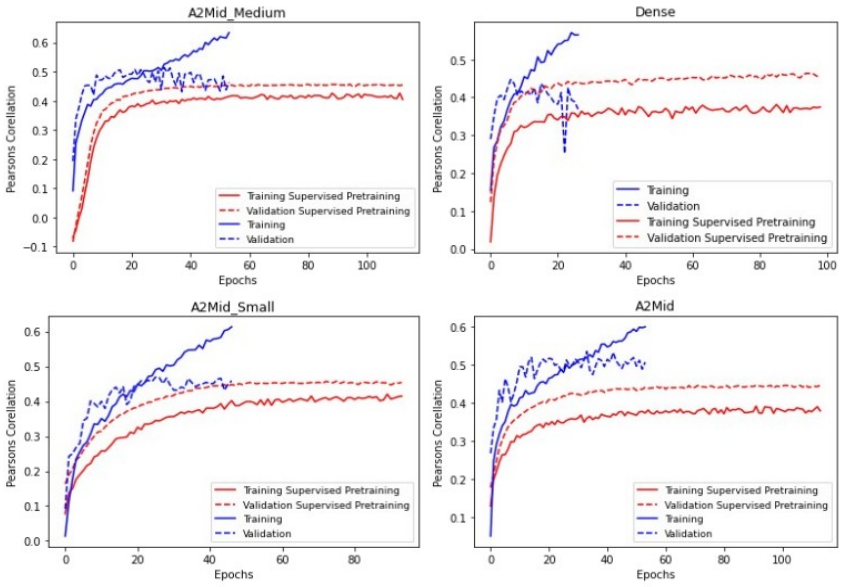
Όπως φαίνεται στο Σχήμα 3.29 η επιλογή του αριθμού των “παγωμένων” επιπέδων είναι πιο δύσκολη υπόθεση καθώς αυξάνεται το μέγεθος του συνόλου δεδομένου που κάνουμε προ-εκπαίδευση. Η επίδοση

των μοντέλων είναι εξίσου καλή ανεξαρτήτως του αριθμού των “παγωμένων” επιπέδων.

3.6.6 Supervised Pre-training

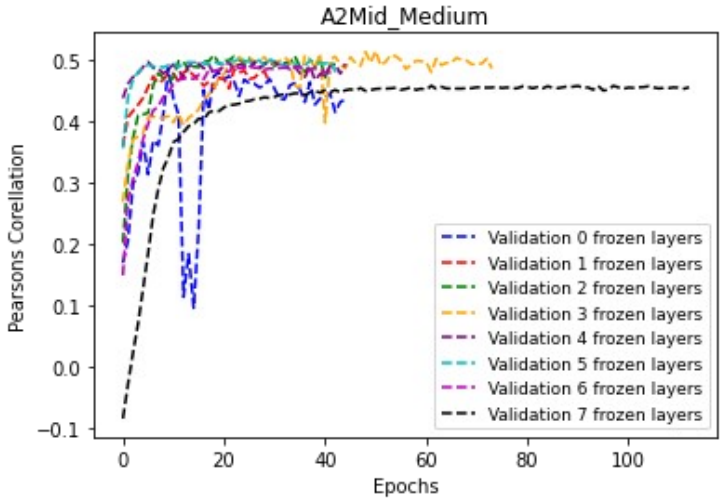
Μια άλλη τεχνική παρόμοια με το Unsupervised Pre-training που βοηθάει στη βελτιστοποίηση της επίδοσης του μοντέλου και στην αντιμετώπιση του φαινομένου overfiting είναι η επιβλεπόμενη προ-εκπαίδευση (Supervised Pre-training). Η διαφορά της είναι στο γεγονός ότι η προ-εκπαίδευση πραγματοποιείται σε δεδομένα τα οποία έχουν ταμπέλα. Καλό είναι η προ-εκπαίδευση να γίνεται σε ταμπέλες η οποίες είναι παρόμοιες με το τελικό στόχο. Με αυτόν το τρόπο το δίκτυο μαθαίνει χρήσιμα χαρακτηριστικά των δεδομένων μας για παρόμοιο σκοπό. Έμπνευση για αυτήν την πρακτική αποτέλεσαν το άρθρα των Aaron van den Oord et al. όπου εφάρμοσαν supervised pre-training σε πολλά προβλήματα ταξινόμησης της μουσικής και το άρθρο των Jordi Pons and Xavier Serra, που χρησιμοποίησαν μοντέλα για την εξαγωγή χαρακτηριστικών από τα δεδομένα.

Εμείς για να πετύχουμε το Supervised Pre-training εκπαιδεύσαμε τα μοντέλα μας σε μεγάλο όγκο τραγουδιών από το Spotify API πάνω στην πρόβλεψη των 12 τιμών των τραγουδιών. Τα σύνολα δεδομένων Pre-training και Large-Pre-training έχουν ταμπέλες τα 12 audio features των τραγουδιών. Οι τιμές αυτές είναι παρεμφερείς με τα ενδιάμεσα χαρακτηριστικά που θέλουμε να κάνουμε πρόβλεψη. Ένα παράδειγμα είναι το χαρακτηριστικό της κλίμακας που είναι κοινό.



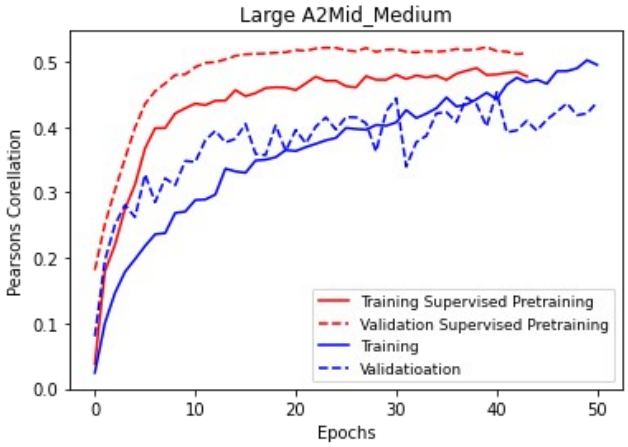
Σχήμα 3.30: Εφαρμογή Supervised Pre-training σε διάφορα μοντέλα στην Pre-training dataset.

Αρχικά κάναμε πειράματα στο μικρό σύνολο δεδομένων, Pre-training (Σχήμα 3.30). “Παγώσαμε” όλα τα επίπεδα και κάναμε αντιστοίχιση στα 7 ενδιάμεσα χαρακτηριστικά. Η επίδοση των μοντέλων δεν ήταν ικανοποιητική. Δοκιμάσαμε να “ξεπαγώσουμε” ορισμένα μοντέλα, κάτι το οποίο έδειξε ότι τα αποτελέσματα μπορούν να βελτιωθούν με τη χρήση Supervised Pre-training, αρκεί να κρατήσουμε τα βάρη ορισμένων επιπέδων αμετάβλητα (Σχήμα 3.31).



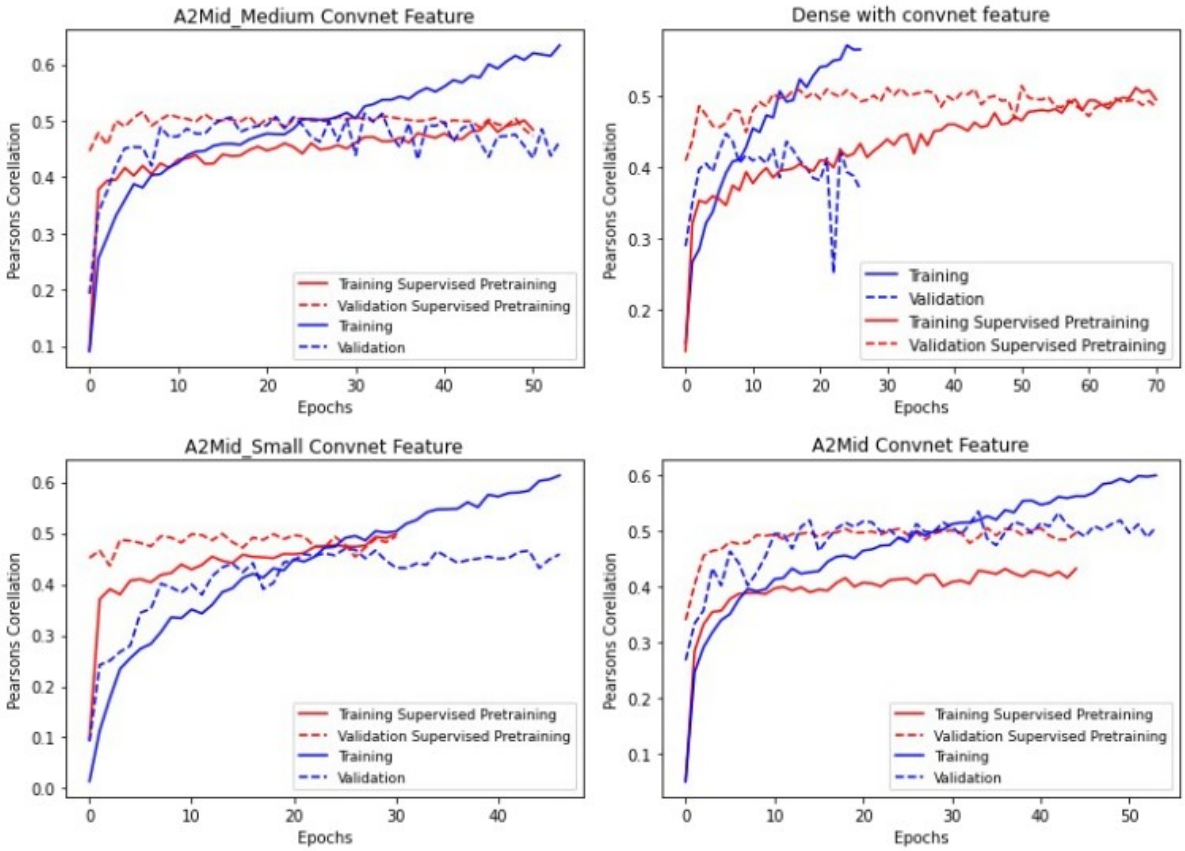
Σχήμα 3.31: Σύγκριση μοντέλων A2Mid_Medium διαφορετικό αριθμό παγωμένων επιπέδων (Large-Pre-training Supervised).

Κάναμε πειράματα και στη μεγάλη Large-Pretraining-dataset. Τα αποτελέσματα είναι καλύτερα σε σύγκριση με τη προ-εκπαίδευση στο μικρό σύνολο δεδομένων.



Σχήμα 3.32: Γραφικές παραστάσεις που αναδεικνύουν την επίδραση του Supervised Pre-training στη μεγάλη dataset.

Ένας διαφορετικός τρόπος προσέγγισης του Supervised-Pretraining ήταν να χρησιμοποιήσουμε και τις ενδιάμεσες εξόδους των επιπέδων και να τις συγκολλήσουμε φτιάχνοντας ένα χαρακτηριστικό το οποίο ονομάζεται convnet feature. Η λογική για την πρακτική αυτήν βρίσκεται στο ότι κάθε επίπεδο εξάγει κάποιο διαφορετικό χαρακτηριστικό του αρχικού μας δεδομένου. Έτσι η αξιοποίηση όλων των χαρακτηριστικών που εξάγουν όλα τα επίπεδα θα δημιουργήσει μια πιο γερή αναπαράστασή του. Προκειμένου να μη βγει μεγάλων διαστάσεων το convnet feature πήραμε το average pool των εξόδων των επιπέδων. Η πρακτική αυτή εφαρμόστηκε από τους Gyorgy Fazekas et al. αποδεικνύοντας με πολλά πειράματα ταξινόμησης της μουσικής την αποτελεσματικότητα του.

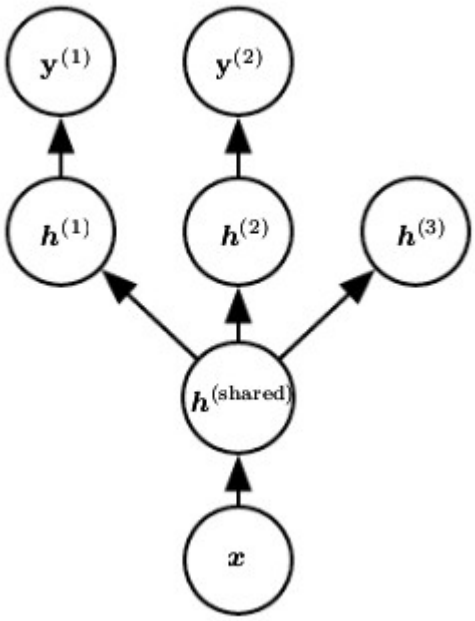


Σχήμα 3.33: Εφαρμογή Supervised Pre-training σε δίαφορα μοντέλα στην Pre-training dataset με χρήση του convnet feature.

Τα αποτελέσματα όπως φαίνεται είναι ικανοποιητικά όσον αφορά τη μείωση του overfitting όσο και στην αύξηση των επιδόσεων. Ενδιαφέρον είναι πως στη περίπτωση αυτή, η τιμή του validation του Pearson Correlation αυξάνεται από νωρίς απότομα.

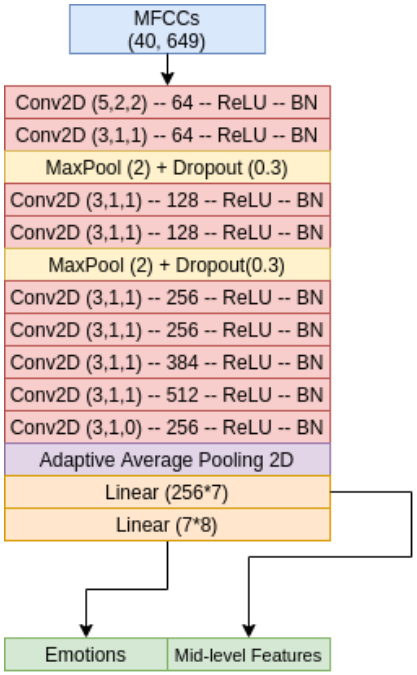
3.6.7 Multitask Learning

Το Multitask learning είναι η τελευταία τεχνική ομαλοποίησης που χρησιμοποιήσαμε. Η τεχνική αυτή βασίζεται στο να μοιράσεις μέρος ενός δικτύου σε δύο προβλήματα. Με αυτό το τρόπο οι τιμές των βαρών του κορμού του δικτύου που μοιράζεται θα αναγκαστούν να πάρουν πιο καλές τιμές πολλές φορές, οδηγώντας σε καλύτερη γενίκευση. Ένα παράδειγμα δίνεται στο Σχήμα 3.34.



Σχήμα 3.34: Παράδειγμα Multi-task Learning. Το x είναι η είσοδος, το $h^{(shared)}$ είναι οι κοινές κρυμμένες μονάδες, τα $h^{(1)}$, $h^{(2)}$, $h^{(3)}$ είναι κρυμμένες μονάδες διαφορετικών προβλημάτων, τα $y^{(1)}$, $y^{(2)}$ είναι έξοδοι διαφορετικών προβλημάτων.

Τα πειράματα που κάναμε βασίστηκαν στο άρθρο που έχουμε ως αναφορά των Verena Haunschmid και Gerhard Widmer και έγιναν στο σύνολο δεδομένων της Aljanaki με αναπαράσταση MFCCs. Το σκεπτικό είναι να κάνουμε πρόβλεψη των ενδιάμεσων χαρακτηριστικών και ταυτόχρονα πρόβλεψη του συναισθήματος και να ανανεώνουμε τα βάρη του μοντέλου σύμφωνα με το άθροισμα των απωλειών των δύο προβλέψεων.



Σχήμα 3.35: Η αρχιτεκτονική του A2Mid2E-Joint.

Mid-level Features	Pearson Correlation
Melody	0.851
Articulation	0.710
Rythm_Complexity	0.566
Rythm_Stability	0.750
Dissonance	0.837
Atonality	0.818
Mode	0.739
Average	0.750

Πίνακας 3.14: Οι τιμές του Pearson Correlation για τα ενδιάμεσα χαρακτηριστικά.

Emotions	Pearson Correlation
Valence	0.772
Energy	0.762
Tension	0.693
Anger	0.708
Fear	0.817
Happy	0.746
Sad	0.685
Tender	0.734
Average	0.730

Πίνακας 3.15: Οι τιμές του Pearson Correlation για τα συναισθήματα.

Τα αποτελέσματα όσον αφορά τις μετρικές ήταν πολύ καλά. Δοκιμάσαμε την ίδια τεχνική και στο σύνολο δεδομένων της Aljanaki με αναπαράσταση spotify χαρακτηριστικών. Σε αυτή την περίπτωση κάναμε πρόβλεψη στις 12 τιμές των spotify χαρακτηριστικών και ύστερα κάναμε πρόβλεψη στα ενδιάμεσα χαρακτηριστικά. Φτιάξαμε το combined loss αθροίζοντας της απώλειες τους και κάναμε με αυτόν τον τρόπο multitask learning. Η μέση τιμή του Pearson correlation ήταν 0.5 για τα 12 στοιχεία και 0.46 για τα ενδιάμεσα χαρακτηριστικά.

Κεφάλαιο 4

Αξιολόγηση Πειραμάτων και Συμπεράσματα

Στο κεφάλαιο αυτό θα μελετήσουμε τα αποτελέσματα των πειραμάτων μας και θα βγάλουμε συμπεράσματα για την αναπαράσταση των τραγουδιών, τις αρχιτεκτονικές των νευρωνικών δικτύων, τις τεχνικές ομαλοποίησης και τη σύνδεση των ενδιάμεσων χαρακτηριστικών με το συναίσθημα. Επίσης θα γίνει σύγκριση των αποτελεσμάτων μας με προηγούμενα προβλήματα.

4.1 Τρόποι Αναπαράστασης της Μουσικής

Όπως έχει προαναφερθεί στα πειράματα μας χρησιμοποιήσαμε δύο τρόπους αναπαράστασης των τραγουδιών. Ο πρώτος ήταν μέσα από τα MFCCs και ο δεύτερος μέσα από τα spotify χαρακτηριστικά. Ο κάθε τρόπος μας έδωσε και διαφορετικά πλεονεκτήματα.

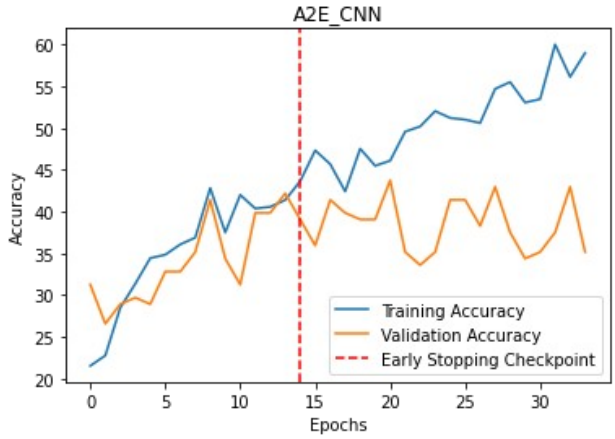
Η αναπαράσταση MFCCs είναι ευρέως γνωστή, για το λόγο ότι είναι αποτελεσματική. Στα δικά μας πειράματα μας έδωσε τα καλύτερα αποτελέσματα. Επίσης έχει το πλεονέκτημα ότι οι τονικότητες που περιγράφει είναι κατανοητές από τον άνθρωπο. Παρόλα αυτά με την αναπαράσταση αυτήν δε μπορούσαμε να εκτελέσουμε πειράματα σε μεγάλου όγκου σύνολα δεδομένων. Ο λόγος είναι ότι ήταν δύσκολο να βρούμε πολλά δεδομένα τραγουδιών. Ακόμα και αν τα καταφέρναμε το μέγεθος τους θα ήταν τεράστιο και δύσκολα διαχειρίσιμο.

Με την αναπαράσταση τραγουδιών με spotify χαρακτηριστικά μπορέσαμε να φτιάξουμε εύκολα, μεγάλα σύνολα δεδομένων, τα οποία τα χρησιμοποιήσαμε στην προ-εκπαίδευση. Ένας λόγος είναι ότι τα δεδομένα του Spotify API είναι πολύ ελαφρά, τα 200.000 τραγούδια είχαν μέγεθος μόλις 4GB. Ένας διαφορετικός λόγος είναι ότι το Spotify API παρέχει 50.000.000 τραγούδια και 5.000 είδη τραγουδιών, δίνοντας μας τη δυνατότητα να συγκροτήσουμε σύνολα δεδομένων έχοντας μόνο τα μετα-δεδομένα τραγουδιών. Παρόλο που τα θετικά που προσφέρει είναι αρκετά, από τα πειράματά μας διαπιστώσαμε ότι δεν προσφέρει καλή αναπαράσταση των τραγουδιών.

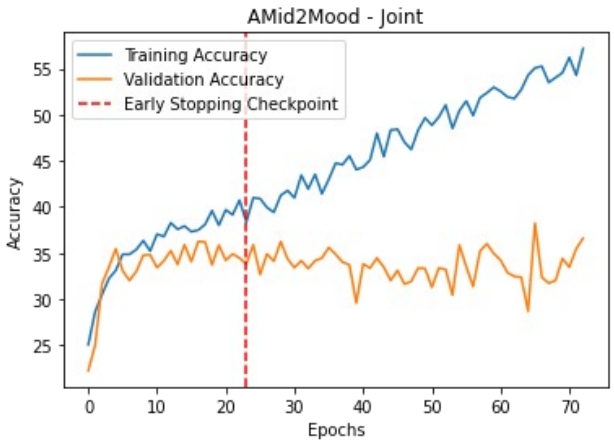
Πέρα από τα πειράματα που κάναμε για τη πρόβλεψη των ενδιάμεσων χαρακτηριστικών, τα υπόλοιπα είχαν χαμηλή επίδοση. Συγκεκριμένα για τη πρόβλεψη των 5 ομάδων του MIREX, πετύχαμε 41% accuracy με απευθείας πρόβλεψη των ομάδων και 32% accuracy με πρόβλεψη των ενδιάμεσων χαρακτηριστικών και ύστερα γραμμικής

σύνδεσης τους με τις 5 ομάδες (Σχήμα 4.1). Επίσης με το σύνολο δεδομένων Mooydlyrics, όπου συγκροτείται από τραγούδια που έχουν ταμπέλες συναισθήματος, τα αποτελέσματα μας κυμαίνονταν στο 35% accuracy (Σχήμα 4.2).

Ένα άλλο πείραμα που δοκιμάσαμε το οποίο ανέδειξε την ανεπάρκεια της αναπαράστασης των spotify χαρακτηριστικών ήταν σχετικά με το remastering. Με τη βοήθεια του Spotify API συγκροτήσαμε ένα σύνολο δεδομένων με 200 τραγούδια, σχετική με το remastering. Τα τραγούδια ήταν κυρίως Rock, Punk και Metal. Τα 100 πρώτα ήταν σε κατάσταση rough mix και τα υπόλοιπα 100 σε κατάσταση remastered. Στόχος αυτού του πειράματος ήταν η δοκιμή του μοντέλου A2Mid_Medium για την παρατήρηση των μεταβολών των ενδιάμεσων χαρακτηριστικών των τραγουδιών που επιφέρει η διαδικασία του remastering. Δυστυχώς λόγω έλλειψης καλής αναπαράστασης, οι τιμές των ενδιάμεσων χαρακτηριστικών των τραγουδιών πριν και μετά τα remastering ήταν παρόμοιες.



Σχήμα 4.1: Το accuracy του CNN μοντέλου πρόβλεψης των 5 ομάδων του MIREX.



Σχήμα 4.2: Το accuracy του CNN μοντέλου πρόβλεψης των 4 ομάδων του Moodylyrics.

4.2 Αρχιτεκτονικές Νευρωνικών Δικτύων

Τις περισσότερες αρχιτεκτονικές νευρωνικών δικτύων τις χρησιμοποιήσαμε στο πρόβλημα της πρόβλεψης ενδιάμεσων χαρακτηριστικών του συνόλου δεδομένων της Aljanaki με αναπαράσταση spotify χαρακτηριστικών. Τα συμπεράσματα που βγάλαμε αφορούν κυρίως τα συνελικτικά δίκτυα και τα πλήρως συνδεδεμένα δίκτυα.

Τα πλήρως συνδεδεμένα δίκτυα είχαν καλά αποτελέσματα αλλά εμφάνιζαν έντονα το φαινόμενο του overfitting. Ο λόγος βρίσκεται στο ότι δεν έχουν τη δυνατότητα να μελετάνε τοπικά χαρακτηριστικά αλλά μελετούν το δεδομένο σαν σύνολο. Το γεγονός αυτό, έδωσε προοπτικές στις τεχνικές ομαλοποίησης να έχουν μεγαλύτερη θετική επίδραση στα μοντέλα, κάτι το οποίο φάνηκε στα πειράματα. Οι τεχνικές ομαλοποίησης στα πλήρως συνδεδεμένα δίκτυα τα οδήγησαν να έχουν αξιόλογα αποτελέσματα.

Όσον αφορά τα συνελικτικά δίκτυα, γενικά εμφάνισαν καλά αποτελέσματα με διαφορετικές επιδόσεις ανάλογες του μεγέθους των δικτύων. Χαρακτηριστικό παράδειγμα είναι πως στα μεγάλα δίκτυα ενώ εμφάνιζαν καλύτερες επιδόσεις εμφάνιζαν επίσης έντονα το φαινόμενο overfitting. Για τα μικρότερα σε μέγεθος δίκτυα, ίσχυε το αντίστροφο. Για τον λόγο αυτόν η επίδραση των τεχνικών ομαλοποίησης φάνηκε κυρίως στα μεγάλα δίκτυα.

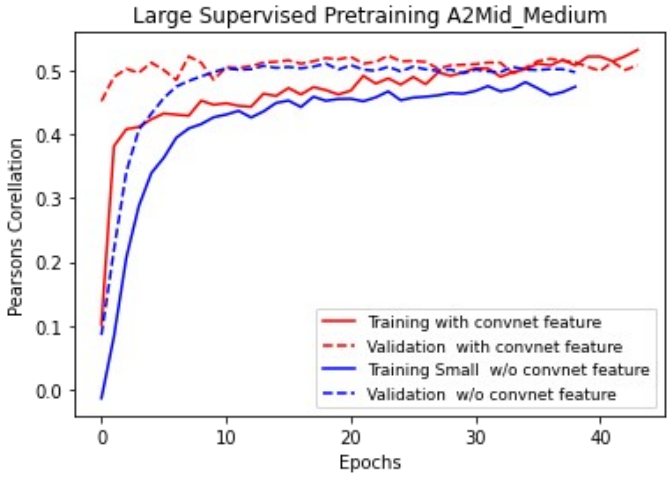
Παρά τις διαφορές που εμφάνισαν τα συνελικτικά και τα πλήρως συνδεδεμένα δίκτυα, με τις τεχνικές ομαλοποίησης καταφέραμε η επίδοση τους να είναι ισάξια καλή, με αυτή του A2Mid_Medium να εμφανίζει τα καλύτερα αποτελέσματα.

Type	Model	Average Pearson Correlation
Dense	Dense_Small	0,503
	Dense_Medium	0,511
	Dense	0,510
CNN	A2Mid_Small	0,504
	A2Mid_Medium	0,517
	A2Mid	0,512
	A2Mid_Large	0,515

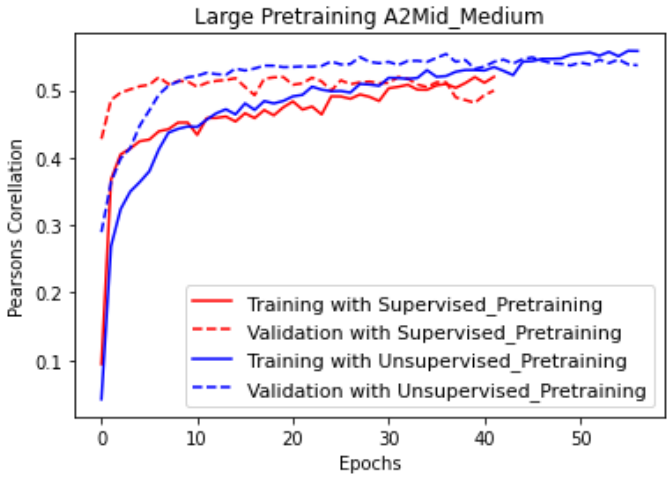
Πίνακας 4.1: Σύγκριση επιδόσεων CNN και ANN μοντέλων.

4.3 Τεχνικές Ομαλοποίησης

Οι τεχνικές ομαλοποίησης που χρησιμοποιήσαμε ήταν χρήσιμες, κάθε μια σε διαφορετικό βαθμό. Όσον αφορά τα dropout επίπεδα, το early stopping και το data augmentation, δεν έχουμε να σχολιάσουμε κάτι πέρα από το ότι βοήθησαν στην αντιμετώπιση του overfitting. Το L2 regularization το χρησιμοποιήσαμε μόνο στις μεγάλες αρχιτεκτονικές με την υπερ-παράμετρο α να έχει μικρή τιμή. Αυτό το κάναμε γιατί η έντονη παρουσία του L2 regularization εμφάνιζε φαινόμενα underfitting. Το Supervised Pretraining αποδείχθηκε ιδιαίτερα χρήσιμο για την αντιμετώπιση του overfitting. Ειδικά η χρήση του convnet χαρακτηριστικού του έδωσε πολύ καλά αποτελέσματα (Σχήμα 4.3). Παρόλα αυτά, η πιο σημαντική τεχνική ομαλοποίησης αποδείχτηκε ότι είναι το Unsupervised-Pretraining (Σχήμα 4.4).

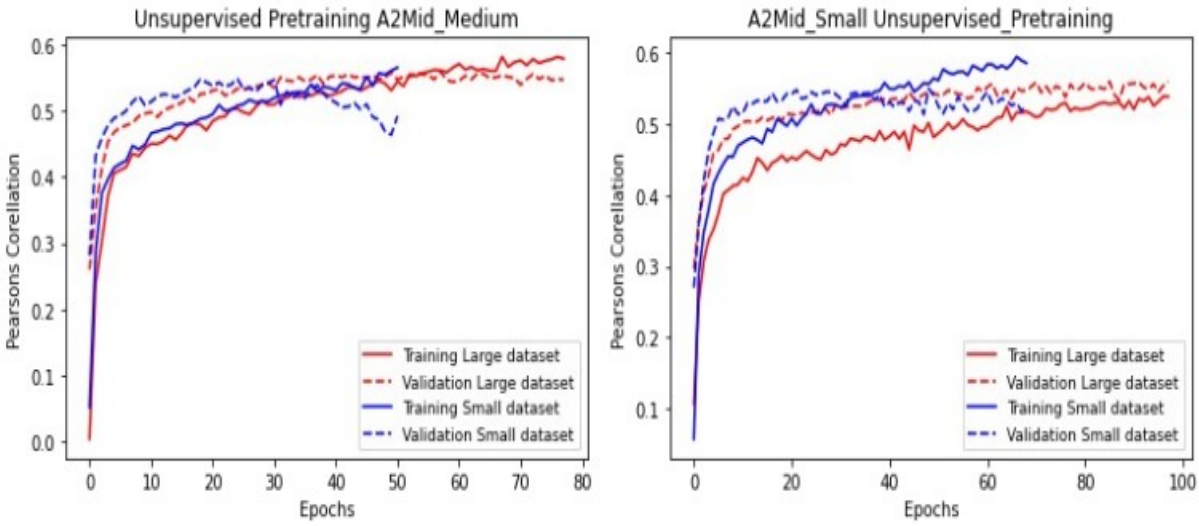


Σχήμα 4.3: A2Mid_Medium που έχει προ-εκπαιδευτεί με Supervised Pre-training με τη χρήση και μη του convnet χαρακτηριστικού.



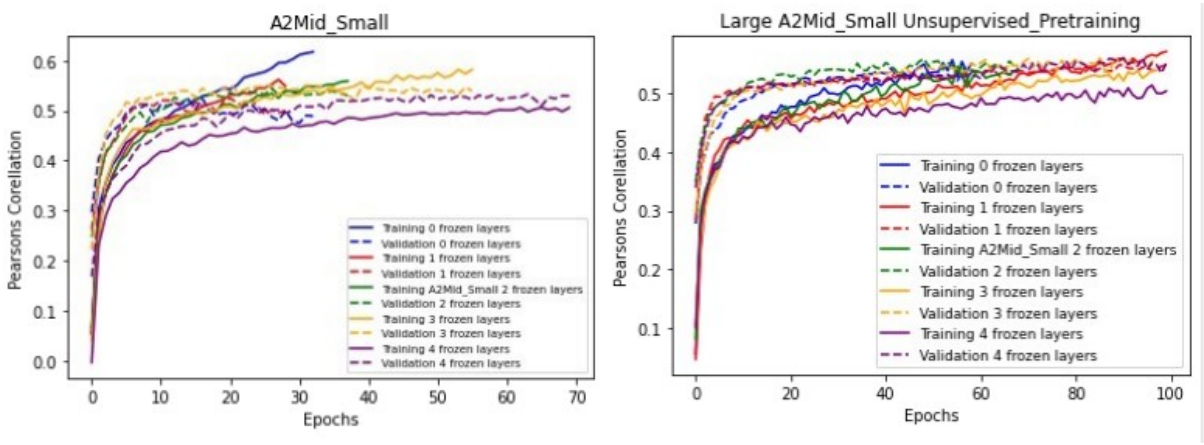
Σχήμα 4.4: A2Mid_Medium που έχει προ-εκπαιδευτεί με Supervised Pre-training και με Unsupervised Pre-training.

Το Unsupervised Pre-training αποτέλεσε τη πιο χρήσιμη τεχνική ομαλοποίησης. Η επίδραση του φάνηκε τόσο στη μείωση του overfitting όσο και στη βελτίωση του Pearson Correlation. Ενδιαφέρον ήταν η διαφορά που υπήρξε ανάλογα με το αν η προ-εκπαίδευση είχε γίνει στο σύνολο δεδομένων Pre-training ή στο σύνολο δεδομένων Large-Pretraining. Όπως φαίνεται και στο Σχήμα 4.5 η προ-εκπαίδευση στο μεγάλο σύνολο δεδομένων βοήθησε περισσότερο σε σχέση με τη προ-εκπαίδευση στο μικρό σύνολο.

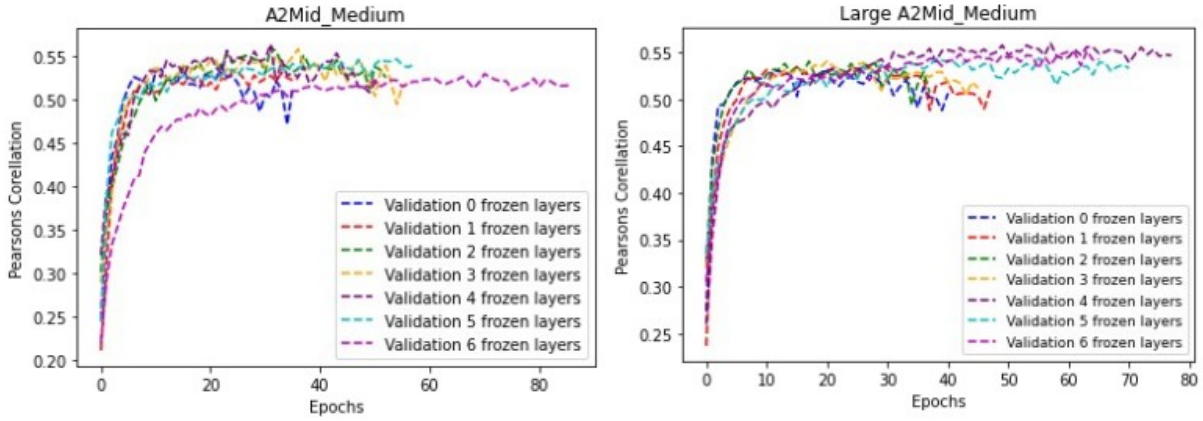


Σχήμα 4.5: Μοντέλα τα οποία έχουν προ-εκπαιδευτεί στη Pre-training dataset και στη Large-Pretraing dataset.

Το Unsupervised Pre-training στο μεγάλο σύνολο δεδομένων φαίνεται να βοήθησε τόσο με την αρχικοποίηση των βαρών όσο και με την εξαγωγή χαρακτηριστικών από τα δεδομένα. Το πρώτο φαίνεται από το πως μη έχοντας “παγώσει” κανένα επίπεδο στην εκπαίδευση του A2Mid_Small, η προ-εκπαίδευση βοήθησε στη μείωση του overfitting (Σχήμα 4.6). Το δεύτερο φαίνεται από το πως έχοντας “παγώσει” όλα τα επίπεδα δεν εμφανίζεται το φαινόμενο του underfitting στην εκπαίδευση του A2Mid_Medium (Σχήμα 4.7). Στις παραπάνω περιπτώσεις όταν η προ-εκπαίδευση γινότανε στο μικρό σύνολο δεδομένων δεν εμφανίζονταν τα φαινόμενα βελτιστοποίησης των μοντέλων.

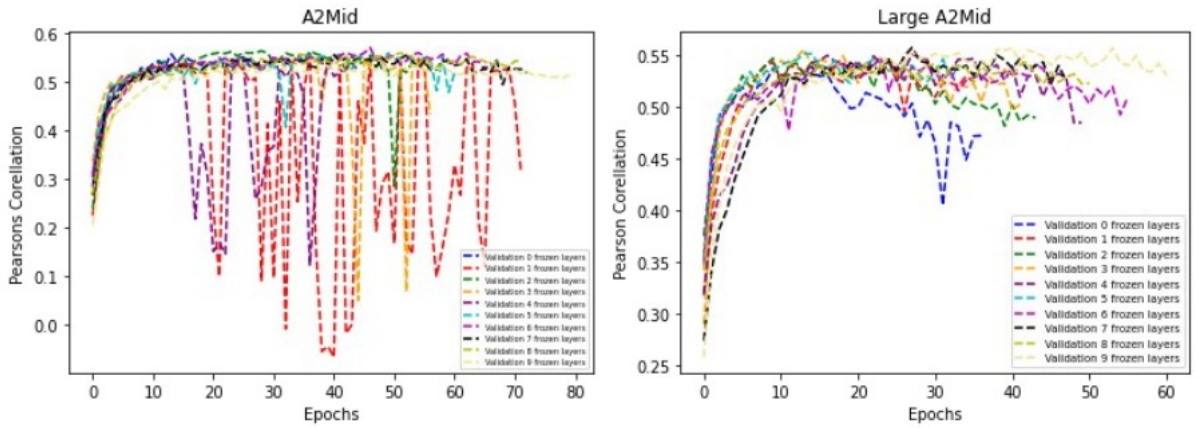


Σχήμα 4.6: Το μοντέλο A2Mid_Small προ-εκπαιδευμένο με τη Pre-training dataset (αριστερά) και με τη Large-Pre-training dataset (δεξιά) με "παγωμένο" διαφορετικό αριθμό επιπέδων κάθε φορά.



Σχήμα 4.7: Το μοντέλο A2Mid_Medium προ-εκπαιδευμένο με τη Pre-training dataset (αριστερά) και με τη Large-Pre-training dataset (δεξιά) με "παγωμένο" διαφορετικό αριθμό επιπέδων κάθε φορά.

Η σημασία του unsupervised pre-training στο μεγάλο σύνολο δεδομένων για την αρχικοποίηση των βαρών φάνηκε και σε άλλα πειράματα. Συγκεκριμένα η προ-εκπαίδευση αρχικοποίησε τα βάρη σε σημείο που δεν μπόρεσαν να αποδράσουν από συγκεκριμένες τιμές, αποφεύγοντας το φαινόμενο overfitting. Με αυτό το τρόπο εξασφαλίστηκε ότι το αποτέλεσμα δεν μπορεί να ξεφύγει. Αυτό φαίνεται στο Σχήμα 4.8 όπου στο A2Mid, όταν η προ-εκπαίδευση του είχε γίνει στο μικρό σύνολο δεδομένων, εμφάνισε φαινόμενα overfitting ενώ όταν η προ-εκπαίδευση είχε γίνει στο μεγάλο, δεν εμφάνισε.



Σχήμα 4.8: Το μοντέλο A2Mid προ-εκπαιδευμένο με τη Pre-training dataset (αριστερά) και με τη Large-Pre-training dataset (δεξιά) με "παγωμένο" διαφορετικό αριθμό επιπέδων κάθε φορά.

4.4 Σύγκριση με παλαιότερες εργασίες

Τα πειράματα μας βασίστηκαν στα πειράματα που έχουν κάνει οι Aljanki et al. και οι Haunschmid et al. στην πρόβλεψη των ενδιάμεσων χαρακτηριστικών στο σύνολο δεδομένων της Aljanaki και στην πρόβλεψη συναισθήματος στο σύνολο δεδομένων, Soundtracks. Τα αποτελέσματα μας ήταν παρόμοια με αυτών. Τις καλύτερες επιδόσεις τις είχαν τα μοντέλα που εφαρμόσαμε σε αναπαράσταση τραγουδιών MFCCs. Τα μοντέλα που εφαρμόσαμε σε αναπαράσταση τραγουδιών με spotify χαρακτηριστικά έδωσαν ικανοποιητικά αποτελέσματα δεδομένου του ότι το σύνολο δεδομένων που γίνανε τα πειράματα είχε το μισό μέγεθος του αρχικού. Λόγω κακής επίδοσης των μοντέλων όσον αφορά την πρόβλεψη συναισθήματος με αναπαράσταση spotify χαρακτηριστικών, δεν κάναμε την αντιστοιχία για αυτά με το συναίσθημα. Παρακάτω δίνονται δύο πίνακες που συγκρίνουν ορισμένα μοντέλα μας με παλαιότερες δουλειές.

Emotions	Haunschmid	A2Mid2E-Joint
Valence	0.82	0.77
Energy	0.78	0.76
Tension	0.82	0.69
Anger	0.76	0.70
Fear	0.79	0.81
Happy	0.65	0.74
Sad	0.64	0.68
Tender	0.72	0.73

Πίνακας 4.2: Οι τιμές του Pearson Correlation για τα συναισθήματα.

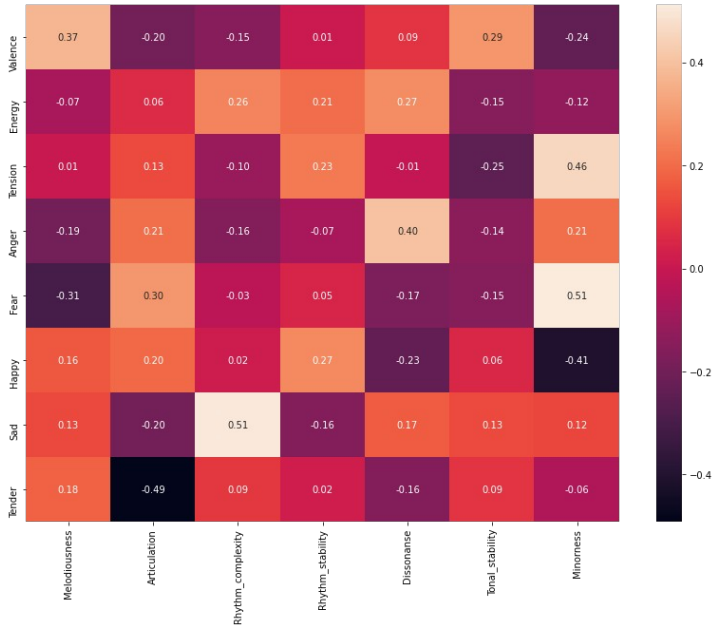
Mid-level feature	Aljanaki	Haunschmid	A2Mid2E-Joint	A2Mid_Medium
Melodiousness	0.70	0.70	0.85	0.58
Articulation	0.76	0.83	0.71	0.76
R.Stability	0.46	0.39	0.56	0.32
R.Complexity	0.59	0.66	0.75	0.54
Dissonance	0.74	0.74	0.83	0.61
Tonal Stability	0.45	0.56	0.81	0.36
Minorness	0.48	0.55	0.73	0.42

Πίνακας 4.3: Οι τιμές του Pearson Correlation για τα ενδιάμεσα χαρακτηριστικά.

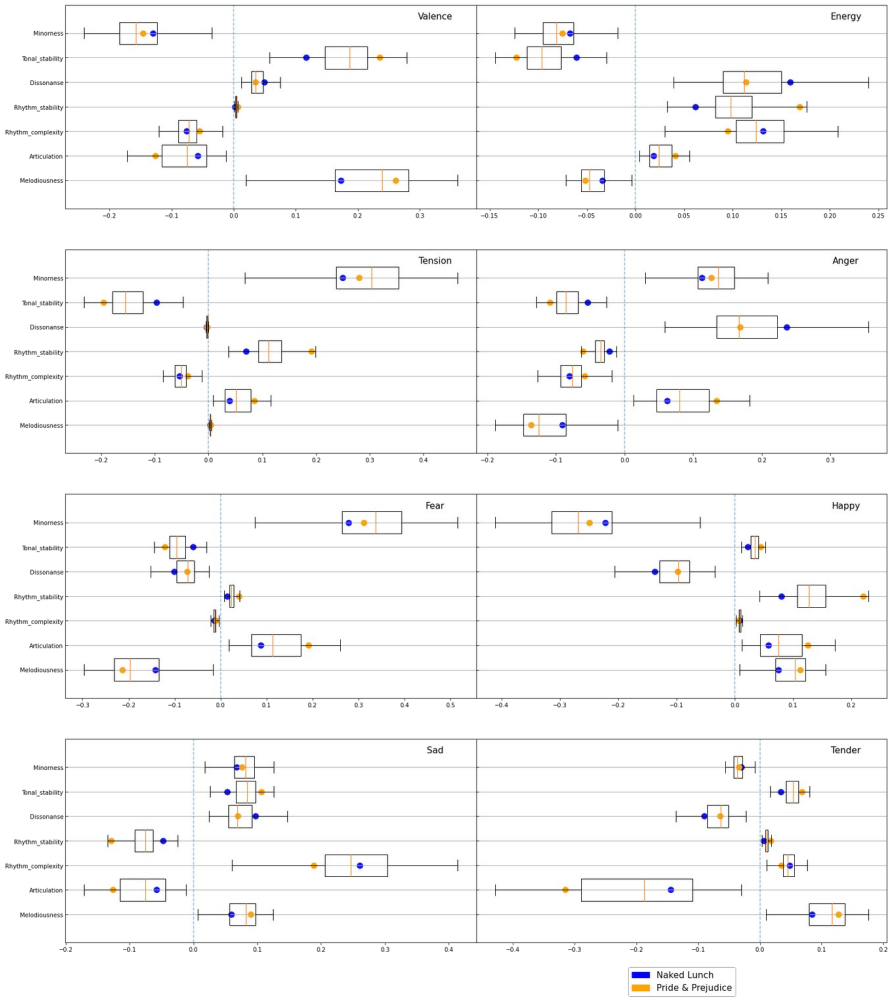
4.5 Συναισθήματα και Ενδιάμεσα Χαρακτηριστικά

Με το τελευταίο επίπεδο του μοντέλου A2Mid2E-Joint έχουμε γραμμική σύνδεση των ενδιάμεσων χαρακτηριστικών με τα συναισθήματα. Η σχέση των δύο μπορεί να αποτυπωθεί από το μέγεθος των βαρών καθώς και από μια τιμή που ονομάζεται effect (Christoph Molnar). Το effect στην προκειμένη περίπτωση είναι το γινόμενο των βαρών με τα ενδιάμεσα χαρακτηριστικά. Στο Σχήμα 4.9 απεικονίζεται η πραγματική σχέση των ενδιάμεσων χαρακτηριστικών με το συναίσθημα. Στο Σχήμα 4.10 απεικονίζονται τα βάρη του μοντέλου A2Mid2E-Joint. Στο Σχήμα 4.11 απεικονίζεται το effect του μοντέλου A2Mid2E-Joint στο σύνολο δεδομένων, Soundtracks.

Σχήμα 4.9: Correlation matrix των πραγματικών τιμών των ενδιάμεσων χαρακτηριστικών με τα συναισθήματα.



Σχήμα 4.10: Correlation matrix των βαρών του μοντέλου A2Mid2E-Joint.



Σχήμα 4.11: Boxplot του effect του μοντέλου A2Mid2E-Joint.

Στα παραπάνω σχήματα φαίνεται η σύνδεση των ενδιάμεσων χαρακτηριστικών με το συναίσθημα. Σε ορισμένες περιπτώσεις υπάρχει συμφωνία μεταξύ των βαρών του μοντέλου (Σχήμα 4.10) και της σχέσης των πραγματικών τιμών (Σχήμα 4.9), για παράδειγμα για το Valence με το Melodiousness ή για το Energy με το Minorness. Παρόλα αυτά σε μερικές περιπτώσεις υπάρχει διαφωνία, για παράδειγμα για τη σχέση του Tension και του Dissonanse. Αυτό οφείλεται στο γεγονός ότι τα βάρη του μοντέλου μελετούν τις τιμές ως σύνολο σε αντίθεση με τη συσχέτιση των πραγματικών τιμών όπου μελετούν τις τιμές ως μονάδα. Σε γενικές γραμμές παρατηρήσαμε συμφωνία της σύνδεσης των ενδιάμεσων χαρακτηριστικών με τα συναισθήματα στους διάφορους τρόπους σύνδεσής τους.

Μελετήσαμε επίσης τη σύνδεση συναισθήματος και ενδιάμεσων χαρακτηριστικών σε δύο τραγούδια. Το ένα προέρχεται από την ταινία Naked Lunch και το άλλο από την ταινία Pride and Prejudice. Το πρώτο τραγούδι εκτελείται από έναν σαξοφωνίστα και το δεύτερο από ορχήστρα. Στα τραγούδια αυτά ενώ έχουν κοινό συναίσθημα που κυριαρχεί, το energy, δεν έχουν κοινά ενδιάμεσα χαρακτηριστικά που το προκαλούν. Το energy του σαξοφωνίστα οφείλεται στο dissonanse και στο rythm_complexity ενώ το energy της ορχήστρας οφείλεται στο rythm_stability και στο articulation. Κάτι το οποίο είναι λογικό αν σκεφτεί κανείς ότι ο σαξοφωνίστας παίζει τζαζ μουσική ενώ η ορχήστρα κλασσική μουσική.

Κεφάλαιο 5

Επίλογος

Στο κεφάλαιο αυτό θα συνοψίσουμε το τι κάναμε σε αυτή τη διπλωματική. Θα περιγράψουμε σύντομα τα σύνολα δεδομένων που χρησιμοποιήσαμε, τις αρχιτεκτονικές που διαλέξαμε, τις τεχνικές ομαλοποίησης που δοκιμάσαμε και τα συμπεράσματα τα οποία βγάλαμε. Επίσης θα περιγράψουμε πιθανές μελλοντικές εργασίες στην κατεύθυνση που όρισε η διπλωματική αυτή.

5.1 Σύνοψη

Ο στόχος της διπλωματικής αυτής ήταν να μελετήσει τη σύνδεση των ενδιάμεσων χαρακτηριστικών με το συναίσθημα, να δοκιμάσει νέες αναπαραστάσεις των τραγουδιών μέσω του Spotify API και να μελετήσει ορισμένες τεχνικές ομαλοποίησης.

Τα πειράματα μας χωρίστηκαν σε δύο σκέλη. Το πρώτο με πειράματα CNN μοντέλων στο σύνολο δεδομένων της Aljanaki με αναπαράσταση MFCCs για τη μελέτη της σύνδεσης ενδιάμεσων χαρακτηριστικών και συναισθημάτων. Το δεύτερο με πειράματα σε πληθώρα αρχιτεκτονικών νευρωνικών δικτύων στο σύνολο δεδομένων της Aljanaki με αναπαράστασή spotify χαρακτηριστικών για τη μελέτη των ενδιάμεσων χαρακτηριστικών και τη γενίκευση των μοντέλων.

Τα αποτελέσματα μας όσον αφορά το πρώτο σκέλος ήταν παρόμοια με τα αποτελέσματα προηγούμενων εργασιών. Ανέδειξαν τη σύνδεση συναισθημάτων με τα ενδιάμεσα χαρακτηριστικά μέσω της μελέτης των βαρών του γραμμικού επιπέδου που τα συνδέει. Τα αποτελέσματα όσον αφορά το δεύτερο σκέλος, ανέδειξαν τη σημασία του Representation Learning ως τεχνική ομαλοποίησης και βελτιστοποίησης των μοντέλων και τις δυνατότητες που παρέχει σε προγραμματιστές το Spotify API. Σε όλα τα παραπάνω πειράματα μελετήθηκε η σύνδεση των παραμέτρων με τις υπερ-παραμέτρους και έγιναν πειράματα για την αυτοματοποιημένη ρύθμιση τους, χρησιμοποιώντας σύγχρονους αλγορίθμους, όπως ο ASHA.

Τα συμπεράσματα από τη διπλωματική αυτή σε σχέση με το Spotify API ήταν ότι ενώ προσφέρει ευκολία στη συγκρότηση οποιουδήποτε συνόλου δεδομένων, πιθανώς δεν προσφέρει επαρκώς καλή αναπαράσταση των τραγουδιών. Έγιναν διάφορα πειράματα που το ανέδειξαν αυτό. Κάποια παραδείγματα είναι η ταξινόμηση του συνόλου δεδομένων Moodylyrics, η ταξινόμηση του συνόλου δεδομένων MIREX-

multimodal και η επίδραση του remastering στα ενδιάμεσα χαρακτηριστικά των τραγουδιών.

5.2 Μελλοντικές εργασίες

Οι μελλοντικές εργασίες βάσει της διπλωματικής αυτής, θα μπορούσαν να πάρουν διαφορετικές κατευθύνσεις. Αρχικά θα είχε ενδιαφέρον η μελέτη της επίδρασης του remastering στα ενδιάμεσα χαρακτηριστικά των τραγουδιών. Απαραίτητη προϋπόθεση για να γίνει αυτό είναι η συγκρότηση ενός συνόλου δεδομένων με τραγούδα πριν και μετά το remastering, αλλά με όσο το δυνατόν καλύτερη αναπαράσταση.

Ένας διαφορετικός δρόμος θα μπορούσε να ήταν η μελέτη των spotify χαρακτηριστικών και της επάρκειας στην αναπαράσταση των τραγουδιών που δίνουν. Αυτό σημαίνει πειράματα που θα μελετούν την επίδοση μοντέλων εκπαιδευμένων με spotify χαρακτηριστικά για διάφορα προβλήματα ταξινόμησης. Είναι σημαντικό να γίνει αυτό γιατί το spotify παρέχει άνεση και ποικιλία όσον αφορά τη συγκρότηση συνόλου δεδομένων, αλλά πιθανόν δεν παρέχει καλή αναπαράσταση τους.

Μια διαφορετική μελλοντική κατεύθυνση θα μπορούσε να ήταν η εκτενέστερη μελέτη των τεχνικών ομαλοποίησης στα μοντέλα με χρήση μεθόδων προ-εκπαίδευσης σε μεγαλύτερα και διαφορετικά σύνολα δεδομένων.

Τέλος, μια κατεύθυνση πιθανώς χρήσιμη, θα μπορούσε να ήταν η σύνδεση των ενδιάμεσων χαρακτηριστικών που μελετήσαμε στη διπλωματική αυτή με Συστήματα Προτάσεων Μουσικής.

Βιβλιογραφία

- Anna Aljanaki ,Mohammad Soleymani. “A DATA-DRIVEN APPROACH TO MID-LEVEL PERCEPTUAL MUSICAL FEATURE MODELING”. 2018.

- Erion Çano, Maurizio Morisio. “MoodyLyrics: A Sentiment Annotated Lyrics Dataset”. 2017.

- Keunwoo Choi, György Fazekas, Mark Sandler, Kyunghyun Cho. “TRANSFER LEARNING FOR MUSIC CLASSIFICATION AND REGRESSION TASKS”. 2017.

- Shreyan Chowdhury, Andreu Vall, Verena Haunschmid, Gerhard Widmer. “TOWARDS EXPLAINABLE MUSIC EMOTION RECOGNITION: THE ROUTE VIA MID-LEVEL FEATURES”. 2019.

- Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, Manuel Moussallam. “MUSIC MOOD DETECTION BASED ON AUDIO AND LYRICS WITH DEEP NEURAL NET”. 2018.

-Dinh, L., Pascanu, R., Bengio, S. & Bengio, Y. “Sharp Minima Can Generalize For Deep Nets”. 2017.

- Tuomas Eerola and Jonna K. Vuoskoski. “A comparison of the discrete and dimensional models of emotion in music”. 2011.

- Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent. “Why Does Unsupervised Pre-training Help Deep Learning?”. 2010.

- “Every Noise At Once”. http://everynoise.com/.

- Anders Friberg, Erwin Schoonderwaldt, Anton Hedblad, Marco Fabiani, and Anders Elowsson. “Using listener-based perceptual features as intermediate representations in music information retrieval”. 2014.

- A Gabrielsson, E Lindström.“The Role of Structure in the Musical Expression of Emotions”. 2010.

- Ian Goodfellow Yoshua Bengio Aaron Courville. “Deep Learning”, 2016.

- Verena Haunschmid, Ethan Manilow, Gerhard Widmer. ”TOWARDS MUSICALLY MEANINGFUL EXPLANATIONS USING SOURCE SEPARATION”. ISMIR. 2020.

- Verena Haunschmid, Shreyan Chowdhury, Gerhard Widmer. “Two-level Explanations in Music Emotion Recognition”. 2019.

- Simon Haykin. ”Νευρωνικά Δίκτυα και Μηχανική Μάθηση”. 2009.

- Kate Hevner. “EXPERIMENTAL STUDIES OF THE ELEMENTS OF EXPRESSION IN MUSIC”. The American Journal of Psychology. 1936.

- Sepp Hochreiter, Fakultät für Informatik, Jürgen Schmidhube. “Flat Minima”. 1997.

- Xiao Hu, J. Stephen Downie. “EXPLORING MOOD METADATA: RELATIONSHIPS WITH GENRE, ARTIST AND USAGE METADATA”. MIREX. 2007.

- Xiao Hu , J. Stephen Downie, Cyril Laurier, Mert Bay, Andreas F. Ehmann. “THE 2007 MIREX AUDIO MOOD CLASSIFICATION TASK: LESSONS LEARNED”. MIREX. 2008.

- Tristan Jehan, David DesRoches. “Analyzer Documentation The Echonest”. 2011.

- M. I. Jordan, T. M. Mitchell. “ Machine learning: Trends, perspectives, and prospects “, 2015.

- Youngmoo E. Kim, Erik M. Schmidt, Raymond Migneco, Brandon G. Morton Patrick Richardson, Jeffrey Scott, Jacquelin A. Speck, and Douglas Turnbull. “MUSIC EMOTION RECOGNITION: A STATE OF THE ART REVIEW”. ISMIR. 2010.

- Diederik P. Kingma, Jimmy Lei Ba. “ADAM : A METHOD FOR STOCHASTIC OPTIMIZATION”. ICLR. 2015.

- Cyril Laurier, Perfecto Herrera. “AUDIO MUSIC MOOD CLASSIFICATION USING SUPPORT VECTOR MACHINE”. MIREX. 2007.

- Cyril Laurier, Jens Grivolla, Perfecto Herrera. “Multimodal Music Mood Classification using Audio and Lyrics”. 2008.

- Cyril Laurier. “Automatic Classification of Musical Mood by Content-Based Analysis”. 2011.

- Yann LeCun, Leon Bottou, Geneviene B. Orr, Klaus-Rober Muller. “Efficient BackProp”, 1998.

- Y LeCun, Y Bengio.“Convolutional networks for images, speech, and time series. The Handbook of Brain Theory and Neural Networks”. Cambridge, MA: MIT Press. 2003.

- Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-Tzur, Moritz Hardt, Benjamin Recht, Ameet Talwalkar. “A SYSTEM FOR MASSIVELY PARALLEL HYPERPARAMETER TUNING”. 2020.

- Tong Liu, Li Han, Liangkai Ma, and Dongwei Guo. “Audio-based deep music emotion recognition”, CDMMS, 2018.

- Christoph Molnar. “Interpretable Machine Learning”. 2019. https://christophm.github.io/interpretable-ml-book/.

- Noam Mor, Lior Wolf, Adam Polyak, Yaniv Taigman. “A UNIVERSAL MUSIC TRANSLATION NETWORK” . ICLR. 2019.

- Robert Neumayer and Andreas Rauber. “Integration of Text and Audio Features for Genre Classification in Music Information Retrieval” , 2007.

- Aäron van den Oord, Sander Dieleman, Benjamin Schrauwen. “TRANSFER LEARNING BY SUPERVISED PRE-TRAINING FOR AUDIO-BASED MUSIC CLASSIFICATION”. ISMIR. 2014.

- R. Panda, R. Malheiro, B. Rocha, A. Oliveira, R. P. Paiva, “Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis”. MIREX. 2013.

- Jordi Pons and Xavier Serra. “RANDOMLY WEIGHTED CNNs FOR (MUSIC) AUDIO CLASSIFICATION”. ICASSP. 2019.

- “Pytorch”. https://github.com/pytorch/. 2018.

- Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. ““Why Should I Trust You?” Explaining the Predictions of Any Classifier”. 2016.

- Russell, J. A. ”A circumplex model of affect. *Journal of Personality and Social Psychology”. 1980. 39*(6), 1161–1178. https://doi.org/10.1037/h0077714

- Stuart Russel, Peter Norvig. “Τεχνητή Νοημοσύνη Μια σύγχρονη προσέγγιση” , 2005.

- Alexander Schindler. “Parallel Convolutional Neural Networks for Music Genre and Mood Classification”. 2016.

- “Spotify API”. https://developer.spotify.com/documentation/web-api/.

- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. 2014.

- LACE WEDIN. “A MULTIDIMENSIONAL STUDY OF PERCEPTUAL - EMOTIONAL QUALITIES IN MUSIC”. 1976.

- ΠΥΡΟΒΟΛΑΚΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ. “Αναγνώριση συναισθήματος με ανάλυση στίχων και ηχητικού σήματος μουσικής βασισμένη σε αρχιτεκτονικές βαθιάς μηχανικής μάθησης”. Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο. 2020.