Εθνικο Μετσοβειο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Διατμηματικο Προγραμμα Μεταπτυχιακων Σπουδων
Επιστημη Δεδομενων και Μηχανικη Μαθηση

Αναγνωριση Ειδους και Εξαγωγη Συναισθηματος απο Μουσικη

Η τρίτη εργαστηριακή αναφορά στα πλαίσια του υποχρεωτικού μαθήματος Αναγνωριση Προτυπων του Διατμηματικού Προγράμματος Μεταπτυχιακών Σπουδών Επιστημη Δεδομενων & Μηχανικη Μαθηση.

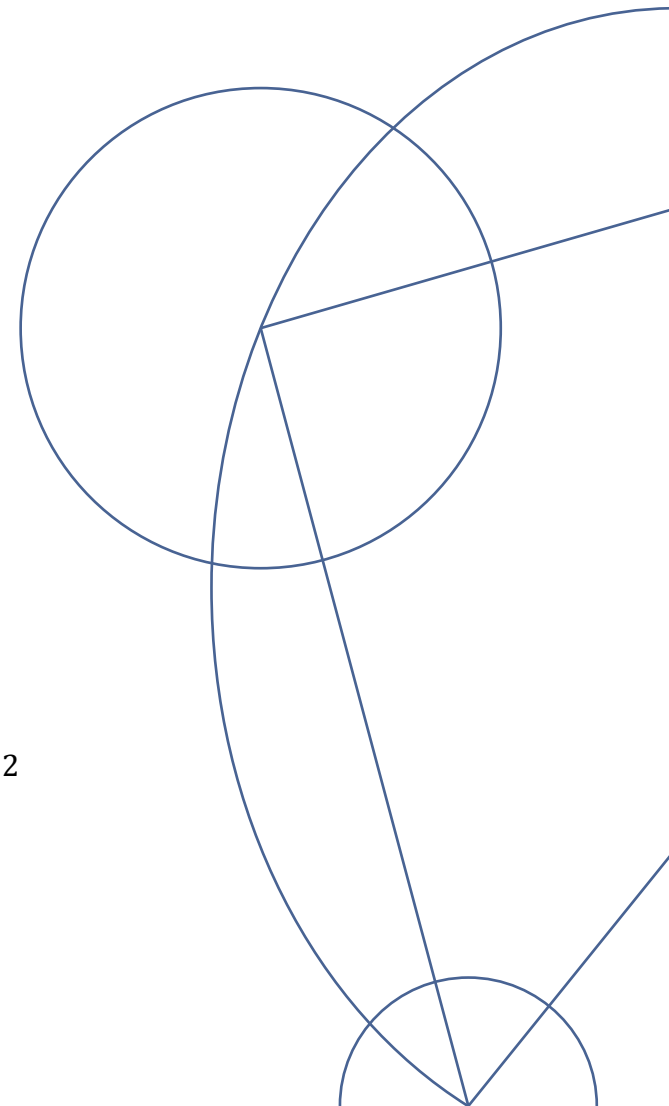
Διδάσκων

Αν. Καθ. Α. Ποταμιανος

Συγγραφή

Νικολαος Σταματης (03400115)
nikolaosstamatis@mail.ntua.gr

Σπυριδων Ρηγας (03400154)
spiridonrigas@mail.ntua.gr



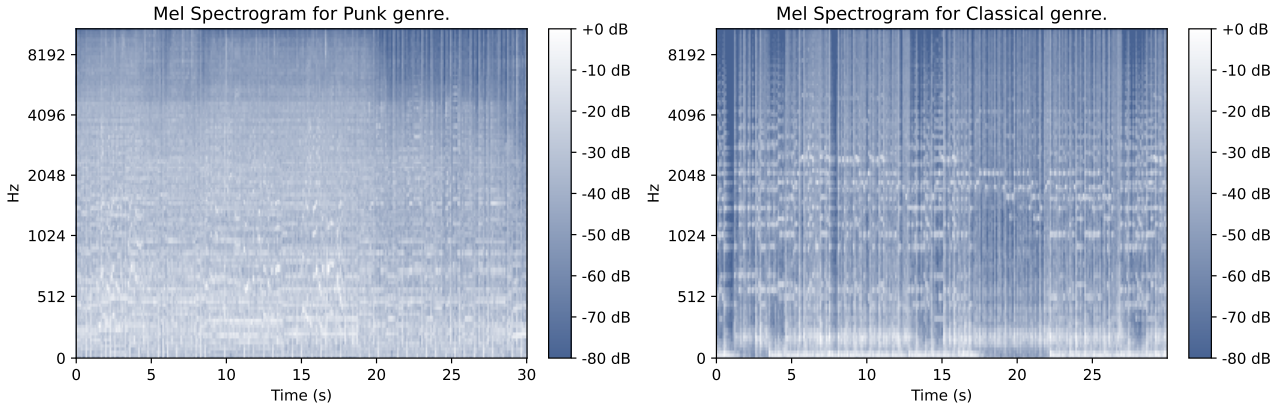
3 Φεβρουαρίου 2022

1 Εισαγωγη

Το περιεχόμενο αυτής της ενότητας αφορά τα Βήματα 1-4 της εργασίας.

Το περιεχόμενο της τρίτης εργαστηριακής άσκησης αποτέλεσε η υλοποίηση συστημάτων επεξεργασίας και αναγνώρισης ήχου, με εφαρμογή σε αναγνώριση είδους (πρώτο μέρος) και εξαγωγή συναισθηματικών διαστάσεων (δεύτερο μέρος) από αποσπάσματα μουσικών κομματιών. Για το πρώτο μέρος της άσκησης αξιοποιήθηκαν 3834 δείγματα του συνόλου Free Music Archive (FMA), χωρισμένα σε 20 κατηγορίες - είδη μουσικής, με σκοπό την εκπαίδευση και αξιολόγηση ταξινομητών. Σε ό,τι αφορά το δεύτερο μέρος της άσκησης, χρησιμοποιήθηκαν 1497 δείγματα με επισημειώσεις των τιμών των συναισθηματικών διαστάσεων valence, energy και danceability για την εκπαίδευση και αξιολόγηση μοντέλων παλινδρόμησης. Σε κάθε περίπτωση, τα δείγματα αποτέλεσαν φασματογραφήματα, τα οποία εξήχθησαν από αποσπάσματα 30 δευτερολέπτων από διαφορετικά τραγούδια.

Για το πρώτο μέρος της άσκησης, αφότου τα δεδομένα μεταφορτώθηκαν, δύο από αυτά επιλέχθηκαν τυχαία (Βήμα 1) προκειμένου να απεικονιστούν τα αντίστοιχα φασματογραφήματα. Το τυχαία επιλεγμένα δεδομένα έχουν κωδικούς 78068 και 47055 και αντιστοιχούν σε μουσική είδους Punk και κλασική μουσική, αντίστοιχα, ενώ τα σχετικά φασματογραφήματα σε κλίμακα mel (mel-spectrograms) φαίνονται στην Εικόνα 1.1.

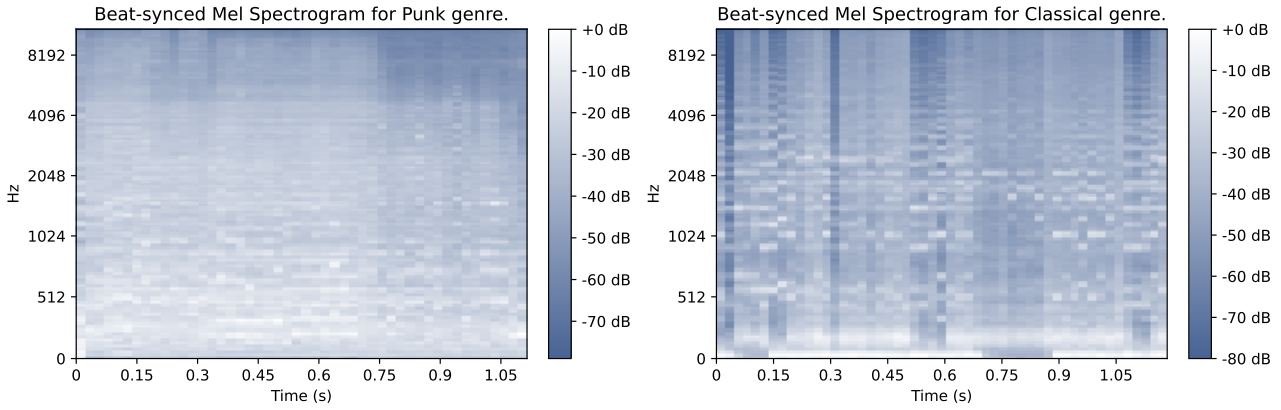


Εικόνα 1.1: Απεικόνιση των mel-φασματογραφημάτων για τα τυχαία επιλεγμένα δεδομένα.

Ένα mel-φασματογράφημα αποτελεί μια τρισδιάσταση αναπαράσταση, όπου ο οριζόντιος άξονας αντιστοιχεί στο χρόνο, ο κατακόρυφος άξονας στη συχνότητα (μετρούμενη σε λογαριθμική κλίμακα mel, αιτιολογώντας έτσι την ονομασία), ενώ ο χρωματικός κώδικας αντιστοιχεί στο πλάτος μιας δεδομένης συχνότητας σε μια δεδομένη χρονική στιγμή, μετρούμενο σε κλίμακα decibel. Τα mel-φασματογραφήματα επιτρέπουν το διαχωρισμό διαφορετικών ήχων ίδιας διάρκειας και συχνότητας, μέσω αναγνώρισης των μοτίβων που σχηματίζονται στην κλίμακα decibel, δηλαδή παρατηρώντας το χρωματικό κώδικα που απεικονίζεται σε σχήματα όπως αυτά της Εικόνας 1.1. Στο παράδειγμα αυτό φαίνεται, για παράδειγμα, πως ενώ τα decibel στην περίπτωση του Punk μουσικού κομματιού πολύ σπάνια πέφτουν κάτω από την τιμή -70, το ίδιο δε μπορεί να ειπωθεί και για το κομμάτι κλασικής μουσικής.

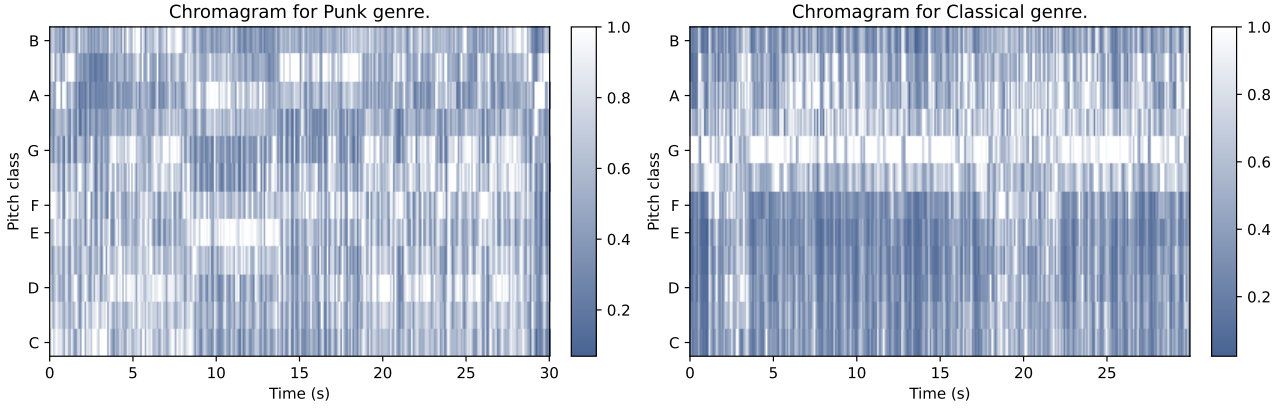
Όπως αναφέρθηκε παραπάνω και αποτυπώνεται στα σχήματα της Εικόνας 1.1, η διάρκεια των μουσικών κομματιών των δεδομένων είναι 30 δευτερόλεπτα, το οποίο σημαίνει πως τα αντίστοιχα φασματογραφήματα συνιστούν ακολουθίες εκατοντάδων time-steps (στην προκείμενη περίπτωση το δείγμα με κωδικό 78068 αποτελείται από 1293 time-steps, ενώ το δείγμα με

κωδικό 47055 αποτελείται από 1291 time-steps). Το υψηλό αυτό μήκος ακολουθιών δημιουργεί την ανάγκη για υψηλούς χρόνους εκπαίδευσης μοντέλων (π.χ. LSTM), αφού η διαδικασία περιλαμβάνει έναν υψηλό αριθμό βαρών τα οποία πρέπει να ληφθούν υπ' όψιν (και να ανανεώνονται συνεχώς μέσω οπισθοδιάδοσης). Ένας τρόπος αντιμετώπισης του συγκεκριμένου ζητήματος (Βήμα 2) ήταν ο συγχρονισμός των φασματογραφημάτων επάνω στο ρυθμό της μουσικής, παίρνοντας μόνο τη διάμεσο ανάμεσα στα σημεία όπου αυτός «χτυπάει». Τα αντίστοιχα mel-φασματογραφήματα χαρακτηρίζονται ως «beat-synced» και απεικονίζονται στα σχήματα της Εικόνας 1.2 για την περίπτωση των δεδομένων με κωδικούς 78068 και 47055.



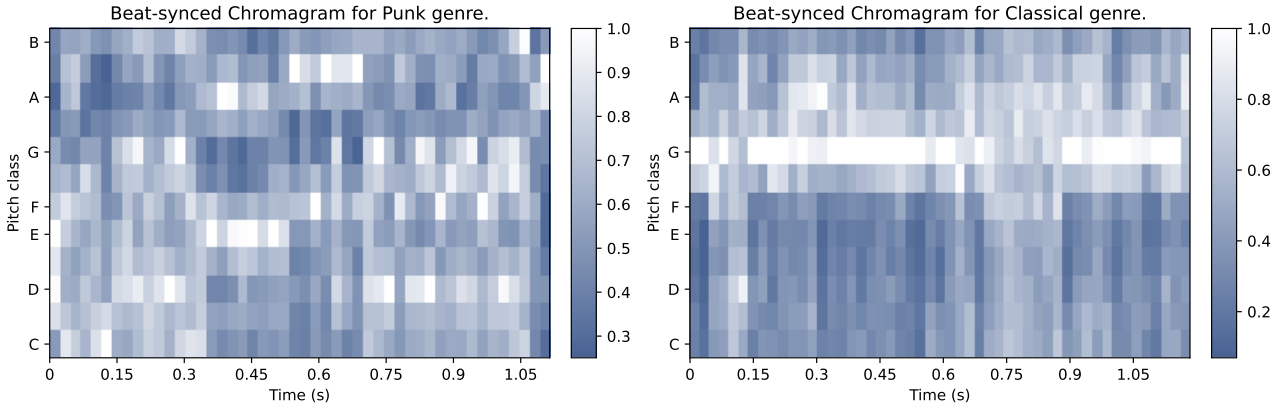
Εικόνα 1.2: Απεικόνιση των beat-synced mel-φασματογραφημάτων για τα τυχαία επιλεγμένα δεδομένα.

Πέρα από mel-φασματογραφήματα, για το σύνολο των δεδομένων εξήχθησαν (Βήμα 3) και τα λεγόμενα χρωμογραφήματα (chromagrams), τα οποία αποτελούν μια ειδική περίπτωση φασματογραφημάτων όπου το πλάτος (χρωματικός κώδικας) υπολογίζεται για ζώνες συχνοτήτων (κατακόρυφος άξονας) που αντιστοιχούν στα 12 ημιτόνια μιας οκτάβας της κλασικής μουσικής. Τα χρωμογραφήματα συνιστούν ένα χρήσιμο εργαλείο για μουσική ανάλυση των αρμονικών και μελοδικών χαρακτηριστικών ενός κομματιού, καθώς και στην αναγνώριση των αλλαγών του ηχοχρώματος και των οργάνων που υπεισέρχονται σε αυτό. Τα σχήματα για τα χρωμογραφήματα και τις beat-synced εκδοχές τους για τα ίδια δεδομένα απεικονίζονται στις Εικόνες 1.3 και 1.4, αντίστοιχα.



Εικόνα 1.3: Απεικόνιση των χρωμογραφημάτων για τα τυχαία επιλεγμένα δεδομένα.

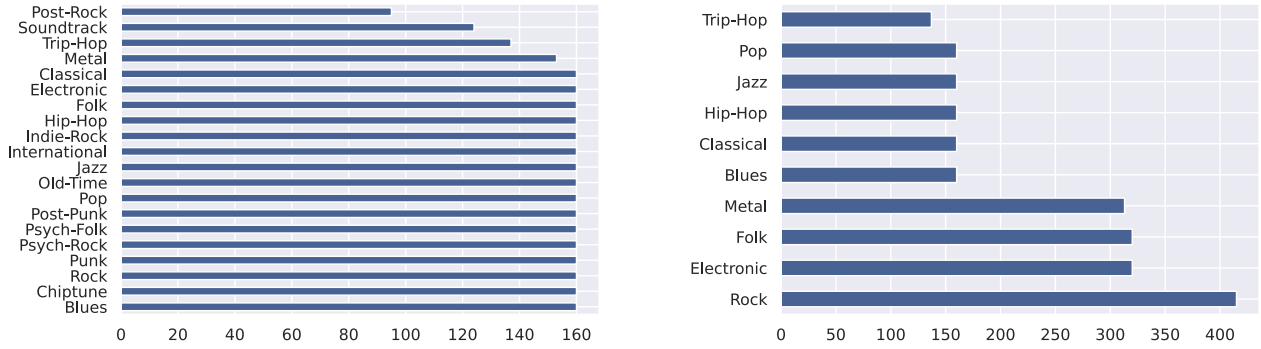
Οι διαφορές μεταξύ ειδών μουσικής είναι εμφανείς και στην περίπτωση των χρωμογραφημάτων, ενώ η επίδραση που έχει η αναγωγή τους στις αντίστοιχες beat-synced εκδοχές είναι ίδια



Εικόνα 1.4: Απεικόνιση των beat-synced χρωμογραφημάτων για τα τυχαία επιλεγμένα δεδομένα.

με πριν, δηλαδή η συρρίκνωση των ακολουθιών time-steps οδηγεί σε φαινόμενα coarse-graining.

Έχοντας μια καθαρή εικόνα της μορφής των δεδομένων, το επόμενο βήμα (Βήμα 4) αποτέλεσε μια βασική προεπεξεργασία τους, πριν αυτά αξιοποιηθούν για την εκπαίδευση ταξινομητών. Ανάμεσα στα 20 είδη μουσικής στα οποία αντιστοιχούν τα δεδομένα υπήρχαν αρκετά συγγενικά είδη, όπως για παράδειγμα Rock και post-Rock, η ένωση των οποίων κρίθηκε σκόπιμη προκειμένου να αποφευχθούν συγχύσεις του ταξινομητή. Παράλληλα, αφαιρέθηκαν από το σύνολο δεδομένων είδη μουσικής που αντιπροσωπεύονταν από λίγα δείγματα και δε μπορούσαν να συγχωνευθούν με συγγενικά είδη. Παρότι το σύνολο δεδομένων ήταν αρχικά ισορροπημένο, η ανάγκη για την αποφυγή σύγχυσης του ταξινομητή που κατέστησε απαραίτητη την προεπεξεργασία αυτή οδήγησε σε ένα μη-ισορροπημένο σύνολο δεδομένων. Το πλήθος δειγμάτων ανά είδος μουσικής πριν και μετά από τη συγκεκριμένη προεπεξεργασία απεικονίζεται στα ιστογράμματα της Εικόνας 1.5.



Εικόνα 1.5: Ιστογράμματα δειγμάτων ανά είδος μουσικής (αριστερά) πριν και (δεξιά) μετά από την προεπεξεργασία των δεδομένων.

Η προεπεξεργασία δεν περιορίστηκε στις προαναφερθείσες συγχωνεύσεις και αφαιρέσεις ειδών, αλλά επεκτάθηκε και στη μορφή των δεδομένων. Συγκεκριμένα, όπως αναλύθηκε στην αναφορά για τη δεύτερη εργαστηριακή άσκηση [LAB2], προκειμένου τα δεδομένα να αξιοποιηθούν για την εκπαίδευση νευρωνικών δικτύων, χωρίστηκαν σε δεδομένα εκπαίδευσης και επικύρωσης (με αναλογία 1/3) και πέρασαν από μια διαδικασία padding. Τέλος, φορτώθηκαν ανά batches σε ξεχωριστούς data loaders, ανάλογα με το αν αντιστοιχούν σε mel-φασματογραφήματα, χρωμογραφήματα ή συνδυασμό αυτών (τα λεγόμενα «fused» δεδομένα), καθώς και με το

εάν έχουν υποστεί beat-syncing ή όχι.

2 Ταξινομηση Μουσικων Κομματιων - LSTM

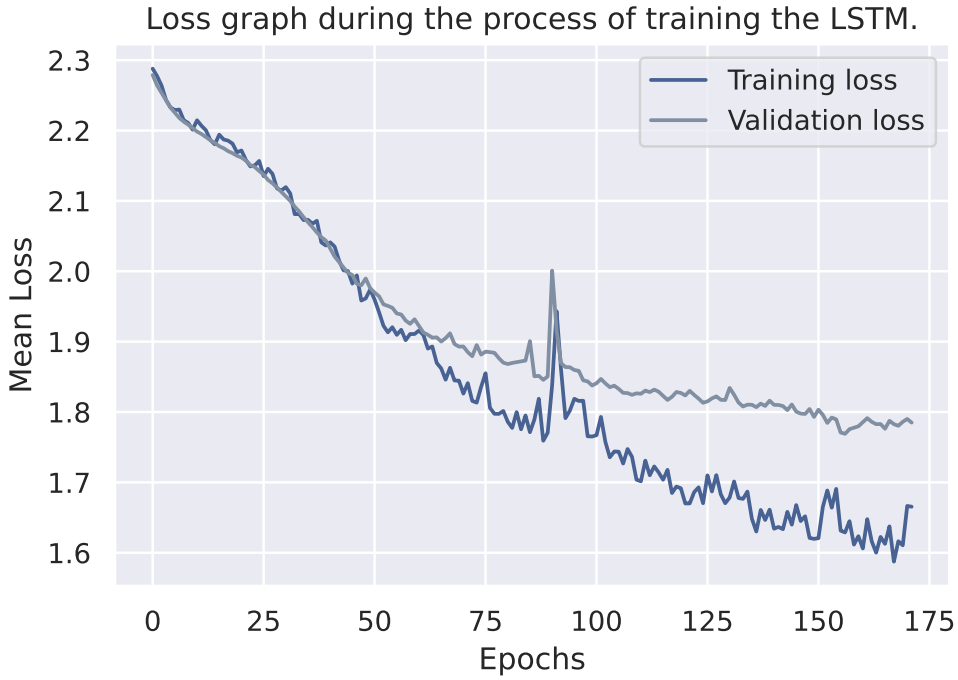
Το περιεχόμενο αυτής της ενότητας αφορά τα Βήματα 5-6 της εργασίας.

Με τα δεδομένα σε μια πλέον αξιοποιήσιμη μορφή, το πρώτο μοντέλο που αναπτύχθηκε για την ταξινόμηση των δεδομένων εκπαίδευσης αποτέλεσε ένα νευρωνικό δίκτυο LSTM (Βήμα 5). Ως ένα sanity check για την εξασφάλιση της λειτουργικότητας του δικτύου, αυτό τέθηκε σε λειτουργία εκπαίδευσης για ένα πολύ μικρό υποσύνολο των δεδομένων, χωρίς επικύρωση, προκειμένου να επαληθευτεί ότι με το πέρας αρκετών εποχών το σφάλμα εκπαίδευσης τείνει στο μηδέν, εξασφαλίζοντας την υπερπροσαρμογή του δικτύου στα δεδομένα (batch overfitting). Η καμπύλη για το σφάλμα εκπαίδευσης απεικονίζεται στην Εικόνα 2.1.

Εικόνα 2.1: Καμπύλη σφάλματος εκπαίδευσης κατά τη δοκιμή υπερπροσαρμογής του δικτύου LSTM.

Το γεγονός πως μετά από μερικές εκατοντάδες εποχές το σφάλμα εκπαίδευσης τείνει στο μηδέν υπέδειξε πως το δίκτυο που αναπτύχθηκε είναι πράγματι λειτουργικό. Έτσι, ρυθμίζοντας την αρχιτεκτονική του ώστε αυτό να είναι bidirectional και να αποτελείται από 2 stacked LSTMs με 100 hidden layers το καθένα, το δίκτυο εκπαιδεύτηκε αρχικά στα mel-φασματογραφήματα του συνόλου των δεδομένων εκπαίδευσης. Προκειμένου να αποφευχθεί η υπερπροσαρμογή του δικτύου στα δεδομένα εκπαίδευσης, στον αλγόριθμο Adam (ο οποίος επιλέχθηκε σύμφωνα με την εκφώνηση ως optimizer) προστέθηκε ένας όρος ομαλοποίησης τύπου L2 (ίσος με 10^{-6}), ενώ επίσης προστέθηκε ένας έλεγχος dropout με πιθανότητα 0.2. Πρόσθετα, η διαδικασία εκπαίδευσης περιλάμβανε ένα κομμάτι επικύρωσης, όπου ο αλγόριθμος early stopping (με patience ίσο με 15) φρόντιζε η εκπαίδευση να τερματιστεί στην περίπτωση που το σφάλμα επικύρωσης έπαυε να μειώνεται. Έτσι, η εκπαίδευση του δικτύου ολοκληρώθηκε έπειτα από 164 εποχές, με τα αποτελέσματα για τις καμπύλες του σφάλματος εκπαίδευσης και επικύρωσης να απεικονίζονται στην Εικόνα 2.2. Αξίζει να τονιστεί πως για το ρυθμό εκμάθησης επιλέχθηκε μια αρκετά

μικρή τιμή (10^{-5}), καθώς σε διαφορετική περίπτωση το σφάλμα επικύρωσης εμφάνιζε ακόμα υψηλότερες διακυμάνσεις από αυτές που φαίνονται στην Εικόνα 2.2.



Εικόνα 2.2: Καμπύλες σφάλματος εκπαίδευσης και επικύρωσης κατά την εκπαίδευση του δικτύου LSTM στα mel-φασματογραφήματα των δεδομένων.

Καθίσταται εμφανές πως το σφάλμα επικύρωσης εμφανίζει μια συνολικά καθοδική τάση κατά την εκπαίδευση του δικτύου, όμως η τιμή στην οποία συγκλίνει δεν είναι αρκετά υψηλή προκειμένου να αναμένει κανείς ικανοποιητικά αποτελέσματα κατά την αξιολόγησή του στην ταξινόμηση άγνωστων δεδομένων.

Ως προς την αξιολόγηση αυτή (Βήμα 6), οι μετρικές που χρησιμοποιήθηκαν για την εκτίμηση της απόδοσης του δικτύου σε ό,τι αφορά την ταξινόμηση σε καθένα από τα 10 υπό μελέτη είδη μουσικής ήταν οι Precision, Recall και F1-Score. Η Precision αποτελεί το πηλίκο του λόγου των δειγμάτων που ορθά ταξινομήθηκαν σε κάποιο είδος μουσικής προς το συνολικό αριθμό δειγμάτων τα οποία ταξινομήθηκαν στο είδος αυτό. Με άλλα λόγια, ισούται με

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

όπου TP = True Positives (σωστές ταξινομήσεις σε κάποιο είδος) και FP = False Positives (δείγματα που εσφαλμένα ταξινομήθηκαν στο είδος αυτό). Σε ό,τι αφορά τη μετρική Recall, αυτή μπορεί κατ' αναλογία με την Precision να οριστεί ως

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

όπου FN = False Negatives (δείγματα που θα έπρεπε να ταξινομηθούν στο συγκεκριμένο είδος, όμως αυτό δε συνέβη). Έτσι, ενώ η Precision εκφράζει την πιθανότητα ένα κομμάτι που ταξινομήθηκε σε κάποιο είδος να ανήκει όντως σε αυτό, η Recall εκφράζει την πιθανότητα ένα κομμάτι που ανήκει σε ένα είδος να ταξινομηθεί ορθά στο είδος αυτό. Τέλος, το F1-Score δεν είναι παρά ο αρμονικός μέσος των Precision και Recall, δηλαδή

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Επειδή οι Precision, Recall και F1-Score αποτελούν μετρικές που αφορούν την ταξινόμηση σε κάθε είδος μουσικής ξεχωριστά, απαραίτητη είναι και η χρήση μετρικών που χαρακτηρίζουν το σύνολο των δεδομένων και άρα μπορούν να αξιολογήσουν τον ταξινομητή συνολικά. Η γνωστότερη εξ αυτών είναι η Accuracy, η οποία δεν είναι παρά ο λόγος των σωστά ταξινομημένων δειγμάτων προς το σύνολο των δειγμάτων. Πέραν αυτής, ορίζονται και οι Micro και Macro Averaged εκδοχές των προαναφερθέντων μετρικών. Η Macro Averaged εκδοχή των Precision, Recall και F1-Score αποτελεί τον απλό μέσο όρο αυτών ως προς όλα τα είδη ταξινόμησης. Από την άλλη, η Micro Average εκδοχή τους αποτελεί ένα βεβαρυμένο μέσο όρο τους, ο οποίος λαμβάνει υπ' όψιν την πιθανή ανισορροπία ανάμεσα στα διάφορα είδη, υπολογίζοντας τις τιμές TP, FP και FN για όλα το σύνολο των δεδομένων. Βάσει αυτών, τα συμπεράσματα που μπορεί κανείς να εξαγάγει είναι τα ακόλουθα:

- Σχετικά υψηλές διαφορές μεταξύ Accuracy και F1-Score αποτελούν ένδειξη ενός μη-ισορροπημένου συνόλου δεδομένων. Στην προκείμενη περίπτωση, όπου όπως έχει ήδη αναλυθεί υπάρχουν μουσικά είδη τα οποία έχουν αισθητά μεγαλύτερη αντιπροσώπευση από άλλα, αναμένονται αποκλίσεις ανάμεσα στις δύο αυτές μετρικές.

- Αντίστοιχη ένδειξη αποτελούν πιθανές διαφορές ανάμεσα στα Micro και Macro Averages των μετρικών.

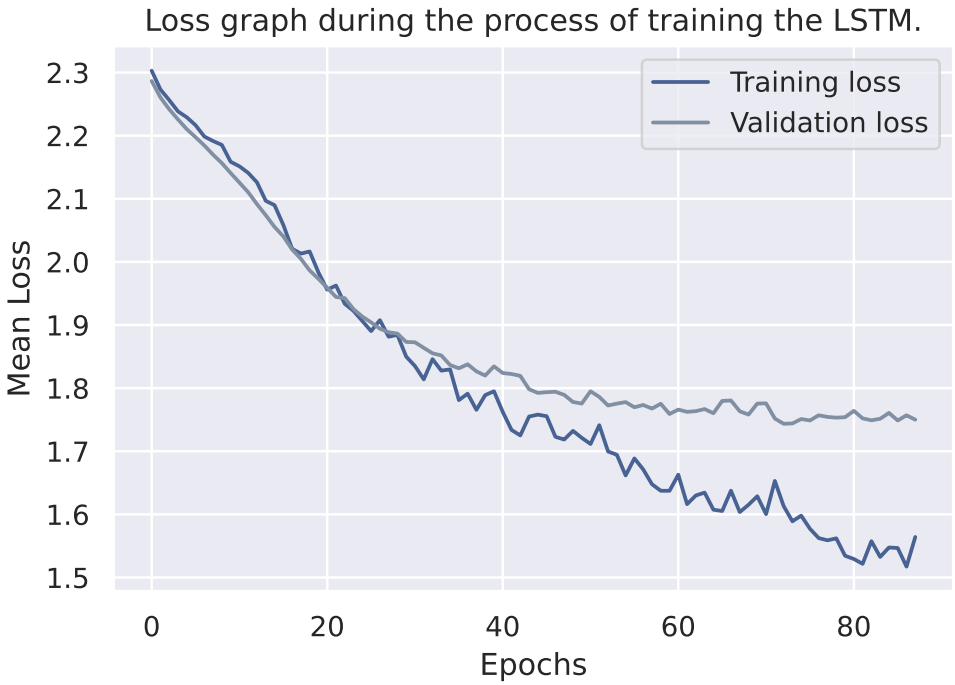
- Το κατά πόσο η Recall ή η Precision συνιστούν καλύτερη μετρική είναι ένα ζήτημα το οποίο εξαρτάται σε μεγάλο βαθμό από το υπό μελέτη πρόβλημα. Για παράδειγμα, ταξινομητές που χρησιμοποιούνται για αναγνώριση ασθενειών οφείλουν να εκτιμούν περισσότερο τα False Negatives σε σχέση με τα False Positives, επομένως η Recall είναι η βέλτιστη επιλογή για την αξιολόγησή τους. Σε τέτοιες περιπτώσεις, η Accuracy και το F1-Score παρέχουν περιορισμένη πληροφορία, αφού το υπό μελέτη πρόβλημα απαιτεί εκ κατασκευής μια προτίμηση σε συγκεκριμένου τύπου σφάλματα ταξινόμησης.

Η απόδοση του εκπαιδευμένου δικτύου στα δεδομένα αξιολόγησης συνοψίζεται βάσει των μετρικών που αναλύθηκαν στον Πίνακα 2.1, στον οποίο σημειώνεται και το πλήθος δειγμάτων κάθε είδους (Support). Η πρώτη σημαντική παρατήρηση είναι πως υπάρχουν 3 είδη μουσικής στα οποία το εκπαιδευμένο δίκτυο δεν ταξινομεί κανένα δείγμα (αυτά με δείκτες 0, 7 και 9). Επιπλέον, όπως ήταν αναμενόμενο, το Accuracy διαφέρει σημαντικά από τις επί μέρους τιμές του F1-Score για κάθε είδος μουσικής, λόγω της σημαντικής ανισορροπίας του συνόλου δεδομένων. Η ανισορροπία αυτή γίνεται αισθητή και από την απόκλιση των Macro και Micro Averaged τιμών των μετρικών η οποία, παρότι μικρή ως απόλυτο μέγεθος, είναι σημαντική δεδομένων των χαμηλών τιμών των μετρικών. Το είδος με τη μεγαλύτερη συνέπεια στα αποτελέσματα των τριών μετρικών αντιστοιχεί στο δείκτη 8 (δηλαδή τη Rock μουσική) και είναι και το είδος με τη μεγαλύτερη εκπροσώπηση, τόσο στα δεδομένα εκπαίδευσης, όσο και στα δεδομένα αξιολόγησης. Τέλος, αν και αισθητά λιγότερο συνεπείς μεταξύ τους, οι μετρικές για τα είδη με δείκτες 2, 3 και 6 (Folk, Electronic και Metal) λαμβάνουν σχετικά υψηλότερες τιμές. Δεν είναι τυχαίο το γεγονός πως οι τρεις αυτές κατηγορίες είναι αυτές με την υψηλότερη εκπροσώπηση αμέσως μετά το είδος Rock.

	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	40
1	0.43	0.47	0.45	40
2	0.33	0.70	0.45	80
3	0.37	0.59	0.46	80
4	0.23	0.07	0.11	40
5	0.25	0.03	0.05	40
6	0.41	0.56	0.47	78
7	0.00	0.00	0.00	40
8	0.30	0.32	0.31	103
9	0.00	0.00	0.00	34
Accuracy			0.35	575
Macro Averaged	0.23	0.27	0.23	575
Micro Averaged	0.27	0.35	0.29	575

Πίνακας 2.1: Συνολική αξιολόγηση του μοντέλου για την ταξινόμηση των mel-φασματογραφημάτων.

Ακολουθώντας ακριβώς την ίδια διαδικασία εκπαίδευσης και αξιολόγησης, η ταξινόμηση πραγματοποιήθηκε και με χρήση των αντίστοιχων beat-synced mel-φασματογραφημάτων. Οι σχετικές καμπύλες σφαλμάτων απεικονίζονται στην Εικόνα 2.3, όπου φαίνεται πως η εκπαίδευση του δικτύου τερματίστηκε μέσω early stopping με το πέρας της 73ης εποχής.



Εικόνα 2.3: Καμπύλες σφάλματος εκπαίδευσης και επικύρωσης κατά την εκπαίδευση του δικτύου LSTM στα beat-synced mel-φασματογραφήματα των δεδομένων.

Σε ό,τι αφορά την απόδοση του δικτύου στα δεδομένα αξιολόγησης, αυτή συνοψίζεται στα περιεχόμενα του Πίνακα 2.2.

	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	40
1	0.43	0.47	0.45	40
2	0.37	0.75	0.49	80
3	0.35	0.59	0.44	80
4	0.34	0.25	0.29	40
5	0.50	0.05	0.09	40
6	0.47	0.51	0.49	78
7	0.00	0.00	0.00	40
8	0.37	0.41	0.39	103
9	0.00	0.00	0.00	34
Accuracy			0.38	575
Macro Averaged	0.28	0.30	0.26	575
Micro Averaged	0.32	0.38	0.32	575

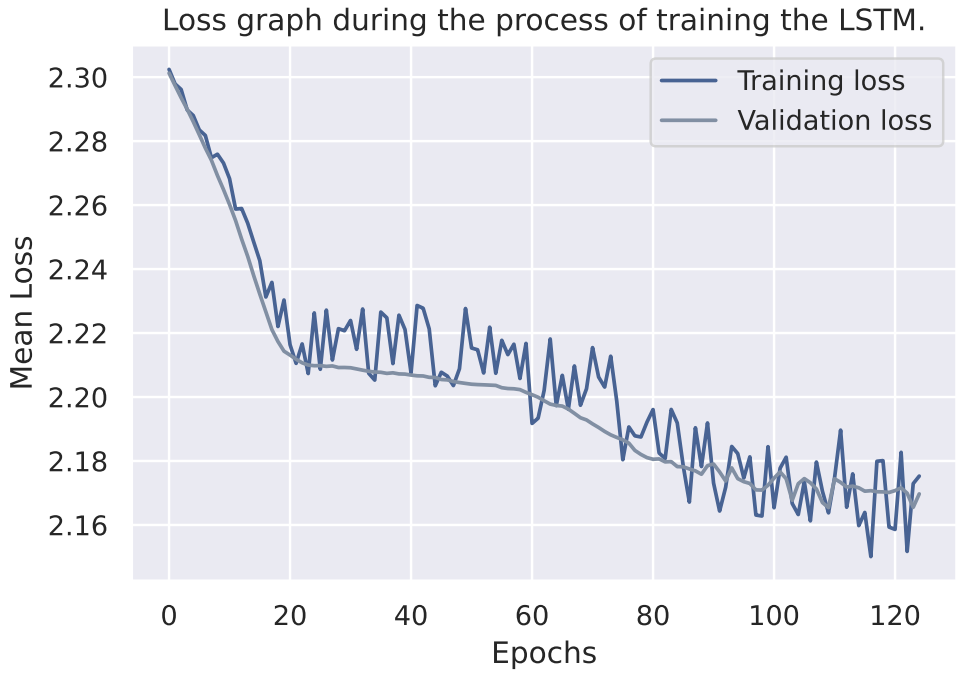
Πίνακας 2.2: Συνολική αξιολόγηση του μοντέλου για την ταξινόμηση των beat-synced mel-φασματογραφημάτων.

Παρότι τα αποτελέσματα δεν είναι ιδιαίτερα ενθαρρυντικά ούτε σε αυτήν την περίπτωση, αξίζει να σημειωθεί πως η βελτίωση σε σχέση με τα non-beat-synced δεδομένα είναι εμφανής. Πρόσθετα, τονίζεται πως η εκπαίδευση του δικτύου απαίτησε πολύ μικρότερο χρόνο, λόγω του γεγονότος πως οι ακολουθίες των beat-synced δεδομένων είχαν μήκος κατά μία τάξη μεγέθους μικρότερο σε σχέση με το μήκος των ακολουθιών των non-beat-synced δεδομένων.

Η ίδια διαδικασία επαναλήφθηκε και για τα χρωμογραφήματα, με τις αντίστοιχες καμπύλες σφαλμάτων να φαίνονται στην Εικόνα 2.4, ενώ τα αποτελέσματα της αξιολόγησης παρατίθενται στον Πίνακα 2.3.

	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	40
1	0.00	0.00	0.00	40
2	0.00	0.00	0.00	80
3	0.19	0.72	0.30	80
4	0.00	0.00	0.00	40
5	0.00	0.00	0.00	40
6	0.22	0.40	0.29	78
7	0.00	0.00	0.00	40
8	0.16	0.20	0.18	103
9	0.00	0.00	0.00	34
Accuracy			0.19	575
Macro Averaged	0.06	0.13	0.08	575
Micro Averaged	0.09	0.19	0.11	575

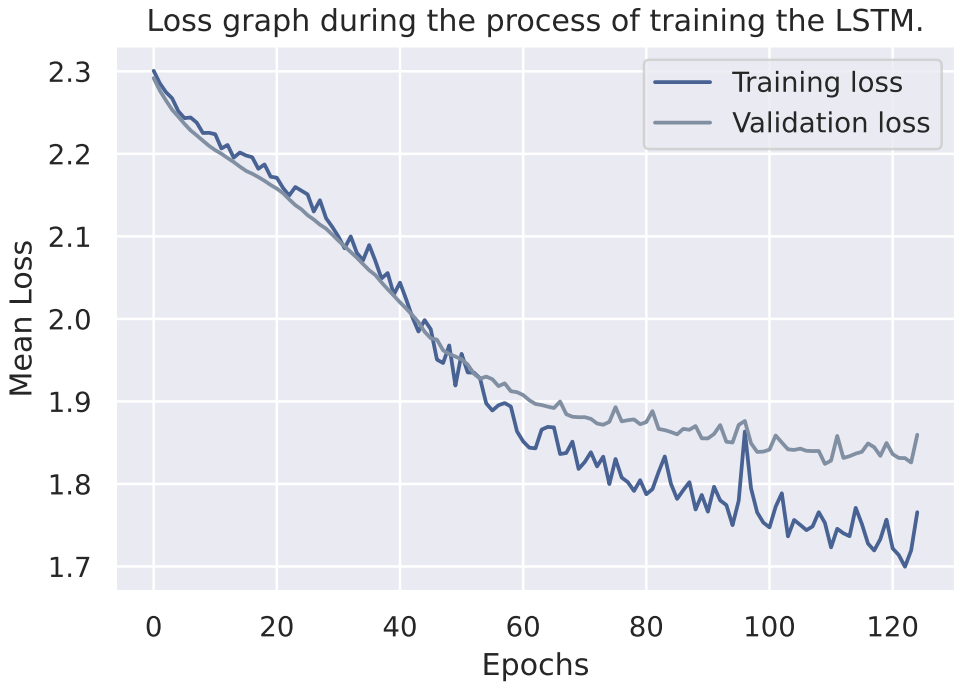
Πίνακας 2.3: Συνολική αξιολόγηση του μοντέλου για την ταξινόμηση των χρωμογραφημάτων.



Εικόνα 2.4: Καμπύλες σφάλματος εκπαίδευσης και επικύρωσης κατά την εκπαίδευση του δικτύου LSTM στα χρωμογραφήματα των δεδομένων.

Σε αντίθεση με τα mel-φασματογραφήματα, γίνεται αντιληπτό πως τα χρωμογραφήματα δεν αποτελούν εξίσου καλά δεδομένα εκπαίδευσης. Η εκπαίδευση του δικτύου χαρακτηρίζεται από υψηλές διακυμάνσεις σε ό,τι αφορά το σφάλμα εκπαίδευσης, ενώ το σφάλμα επικύρωσης δεν καταφέρνει να ξεπεράσει καν το κατώφλι του 2.15, τη στιγμή που στις προηγούμενες περιπτώσεις έφτασε μέχρι την τιμή ~ 1.75. Σε σχέση με τα δεδομένα αξιολόγησης, το δίκτυο πραγματοποίησε ταξινομήσεις μόνο σε τρία είδη μουσικής (τρία από τα τέσσερα είδη με την υψηλότερη αντιπροσώπευση), πετυχαίνοντας μόλις το μισό Accuracy σε σχέση με αυτό που επετεύχθη αξιοποιώντας τα beat-synced mel-φασματογραφήματα. Αντίστοιχα απογοητευτική είναι η απόδοσή του και ως προς τις υπόλοιπες μετρικές, με τις averaged εκδοχές του Precision να αντιστοιχούν σε μονοψήφια ποσοστά. Με εξαίρεση την τιμή για το Recall στο είδος μουσικής με δείκτη 2, η οποία ήταν η υψηλότερη μετρική και για τα αποτελέσματα των mel-φασματογραφημάτων, όλες οι μετρικές υποδεικνύουν πως τα χρωμογραφήματα από μόνα τους δεν επαρκούν για την κατασκευή ενός καλού ταξινομητή.

Κλείνοντας την ανάλυση των κύκλων εκπαίδευσης και ταξινόμησης με χρήση του δικτύου LSTM, παρατίθενται τα αποτελέσματα για την περίπτωση των fused δεδομένων, δηλαδή τη συνένωση των mel-φασματογραφημάτων με τα χρωμογραφήματα. Οι καμπύλες για το σφάλμα εκπαίδευσης και επικύρωσης φαίνονται στην Εικόνα 2.5, ενώ στον Πίνακα 2.4 παρατίθενται τα αποτελέσματα για τις μετρικές αξιολόγησης. Παρότι η εκπαίδευση του δικτύου διήρκησε για λιγότερες εποχές σε σχέση με την περίπτωση των αμιγώς mel-φασματογραφημάτων, τα τελικά αποτελέσματα είναι παρόμοια για τις δύο περιπτώσεις, ως προς όλες τις απόψεις. Το γεγονός αυτό είναι αναμενόμενο: η προηγούμενη διερεύνηση υπέδειξε πως τα χρωμογραφήματα δεν αποτελούν καλά δεδομένα εκπαίδευσης για τα δίκτυα LSTM. Έτσι, η επίδοση του δικτύου που εκπαιδεύεται στα fused δεδομένα είναι λογικό να αποδίδεται σχεδόν αποκλειστικά στην εκπαίδευσή του στο κομμάτι των δεδομένων που απαρτίζονται από τα mel-φασματογραφήματα.



Εικόνα 2.5: Καμπύλες σφάλματος εκπαίδευσης και επικύρωσης κατά την εκπαίδευση του δικτύου LSTM στα fused δεδομένα.

	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	40
1	0.39	0.53	0.45	40
2	0.30	0.68	0.42	80
3	0.35	0.46	0.40	80
4	1.00	0.03	0.05	40
5	1.00	0.03	0.05	40
6	0.41	0.49	0.45	78
7	0.00	0.00	0.00	40
8	0.29	0.40	0.34	103
9	0.00	0.00	0.00	34
Accuracy			0.34	575
Macro Averaged	0.37	0.26	0.21	575
Micro Averaged	0.37	0.34	0.27	575

Πίνακας 2.4: Συνολική αξιολόγηση του μοντέλου για την ταξινόμηση των fused δεδομένων.

Μια εναλλακτική προσέγγιση στην ταξινόμηση κομματιών σε μουσικά είδη είναι η αντιμετώπιση των φασματογραφημάτων ως εικόνες, όπου αντί για pixels στον άξονα του «μήκους» και του «πλάτους», υπάρχουν τα time-steps και οι συχνότητες. Στα πλαίσια αυτά, αποφασίστηκε η εκπαίδευση Συνελικτικών Νευρωνικών Δικτύων για την ταξινόμηση των δεδομένων.

3 Ταξινομηση Μουσικων Κομματιων - CNN

Το περιεχόμενο αυτής της ενότητας αφορά το Βήμα 7 της εργασίας.

Τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNN) αποτελούν ιδιαίτερα δημοφιλή εργαλεία για την επεξεργασία και την αναγνώριση εικόνας. Τα κύρια δομικά στοιχεία ενός CNN που το διαφοροποιούν από τα συνήθη νευρωνικά δίκτυα αποτελούν τα συνελικτικά επίπεδά του (convolutional layers) σε συνδυασμό με τα επίπεδα υποδειγματοληψίας (pooling).

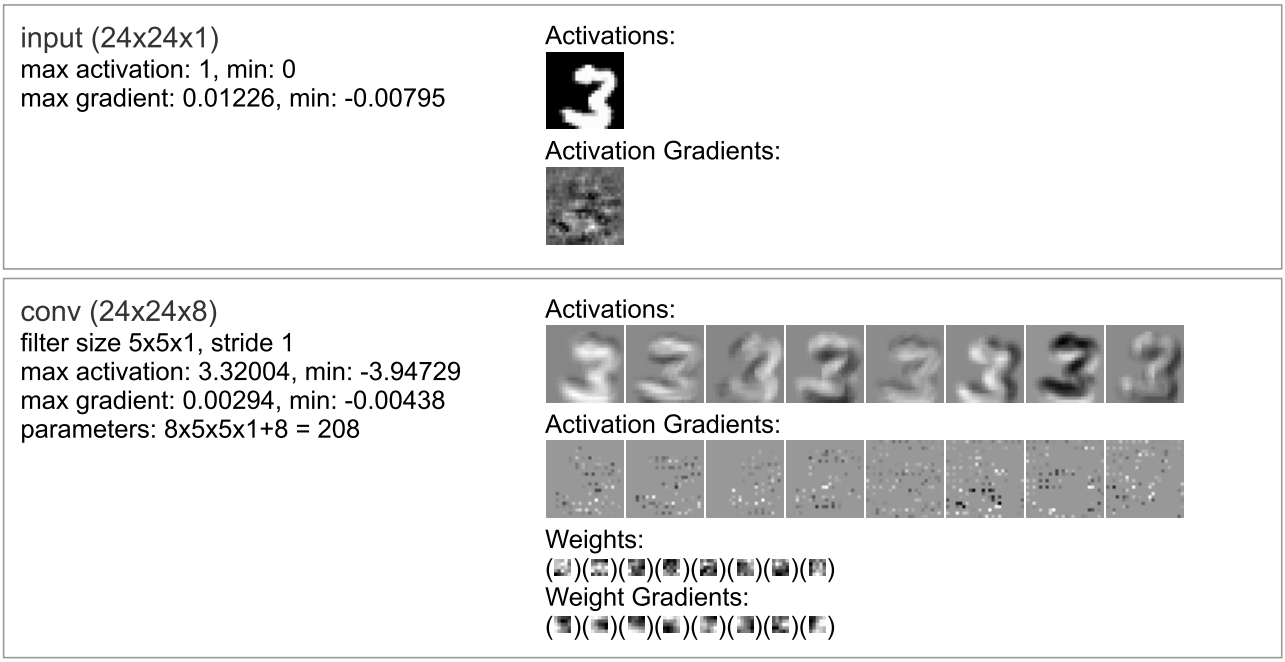
Σε ένα convolutional layer υπολογίζεται η συνέλιξη της εισόδου με κάποιον σταθερό πυρήνα (kernel) μικρότερης διάστασης. Αυτό κατά κάποιον τρόπο αφαιρεί το θόρυβο και ομαλοποιεί τα δεδομένα εισόδου. Στο επίπεδο του pooling, η είσοδος διαμερίζεται σε μικρότερα κομμάτια (ξένα ή μη) πάνω στα οποία κρατείται μονάχα μία τιμή, πχ. η μέγιστη τιμή όλων των επιμέρους κομματιών (γνωστό και ως max pooling). Με αυτόν τον τρόπο μειώνεται η διάσταση της εισόδου, καθώς κάθε γειτονιά δεδομένων αντιπροσωπεύεται πλέον από μία μόνο τιμή. Γεωμετρικά, θα μπορούσε κανείς να πει πως, ενώ ένα convolutional layer ομαλοποιεί την εικόνα, ένα pooling layer κρατά ένα πιο πρόχειρο περίγραμμά της που βασίζεται στα πιο αντιπροσωπευτικά features της, ενώ παράλληλα μειώνει σημαντικά τη διάστασή της.

Ένα από τα σημαντικότερα αίτια για την επιτυχία των CNN στην αναγνώριση εικόνας εντοπίζεται στο γεγονός πως τα συνελικτικά επίπεδά τους είναι αναλλοίωτα στις μεταφορές. Έτσι, γεωμετρικά χαρακτηριστικά της εικόνας, όπως γωνίες, πλευρές, κ.α., τα οποία εμφανίζονται στην είσοδο, παραμένουν αναλλοίωτα και κατά την έξοδο τους από το επίπεδο [Cal20, p. 521]. Το ίδιο ισχύει και για τα επίπεδα του pooling, αλλά μόνο τοπικά [Cal20, p. 511]. Παράλληλα, έχουν σημαντικά λιγότερα βάρη, και επομένως πιο εύκολη εκπαίδευση σε σχέση με ένα πλήρως συνεκτικό δίκτυο αντίστοιχου μεγέθους.

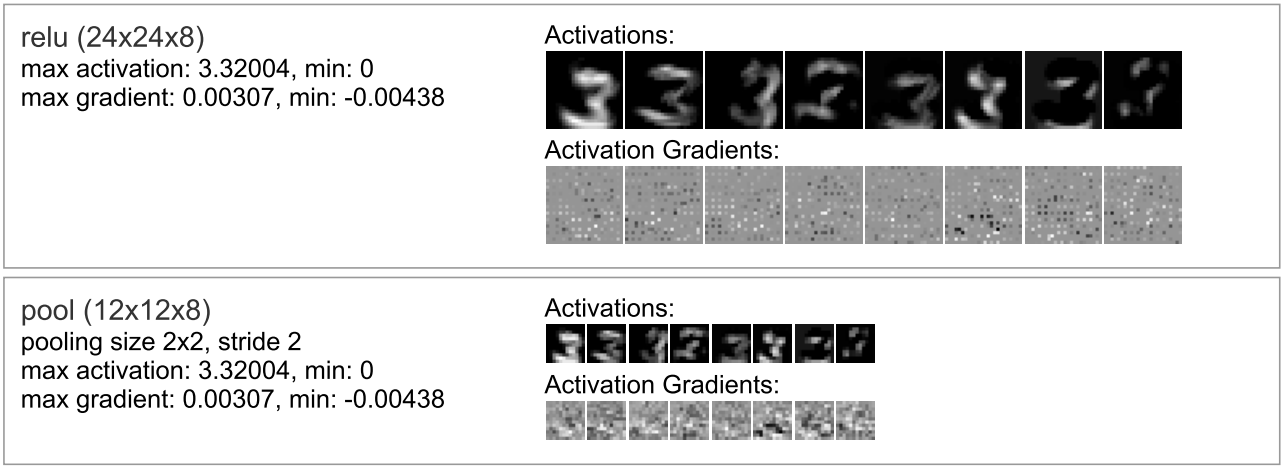
Στα υπόλοιπα δομικά στοιχεία ενός CNN συμπεριλαμβάνονται οι συναρτήσεις ενεργοποίησης που δρουν στην έξοδο των συνελικτικών και pooling επιπέδων. Η πιο συνηθισμένη επιλογή είναι η Rectified Linear Unit (ReLU), $f(x) = \max\{0, x\}$. Το πλεονέκτημά της έναντι των σιγμοειδών συναρτήσεων είναι ότι αποφεύγεται ο κορεσμός, καθώς $\lim_{x\to\infty} f(x) = +\infty$, το οποίο έχει ως αποτέλεσμα τα δίκτυα να εκπαιδεύονται πολύ γρηγορότερα.

Μια ακόμα συνηθισμένη πρακτική είναι αυτή του batch normalization, κατά την οποία τα δεδομένα κανονικοποιούνται προτού εφαρμοστεί σε αυτά η συνάρτηση του κάθε επιπέδου. Ο λόγος είναι ότι, ιδίως σε βαθιά δίκτυα, μικρές αλλαγές στα βάρη των υψηλότερων στρωμάτων μπορεί να δημιουργήσουν μεγάλες αλλαγές σε αυτά των υπολοίπων, αλλάζοντας επίσης και τις αντίστοιχες κατανομές. Η κανονικοποίηση εξασφαλίζει ότι τα βάρη εξακολουθούν να βλέπουν δεδομένα ίδιας τάξης μεγέθους με αυτά που έβλεπαν πριν την ανανέωσή τους. Τέλος, στο προτελευταίο επίπεδο συμπεριλαμβάνεται ένα fully connected layer, το οποίο συνδυάζει τις τοπικές πληροφορίες των προηγούμενων επιπέδων για να δώσει την τελική πρόβλεψη για ολόκληρη την αρχική εικόνα.

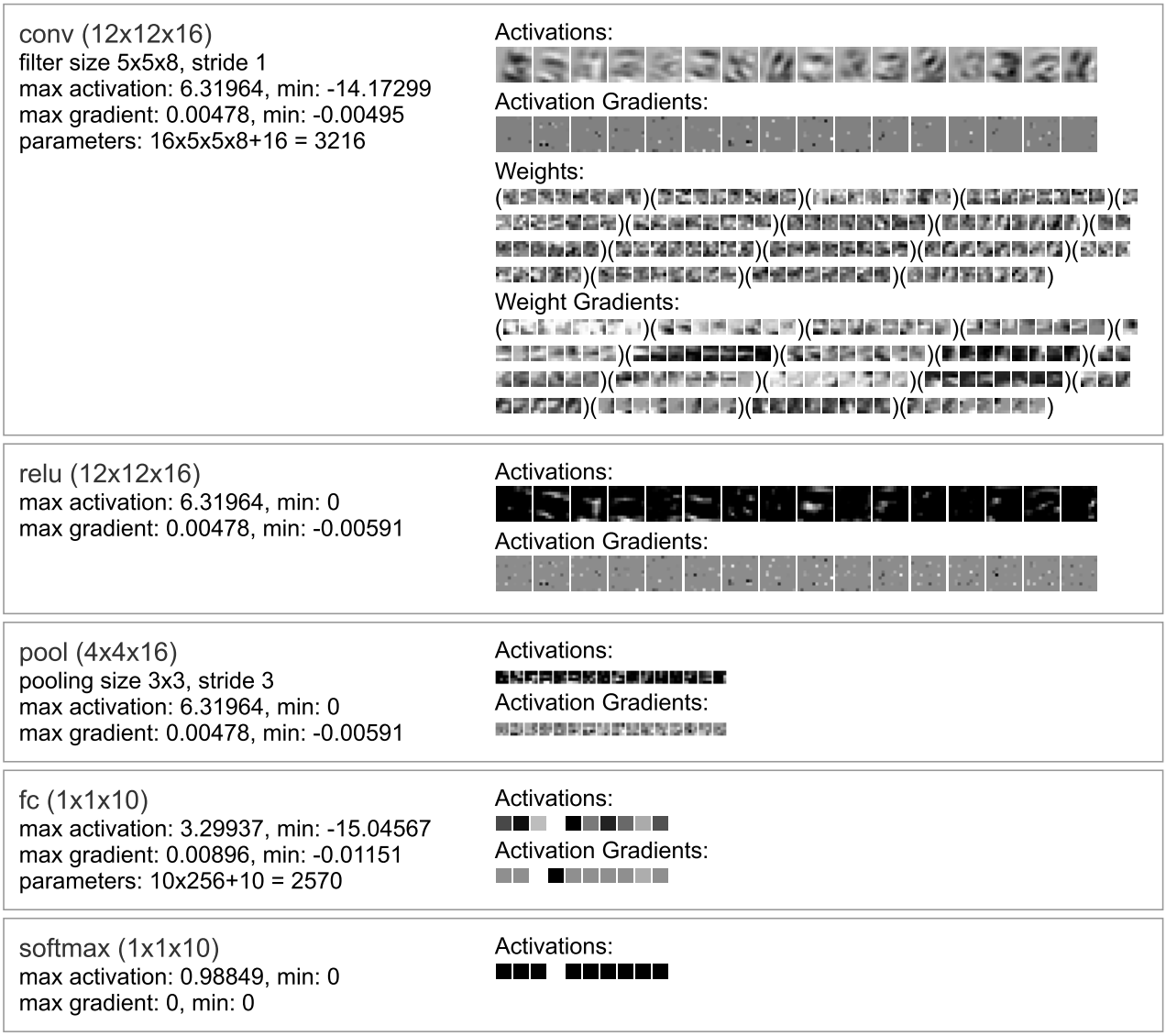
Στη σελίδα ConvNetJS δόθηκε η δυνατότητα εκπαίδευσης ενός CNN στα δεδομένα του MNIST, τα οποία περιλαμβάνουν εικόνες ψηφίων από το 0 εως το 9, παρόμοιες με αυτές της πρώτης εργαστηριακής άσκησης. Το default δίκτυο της ιστοσελίδας πέτυχε εξαιρετικά αποτελέσματα, τα οποία όμως δε θα σχολιαστούν λεπτομερώς παρακάτω, αφού το κύριο αντικείμενο ενδιαφέροντος αποτελούν τα δομικά συστατικά του δικτύου και η διαισθητική ερμηνεία των εξόδων του σε κάθε επίπεδο. Στο πλαίσιο αυτό, παρατίθενται ακολούθως οι Εικόνες 3.1 - 3.3, στις περιγραφές των οποίων αναλύονται τα δομικά στοιχεία του CNN που χρησιμοποιήθηκε και το πώς αυτά επεξεργάζονται, τροποποιούν και εν τέλει ταξινομούν ένα δεδομένο (εικόνα ψηφίου) εισόδου.



Εικόνα 3.1: Αρχικά, κάθε εικόνα είχε μέγεθος 28×28, όμως προτού τροφοδοτηθεί στο πρώτο επίπεδο της αφαιρούταν με τυχαίο τρόπο ένα 4×4 πλαίσιο. Έτσι, η πρώτη είσοδος είχε πάντα διάσταση 24×24. Κατά το πρώτο συνελικτικό επίπεδο τα γεωμετρικά χαρακτηριστικά του ψηφίου 3 διατηρήθηκαν, ενώ ταυτόχρονα ομαλοποιήθηκε κάπως το περίγραμμά του, αποτέλεσμα του αναλλοίωτου των συνελικτικών επιπέδων στις μεταφορές και της πράξης της συνέλιξης, αντίστοιχα.



Εικόνα 3.2: Στη συνέχεια εφαρμόστηκε στο αποτέλεσμα η συνάρτηση ενεργοποίησης ReLU, η οποία διατήρησε ταυτοτικά όσες τιμές βρίσκονταν πάνω από ένα ορισμένο κατώφλι (threshold), ενώ μηδένισε όλες τις υπόλοιπες. Έτσι, κάποια κομμάτια του αριθμού 3, αλλά και της υπόλοιπης εικόνας, παρέμειναν ακριβώς ίδια, ενώ όλα τα υπόλοιπα έγιναν μαύρα. Έπειτα, στο pooling layer, εφαρμόστηκε ένα παράθυρο 2×2 με βήμα 2, δηλαδή συνολικά 12×12 υποπαράθυρα ξένα μεταξύ τους, καθένα εκ των οποίων κρατούσε μόνο τη μέγιστη τιμή του 2×2 υποπίνακά του. Έγινε εμφανής όχι μόνο η μείωση της διάστασης της εικόνας εξόδου, αλλά και η διατήρηση των κύριων γεωμετρικών χαρακτηριστικών του ψηφίου 3, απόρροια της ιδιότητας του τοπικού αναλλοίωτου για το pooling operation.

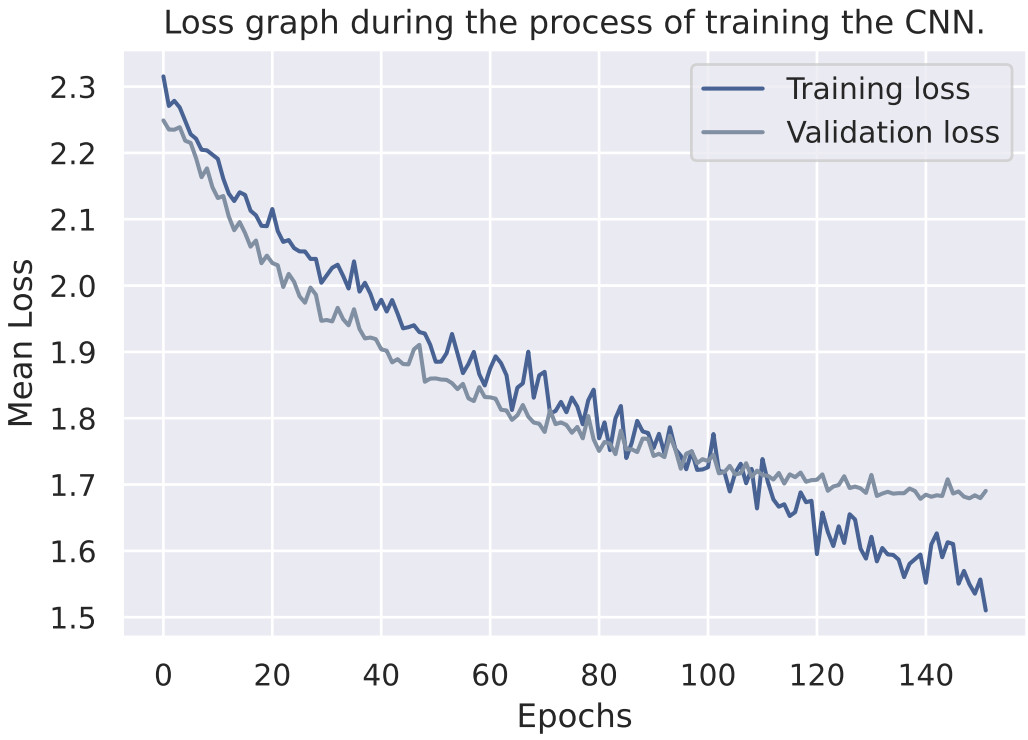


Εικόνα 3.3: Προχωρώντας, οι διαστάσεις γίνονταν ολοένα και μικρότερες, με αποτέλεσμα η ανθρώπινη αντίληψη να αποτυγχάνει να διακρίνει κάποιο εμφανές σχήμα ή μοτίβο. Το δίκτυο είχε κρατήσει τοπικά την κύρια πληροφορία που το ενδιέφερε γύρω από τα πιο αντιπροσωπευτικά σημεία. Τέλος, στο FC επίπεδο, το δίκτυο συνδύαζε όλες τις προηγούμενες πληροφορίες ώστε να κατατάξει τη δοθείσα εικόνα σε μία από τις 10 κατηγορίες. Στην έξοδο της softmax επιλεγόταν αυτή με τη μεγαλύτερη πιθανότητα (λευκή).

Βάσει της παραπάνω ανάλυσης, αναπτύχθηκε ένα CNN με σκοπό την ταξινόμηση των δεδομένων του συνόλου FMA. Το πρώτο μέρος του CNN αποτελούταν από 4 επίπεδα, καθένα εκ των οποίων πρώτα πραγματοποιούσε μια 2D συνέλιξη, στη συνέχεια ένα batch normalization (το οποίο ήταν configurable κατά την κλήση της κλάσης), κατόπιν κλήση της ReLU και τελικά ένα max pooling. Σε κάθε περίπτωση, το max pooling πραγματοποιούταν σε 2×2 υποπίνακες του input, ενώ ο πυρήνας των 2D συνελίξεων ήταν 3×3. Σε ό,τι αφορά το βάθος των συνελίξεων, το πρώτο επίπεδο είχε έξοδο 4, το δεύτερο είχε έξοδο 16, το τρίτο είχε έξοδο 64, ενώ το τελικό είχε έξοδο 256. Στην έξοδο του τελικού επιπέδου πραγματοποιούταν ένα flattening, ώστε το output να περάσει στο δεύτερο μέρος του CNN, το οποίο αποτελούταν από τρία επίπεδα FC. Στην έξοδο καθενός εξ αυτών (πέραν του τελευταίου) εφαρμόστηκε ενεργοποίηση ReLU, ενώ επίσης προστέθηκε το ενδεχόμενο dropout με πιθανότητα 0.23. Το πρώτο FC επίπεδο είχε έξοδο 128, το

δεύτερο είχε έξοδο 64, ενώ το τελικό είχε έξοδο 10, όσα δηλαδή και τα είδη προς ταξινόμηση.

Αφότου η βασική λειτουργικότητα του CNN ελέγχθηκε μέσω batch overfitting (μάλιστα, η υπερπροσαρμογή του CNN πραγματοποιήθηκε πολύ γρηγορότερα σε σχέση με αυτήν του LSTM, χάρη στα ιδιαίτερα χαρακτηριστικά του καθώς και χάρη στο batch normalization), αξιοποιήθηκε για την ταξινόμηση των δεδομένων, με τα αποτελέσματα για τις καμπύλες σφαλμάτων και τα αποτελέσματα αξιολόγησης να απεικονίζονται στις Εικόνες 3.4 - 3.7.

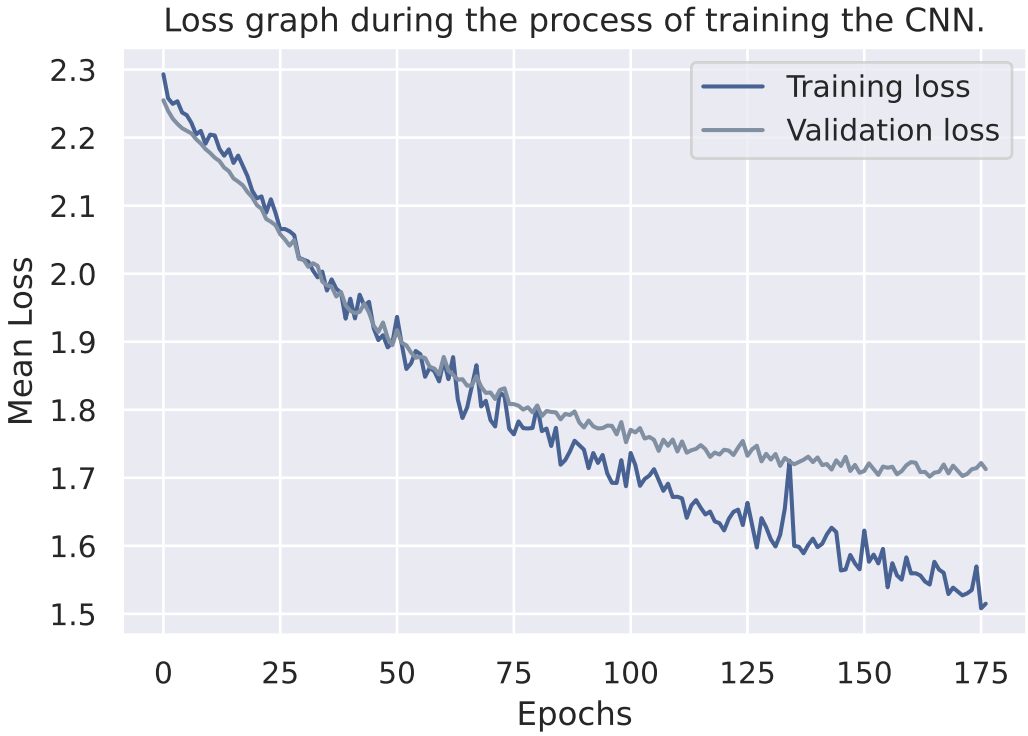


	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	40
1	0.43	0.68	0.52	40
2	0.40	0.70	0.51	80
3	0.34	0.65	0.45	80
4	0.39	0.40	0.40	40
5	0.00	0.00	0.00	40
6	0.58	0.46	0.51	78
7	0.00	0.00	0.00	40
8	0.42	0.49	0.45	103
9	0.00	0.00	0.00	34
Accuracy			0.41	575
Macro Averaged	0.26	0.34	0.28	575
Micro Averaged	0.32	0.41	0.35	575

Εικόνα 3.4: Καμπύλες σφαλμάτων και Πίνακας αξιολόγησης για το δίκτυο CNN στα mel-φασματογραφήματα των δεδομένων.

Σε ό,τι αφορά την αξιολόγηση των mel-φασματογραφημάτων, γίνεται εμφανές πως, παρότι η επίδοση του CNN δεν ήταν πολύ ικανοποιητική, ξεπέρασε αυτήν του LSTM για το ίδιο σύνολο

δεδομένων ως προς όλα τα κριτήρια. Αξίζει, επίσης, να αναφερθεί πως το CNN απέφυγε τις ταξινομήσεις στα ίδια είδη μουσικής που απέφυγε και το LSTM.

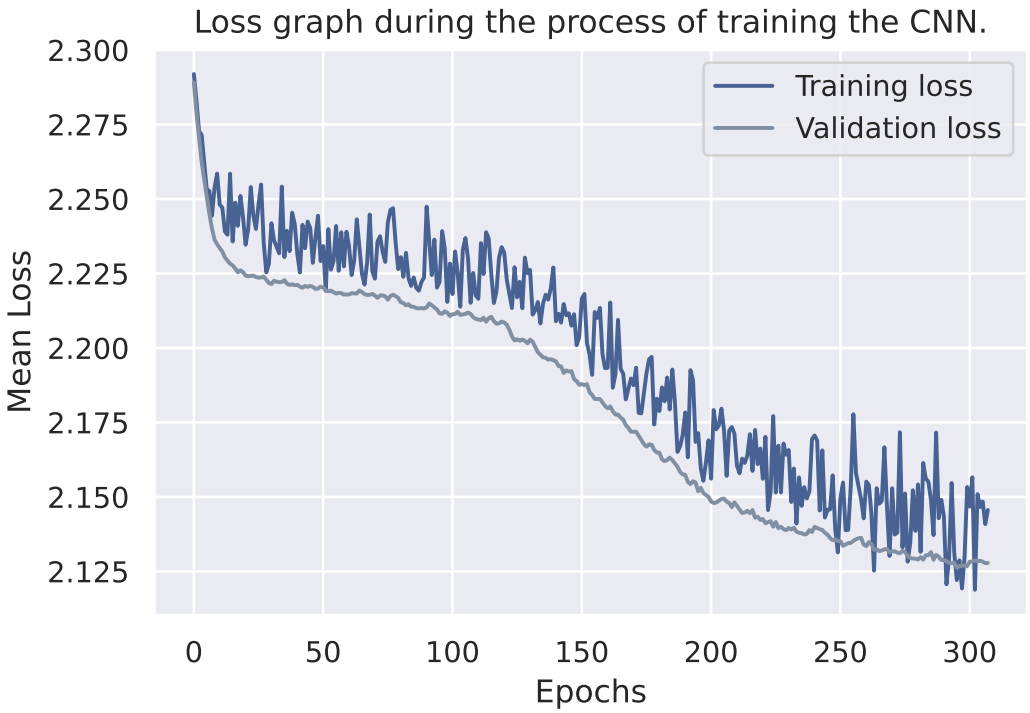


	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	40
1	0.43	0.65	0.51	40
2	0.38	0.61	0.47	80
3	0.42	0.55	0.48	80
4	0.35	0.45	0.39	40
5	0.17	0.05	0.08	40
6	0.50	0.58	0.54	78
7	0.00	0.00	0.00	40
8	0.40	0.42	0.41	103
9	0.26	0.15	0.19	34
Accuracy			0.40	575
Macro Averaged	0.29	0.35	0.31	575
Micro Averaged	0.33	0.40	0.36	575

Εικόνα 3.5: Καμπύλες σφαλμάτων και Πίνακας αξιολόγησης για το δίκτυο CNN στα beat-synced mel-φασματογραφήματα των δεδομένων.

Σε ό,τι αφορά τα beat-synced mel-φασματογραφήματα, το CNN σημείωσε αποτελέσματα όμοια με αυτά στα non-beat-synced δεδομένα. Η επίδοσή του παρέμεινε καλύτερη από αυτή του LSTM κατά την αξιολόγησή του στα αντίστοιχα δεδομένα. Αξίζει στο σημείο αυτό να αναφερθεί πως η ίδια η δομή των CNN επιβάλλει αυτά να εκπαιδεύονται για την ταξινόμηση εικόνων (ή, εν γένει, 2D αντικειμένων) συγκεκριμένου μήκους και πλάτους. Στην περίπτωση των beat-synced mel-φασματογραφημάτων, το padding που πραγματοποιήθηκε στα δεδομένα εκπαίδευσης επέβαλε

στα time-steps των δεδομένων να παίρνουν 129 τιμές. Από την άλλη, η μέγιστη ακολουθία για τα δεδομένα αξιολόγησης της ίδιας κατηγορίας είχε μήκος 116 time-steps. Έτσι, προκειμένου τα δεδομένα να μπορέσουν να ταξινομηθούν από το CNN, πραγματοποιήθηκε σε αυτά επιπλέον padding, προκειμένου το μήκος των ακολουθιών τους να ισούται με 129 time-steps. Στην αντίστροφη περίπτωση, όπου δηλαδή η ακολουθία μέγιστου μήκους time-steps ανήκει στα δεδομένα αξιολόγησης και όχι στα δεδομένα εκπαίδευσης, ο μόνος τρόπος τα αντίστοιχα δεδομένα να ταξινομηθούν θα ήταν η αποκοπή ορισμένων time-steps.

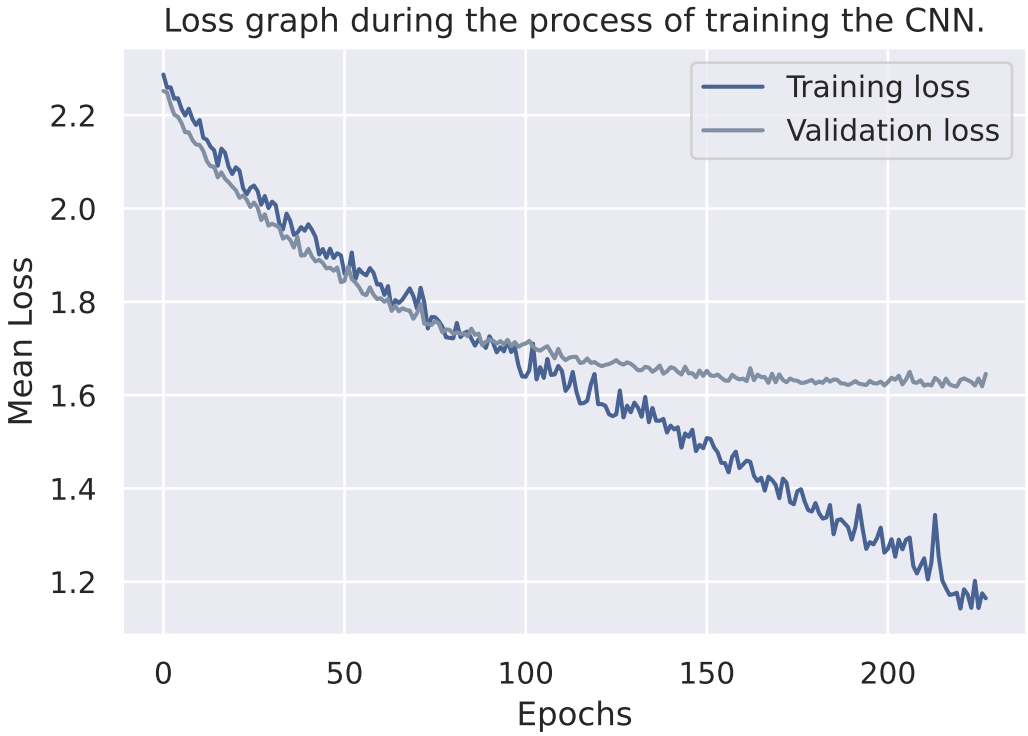


	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	40
1	0.00	0.00	0.00	40
2	0.00	0.00	0.00	80
3	0.21	0.62	0.32	80
4	0.00	0.00	0.00	40
5	0.00	0.00	0.00	40
6	0.26	0.53	0.35	78
7	0.00	0.00	0.00	40
8	0.23	0.41	0.29	103
9	0.00	0.00	0.00	34
Accuracy			0.23	575
Macro Averaged	0.07	0.16	0.10	575
Micro Averaged	0.11	0.23	0.14	575

Εικόνα 3.6: Καμπύλες σφαλμάτων και Πίνακας αξιολόγησης για το δίκτυο CNN στα χρωμογραφήματα των δεδομένων.

Προχωρώντας στην επίδοση της ταξινόμησης βάσει χρωμογραφημάτων, αν και ξανά παρατη-

ρήθηκε μια σχετική βελτίωση σε σχέση με το LSTM, δεν παύει να ισχύει πως αυτά αποτελούν το χειρότερο σύνολο δεδομένων για την εκπαίδευση των δικτύων.



	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	40
1	0.49	0.70	0.58	40
2	0.48	0.57	0.52	80
3	0.36	0.64	0.46	80
4	0.33	0.42	0.37	40
5	0.13	0.05	0.07	40
6	0.54	0.55	0.54	78
7	0.00	0.00	0.00	40
8	0.39	0.51	0.44	103
9	0.00	0.00	0.00	34
Accuracy			0.42	575
Macro Averaged	0.27	0.35	0.30	575
Micro Averaged	0.33	0.42	0.36	575

Εικόνα 3.7: Καμπύλες σφαλμάτων και Πίνακας αξιολόγησης για το δίκτυο CNN στα fused δεδομένα.

Τέλος, στα fused δεδομένα το CNN σημείωσε το μεγαλύτερο μέχρι το σημείο αυτό score ως προς όλες τις μετρικές, αν και σε πρακτικό επίπεδο αυτό είναι ίδιο με το score που πέτυχε και στα mel- και beat-synced mel-φασματογραφήματα.

Συνολικά, το CNN πέτυχε υψηλότερες επιδόσεις στην ταξινόμηση των δεδομένων του συνόλου FMA, σημειώνοντας μια αύξηση $\sim 5-6\%$ στην Accuracy, σε σχέση με το LSTM. Επιπλέον, οι επιδόσεις αυτές δεν εμφάνισαν τόσο υψηλή διακύμανση στα διαφορετικά είδη δεδομένων (mel,

fused και beat-synced mel), σε αντίθεση με τις αντίστοιχες του LSTM, όπου υπήρχαν αποκλίσεις της τάξης του 4%.

4 Εκτιμηση Συναισθηματος μεσω Παλινδρομησης

Το περιεχόμενο αυτής της ενότητας αφορά τα Βήματα 8-10 της εργασίας.

Προχωρώντας στο δεύτερο μέρος της εργαστηριακής άσκησης, οι κλάσεις των LSTM και CNN που αναπτύχθηκαν κατά το πρώτο μέρος γενικεύτηκαν, ώστε τα δίκτυα να μπορούν να ανταπεξέλθουν και σε προβλήματα παλινδρόμησης (Βήμα 8). Συγκεκριμένα, αλλάζοντας τη συνάρτηση κόστους τους και προσαρμόζοντας λίγο την αρχιτεκτονική τους (π.χ. διαστάσεις της τελικής εξόδου), τα δίκτυα τροποποιήθηκαν ώστε να μπορούν να προβλέπουν τιμές για τους συναισθηματικούς δείκτες valence (θετικό ή αρνητικό συναίσθημα), energy (ισχύς συναισθήματος) και danceability (πόσο χορευτικό είναι το τραγούδι). Τα διαθέσιμα δεδομένα[1] χωρίστηκαν σε δεδομένα εκπαίδευσης, επικύρωσης και αξιολόγησης σε αναλογία 60:25:15, μιας και για το συγκεκριμένο πρόβλημα οι επισημειώσεις των πραγματικών δεδομένων αξιολόγησης δεν ήταν διαθέσιμες. Κατόπιν, εκπαιδεύτηκαν τρία μοντέλα (ένα για κάθε συναισθηματικό δείκτη) από κάθε είδος δικτύου (LSTM και CNN). Σε ό,τι αφορά τα CNN, οι τιμές των χαρακτηριστικών τους διατηρήθηκαν ίδιες με αυτές του προβλήματος ταξινόμησης, απλώς ως συνάρτηση κόστους χρησιμοποιήθηκε η L1Loss. Από την άλλη, λόγω της προηγουμένως κατώτερης επίδοσής τους, για τα LSTM δίκτυα πραγματοποιήθηκε πρώτα μια διερεύνηση πλέγματος (grid search), χωρίς τα δίκτυα να έρθουν σε επαφή με τα δεδομένα αξιολόγησης, προκειμένου να προσδιοριστούν οι βέλτιστες τιμές για ορισμένες υπερπαραμέτρους: το πλήθος των hidden layers, την πολλαπλότητα των stacked LSTMs, το ρυθμό εκμάθησης και τη συνάρτηση κόστους. Τα αποτελέσματα του grid search παρατίθενται στον Πίνακα 4.1, όπου φαίνονται επίσης και οι επιδόσεις των δικτύων στο πρόβλημα παλινδρόμησης, χρησιμοποιώντας ως μετρική τη συσχέτιση Spearman.

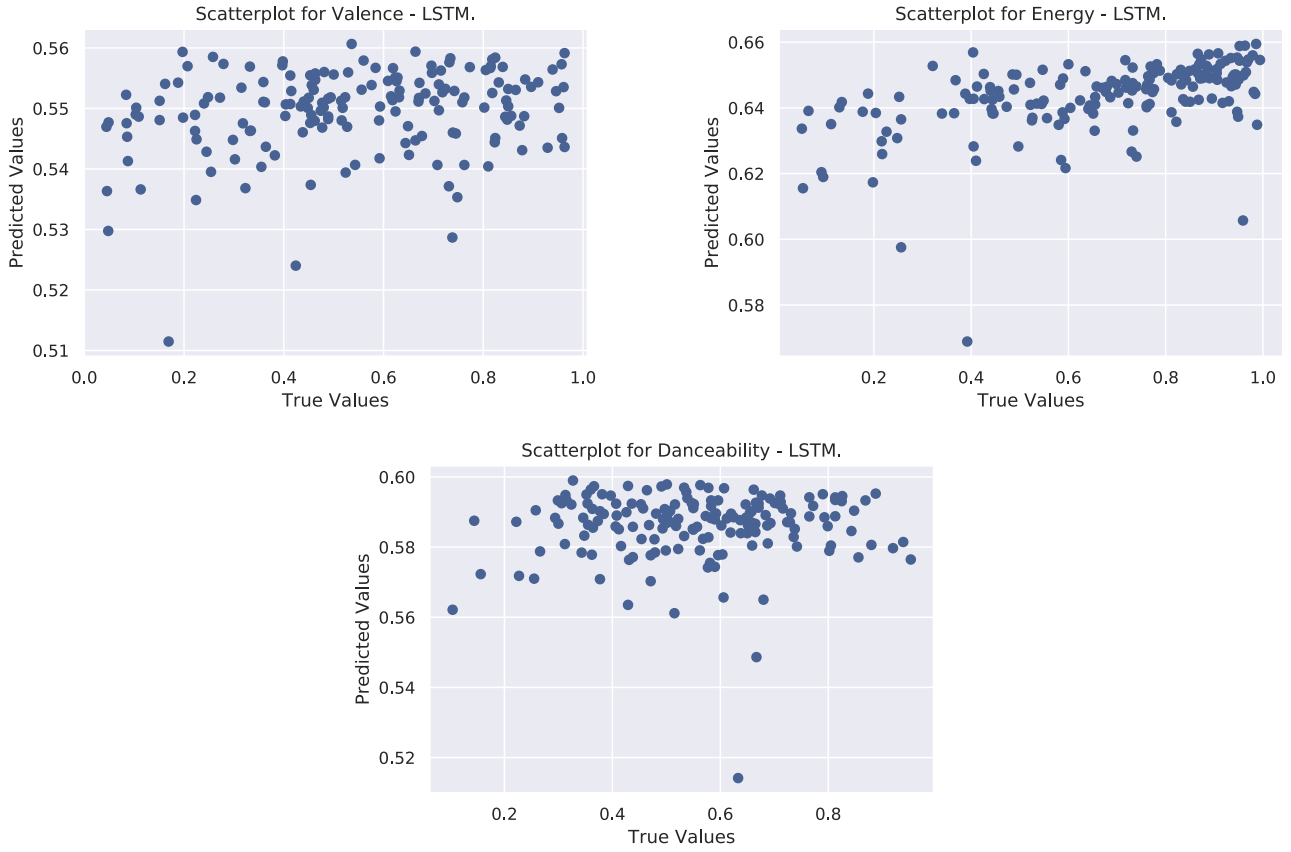
	NN-Type	Grid Search Specs	Spearman correlation (x100)	p-value
Valence	LSTM	600/5/10^{-5}/ `L1Loss`	21.76	0.007
	CNN		43.71	<0.001
Energy	LSTM	600/5/10^{-5}/ `SmoothL1Loss`	56.3	<0.001
	CNN		59.16	<0.001
Danceability	LSTM	600/5/10^{-6}/ `L1Loss`	4.14	0.59
	CNN		58.9	<0.001

Πίνακας 4.1: Συνοπτικά αποτελέσματα για το πρόβλημα παλινδρόμησης. Για την περίπτωση των LSTM πραγματοποιήθηκε ένα υποτυπώδες grid search επάνω στις υπερπαραμέτρους `rnn_size`, `num_layers`, `learning_rate` και `loss_function`. Στη στήλη Grid Search Specs του πίνακα παρατίθενται οι τιμές αυτών των παραμέτρων (με την ίδια σειρά) για τις οποίες προέκυψε το υψηλότερο validation score.

Σε ό,τι αφορά τα δίκτυα LSTM, είναι φανερό πως η υψηλότερη επίδοση σημειώθηκε στον υπολογισμό του χαρακτηριστικού energy, όπου παρατηρήθηκε επίδοση 56.3% με p-value (ένας

[1] Στην περίπτωση της παλινδρόμησης αξιοποιήθηκαν μόνο τα beat-synced mel-φασματογραφήματα, αφού συνδύαζαν υψηλές επιδόσεις στο πρόβλημα ταξινόμησης και για τους δύο τύπους δικτύων με σημαντικά μειωμένους χρόνους εκπαίδευσης.

δείκτης εμπιστοσύνης του αποτελέσματος) μικρότερο του ορίου σημαντικότητας 0.001. Η επίδοσή του δεν ήταν αντίστοιχη και στα δύο άλλα χαρακτηριστικά, με το danceability να αντιστοιχεί σε score μόλις 4.14%. Αξίζει, βέβαια, να σημειωθεί πως η επίδοση του LSTM εμφάνιζε πολύ σημαντικές διακυμάνσεις από πείραμα σε πείραμα, γεγονός που αποτυπώνεται και από το πολύ υψηλό p-value του αποτελέσματος για τη συσχέτιση Spearman στο danceability (υπήρχαν πειράματα στα οποία το score στο danceability ήταν της τάξης του $\sim 20\%$). Προκειμένου η επίδοση του LSTM να αξιολογηθεί συνολικά και για τους τρεις συναισθηματικούς δείκτες, η μέση συσχέτιση Spearman υπολογίστηκε ως ο απλός μέσος όρος των επί μέρους και βρέθηκε ίση με 27.49%. Στην Εικόνα 4.1 απεικονίζονται τα διαγράμματα διασποράς των τριών LSTM για τις προβλέψεις καθενός εκ των τριών συναισθηματικών δεικτών.

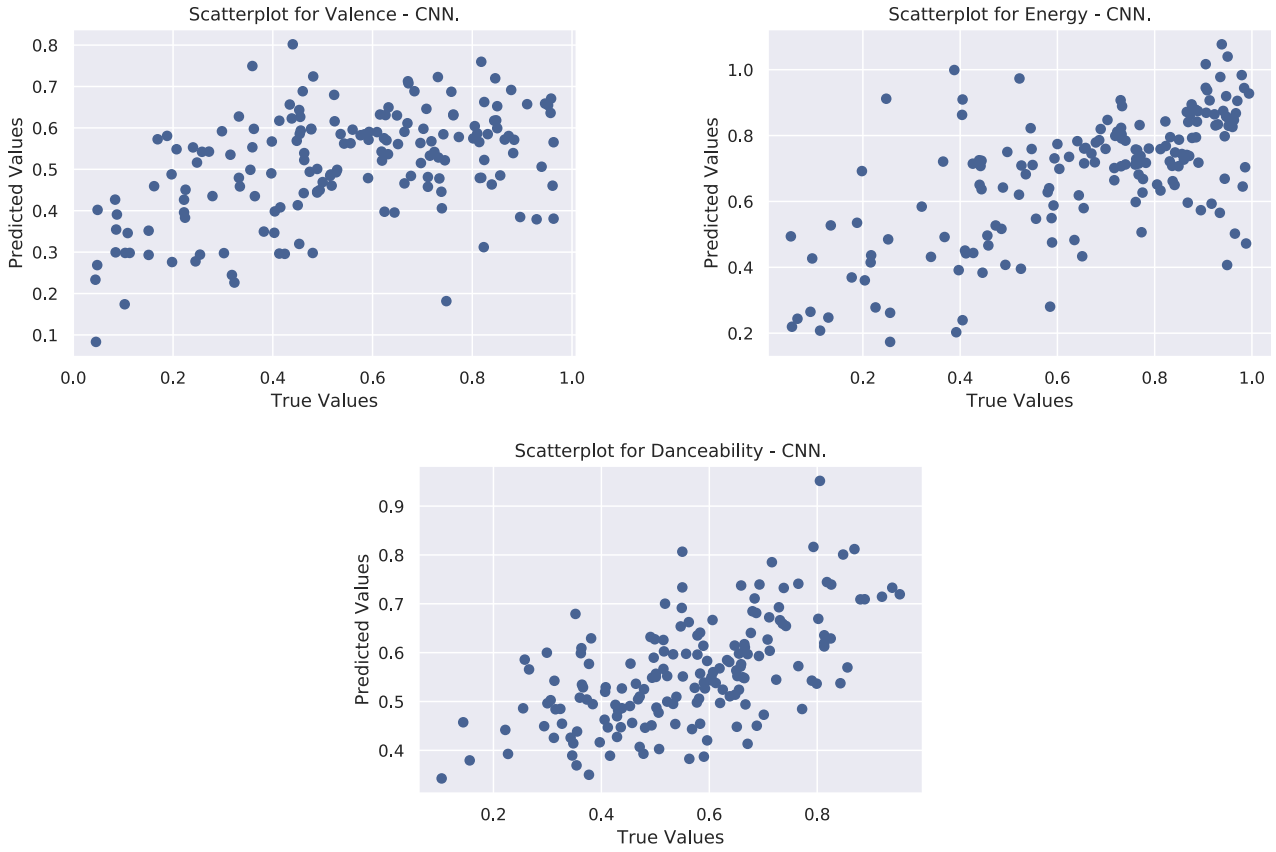


Εικόνα 4.1: Διαγράμματα διασποράς για τις προβλέψεις των LSTM που εκπαιδεύτηκαν για το πρόβλημα παλινδρόμησης, με τον οριζόντιο και τον κατακόρυφο άξονα να αντιστοιχούν στις πραγματικές και τις προβλεπόμενες από τα δίκτυα τιμές των χαρακτηριστικών, αντίστοιχα.

Είναι άξιο αναφοράς το γεγονός πως τα LSTM δεν κατάφεραν να προβλέψουν τιμές μεγαλύτερες του 0.6 για οποιοδήποτε από τα τρία χαρακτηριστικά, με μια συσσώρευση να παρατηρείται στο διάστημα [0.5,0.6].

Από την άλλη, όπως φαίνεται από τα αποτελέσματα του Πίνακα 4.1, οι επιδόσεις των CNN ήταν σημαντικά υψηλότερες (και συνεπέστερες από πείραμα σε πείραμα, παρότι κι εκεί παρατηρήθηκαν διακυμάνσεις) από αυτές των LSTM και για τους τρεις συναισθηματικούς δείκτες. Τα score των δικτύων CNN συνοδεύονταν σε κάθε περίπτωση από p-values μικρότερα του ορίου 0.001, με τα score για τα χαρακτηριστικά energy και danceability να φτάνουν σχεδόν το 60%. Και σε αυτήν την περίπτωση, αν και με μικρή διαφορά, ο συναισθηματικός δείκτης που προσεγγίστηκε καλύτερα από το δίκτυο ήταν το energy, ενώ συνολικά τα CNN είχαν μέση συσχέτιση Spearman ίση με 53.92%, δηλαδή σχεδόν διπλάσια από την αντίστοιχη των LSTM.

Πρόσθετα, όπως φαίνεται και από τα διαγράμματα διασποράς της Εικόνας 4.2, οι προβλέψεις των CNN δεν περιορίστηκαν σε ένα μικρό διάστημα, αλλά εν γένει σε ολόκληρο το εύρος [0,1]. Τέλος, αξίζει να αναφερθεί πως η κατανομή των σημείων στα διαγράμματα διασποράς των CNN ήταν πολύ πιο κοντά σε μια κατανομή τύπου $y = x$ (η οποία θα αντιστοιχούσε σε απολύτως ακριβείς εκτιμήσεις) από ότι στα διαγράμματα διασποράς των LSTM.



Εικόνα 4.2: Διαγράμματα διασποράς για τις προβλέψεις των CNN που εκπαιδεύτηκαν για το πρόβλημα παλινδρόμησης, με τον οριζόντιο και τον κατακόρυφο άξονα να αντιστοιχούν στις πραγματικές και τις προβλεπόμενες από τα δίκτυα τιμές των χαρακτηριστικών, αντίστοιχα.

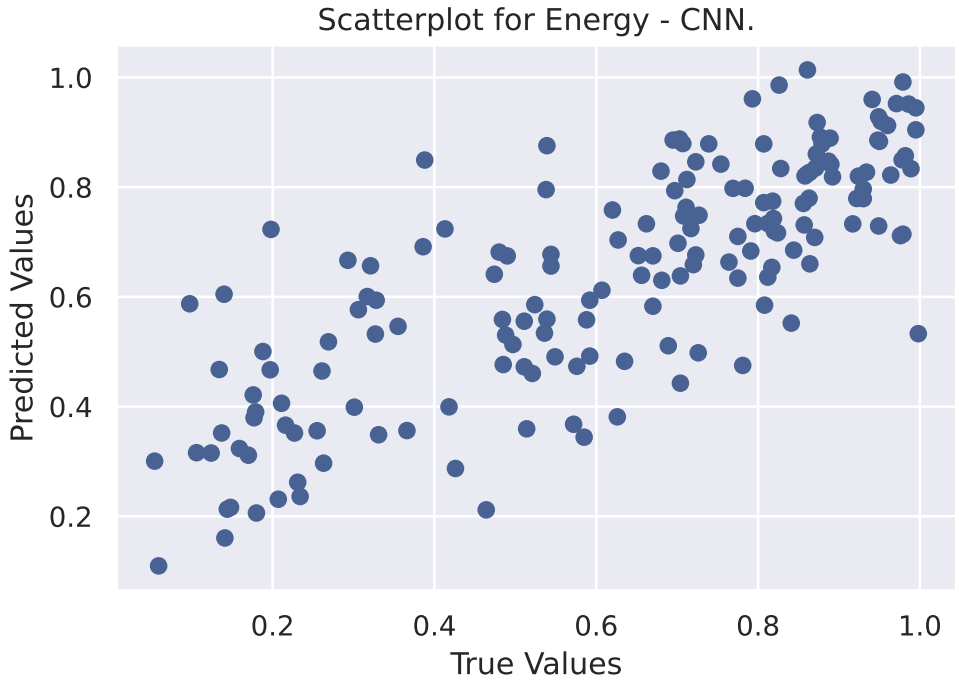
Έτσι, με ασφάλεια καταλήγει κανείς στο συμπέρασμα πως και στην περίπτωση της παλινδρόμησης τα CNN έχουν συνολικά υψηλότερες επιδόσεις, αυτή τη φορά, μάλιστα, με μεγάλη διαφορά.

Ένας από τους ποικίλους παράγοντες στους οποίους θα μπορούσαν ενδεχομένως να αποδοθούν οι μη-ιδανικές επιδόσεις των δικτύων (ειδικά των LSTM), είναι το περιορισμένο πλήθος δειγμάτων του συνόλου δεδομένων. Παρότι το σύνολο αποτελείται από σχεδόν 1500 δείγματα, μόλις το 60% αυτών (περίπου 900 δείγματα) αξιοποιείται για την εκπαίδευση των δικτύων, αφού τα υπόλοιπα είναι απαραίτητα για τη διαδικασία επικύρωσης, καθώς και για την αξιολόγηση. Παρότι 900 δείγματα θα επαρκούσαν στην περίπτωση άλλων ταξινομητών ή μοντέλων παλινδρόμησης, γενικά στη βαθιά μάθηση δεν είναι σπάνια ακόμα και η αξιοποίηση δεκάδων χιλιάδων δεδομένων για την εκπαίδευση δικτύων, όποτε αυτά είναι διαθέσιμα, καθώς ο συνδυασμός πολλών παραμέτρων (το οποίο ισχύει σχεδόν εξ ορισμού στη βαθιά μάθηση) με περιορισμένο αριθμό δεδομένων οδηγεί συχνά σε φαινόμενα υπερπροσαρμογής. Το πρόβλημα των περιορισμένων δεδομένων μπορεί να παρακαμφθεί εάν κανείς στραφεί στην έννοια του transfer learning [YCBL].

Το transfer learning ξεκινά με την εκπαίδευση ενός βασικού (base) δικτύου σε ένα αρχικό σύνολο υψηλού αριθμού δεδομένων για την επιτέλεση ενός συγκεκριμένου σκοπού (task). Κα-

τόπιν, οι εκπαιδευμένες παράμετροί του διαμορφώνουν αυτούσιες τον κορμό ενός δεύτερου δικτύου, το οποίο επεκτείνεται με επιπλέον επίπεδα και εκπαιδεύεται στο επιθυμητό σύνολο δεδομένων για το επιθυμητό task. Η εκπαίδευση αυτή πραγματοποιείται είτε ανανεώνοντας περαιτέρω τις παραμέτρους που εκπαιδεύτηκαν στο πρώτο σύνολο (fine-tuning), είτε κρατώντας τες σταθερές και ανανεώνοντας μόνο τα βάρη των νέων επιπέδων (freezing). Εφόσον τα δύο tasks είναι όμοια και εφόσον ο βασικός κορμός του δικτύου και η επέκτασή του εκπαιδεύονται στην αναγνώριση πιο γενικών (general) και πιο ειδικών (specific) χαρακτηριστικών, αντιστοίχως, η διαδικασία transfer learning οδηγεί στην ανάπτυξη δικτύων που γενικεύουν ευκολότερα και δεν πραγματοποιούν έντονο overfitting, με αποτέλεσμα να αποδίδουν συνολικά καλύτερα.

Στο πλαίσιο αυτό, αποφασίστηκε η εφαρμογή μεθόδων transfer learning (Βήμα 9) χρησιμοποιώντας ως κορμό το δίκτυο CNN που εκπαιδεύτηκε στα δεδομένα του συνόλου FMA, προκειμένου να διερευνηθεί το ενδεχόμενο βελτίωσης της επίδοσής του στο πρόβλημα παλινδρόμησης. Το CNN επιλέχθηκε έναντι του LSTM, λόγω της υψηλής διακύμανσης που παρουσίασε το δεύτερο στο πρόβλημα παλινδρόμησης, καθώς και της συνολικά χειρότερης επίδοσής του (βλ. Πίνακα 4.1). Δεδομένου πως το σύνολο FMA περιλαμβάνει πάνω από διπλάσια δεδομένα σε σχέση με το δεύτερο σύνολο δεδομένων, η εφαρμογή transfer learning είναι δικαιολογημένη. Επιπλέον, παρότι τα δύο tasks διαφέρουν αρκετά (στην πρώτη περίπτωση πραγματοποιείται ταξινόμηση τραγουδιών ενώ στη δεύτερη παλινδρόμηση για την εκτίμηση συναισθηματικών δεικτών), εάν κανείς θεωρήσει πως τα συνελικτικά επίπεδα εκπαιδεύονται για την αναγνώριση των general χαρακτηριστικών, δε θα έπρεπε να περιμένει μείωση της απόδοσης εάν η επέκταση για το task παλινδρόμησης πραγματοποιηθεί στα FC επίπεδα.

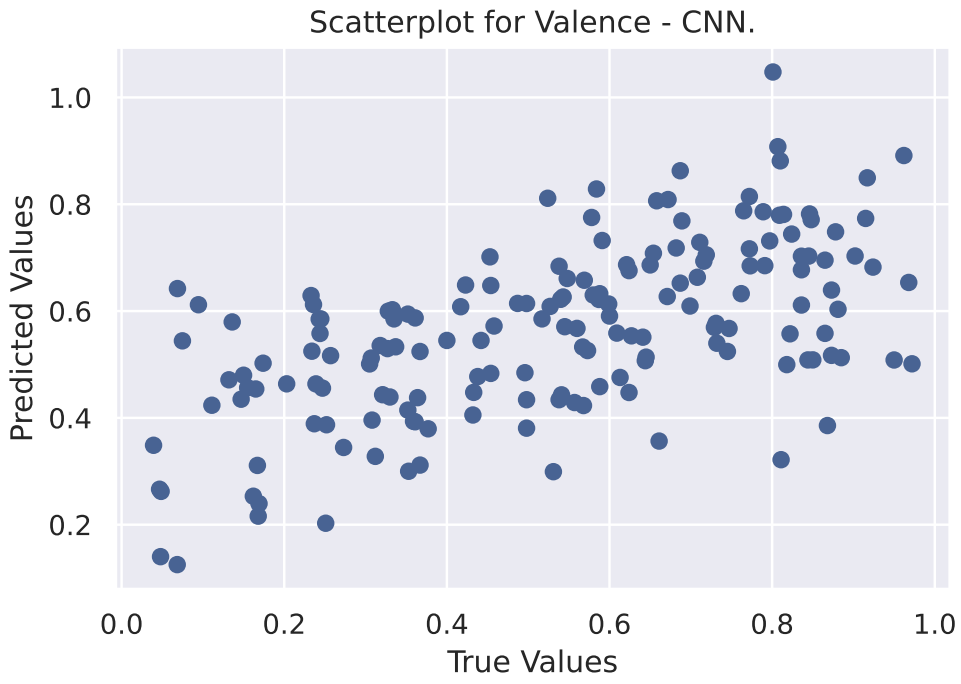


Εικόνα 4.3: Διάγραμμα διασποράς για τις προβλέψεις του CNN που εκπαιδεύτηκε μέσω transfer learning για τον προσδιορισμό του συναισθηματικού δείκτη energy.

Έτσι, από το CNN που εκπαιδεύτηκε στα δεδομένα του συνόλου FMA διατηρήθηκαν όλες οι παράμετροι εκτός από τα βάρη των 2 τελευταίων FC επιπέδων, ενώ η διάσταση του τελευταίου FC επιπέδου μετατράπηκε από 10 σε 1, προκειμένου να πραγματοποιείται παλινδρόμηση. Ως προς προσδιορισμό συναισθηματικός δείκτης επιλέχθηκε ο energy, στον οποίο το Spearman Correlation Score ήταν 58.9% για το CNN που εκπαιδεύτηκε αυστηρά για παλινδρόμηση. Αρ-

χικά, επιχειρήθηκε το transfer learning μέσω freezing, δηλαδή κρατώντας παγωμένες τις τιμές όλων των βαρών με εξαίρεση αυτά των τελευταίων δύο επιπέδων. Η τεχνική αυτή δεν απέφερε ικανοποιητικά αποτελέσματα, συνεπώς η διαδικασία επαναλήφθηκε μέσω fine tuning, δηλαδή επιτρέποντας σε όλα τα βάρη να ανανεωθούν κατά την εκπαίδευση του δικτύου στο επιθυμητό σύνολο δεδομένων. Οι προβλέψεις που πραγματοποίησε το CNN για το δείκτη energy φαίνονται στο διάγραμμα διασποράς της Εικόνας 4.3. Η συσχέτιση Spearman για την επίδοση του δικτύου υπολογίστηκε ίση με 76.8% με p-value μικρότερο από 0.001, πράγμα το οποίο κατέστησε πλήρως επιτυχημένη την απόπειρα εφαρμογής transfer learning στο πρόβλημα παλινδρόμησης.

Μάλιστα, δεδομένης της επιτυχίας της μεθόδου transfer learning στον προσδιορισμό του energy, διερευνήθηκε το ενδεχόμενο βελτίωσης και στον προσδιορισμό του valence, το οποίο ήταν ο συναισθηματικός δείκτης με τη μικρότερη συσχέτιση Spearman σε ό,τι αφορά το CNN που εκπαιδεύτηκε για παλινδρόμηση (μόλις 43.71%). Ακολουθώντας τα ίδια ακριβώς βήματα, το CNN που εκπαιδεύτηκε μέσω transfer learning πραγματοποίησε για το valence τις προβλέψεις που απεικονίζονται στο διάγραμμα διασποράς της Εικόνας 4.4.

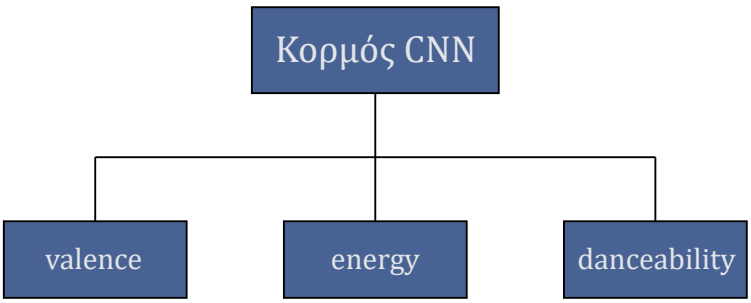


Εικόνα 4.4: Διάγραμμα διασποράς για τις προβλέψεις του CNN που εκπαιδεύτηκε μέσω transfer learning για τον προσδιορισμό του συναισθηματικού δείκτη valence.

Πράγματι, και στην περίπτωση του valence το transfer learning απέδωσε καρπούς, αφού αύξησε τη συσχέτιση Spearman στην τιμή 57.04% με p-value μικρότερο από το όριο 0.001.

Το τελικό μέρος της εργαστηριακής άσκησης και κατ' επέκταση της παρούσας εργαστηριακής αναφοράς αποτέλεσε η εκπαίδευση ενός δικτύου CNN με multitask learning [MTSK]. Τα CNN που εκπαιδεύτηκαν, είτε αμιγώς για παλινδρόμηση, είτε μέσω transfer learning, είχαν σε κάθε περίπτωση ένα task: τον προσδιορισμό τιμών για έναν συναισθηματικό δείκτη τη φορά. Αυτός ήταν και ο λόγος για τον οποίο στο Βήμα 8 αναπτύχθηκαν 3 διαφορετικά CNN (ένα για κάθε δείκτη) ή για τον οποίο στο Βήμα 9 αναπτύχθηκαν 2 διαφορετικά CNN (ένα για το energy και ένα για το valence). Μια διαφορετική προσέγγιση είναι η κατασκευή ενός μοντέλου ικανό να προβλέπει ταυτόχρονα και τους τρεις συναισθηματικούς δείκτες. Τα ωφέλη αυτής της προσέγγισης δεν περιορίζονται μόνο στην εξοικονόμηση χώρου και χρόνου, καθώς τη θέση τριών μοντέλων παίρνει μόνο ένα, αλλά επεκτείνονται θεωρητικά και στην απόδοση: όταν ένα μοντέλο

εκπαιδεύεται ταυτόχρονα για διαφορετικά tasks, τα οποία όμως είναι συσχετισμένα μεταξύ τους, αναμένεται να μάθει καλύτερα τα σημαντικά χαρακτηριστικά στα οποία πρέπει να δώσει έμφαση, αφού αυτά είναι λογικό να αφορούν περισσότερα από ένα tasks. Για παράδειγμα, κάποιο χαρακτηριστικό των δεδομένων που οδηγεί σε υψηλό valence (πολύ θετικό συναίσθημα) είναι συχνά αναμενόμενο να οδηγεί και σε υψηλό danceability. Έτσι, θα ανέμενε κανείς μια προσέγγιση μέσω multitask learning να βελτιώσει τα αποτελέσματα του Βήματος 8. Για το σκοπό αυτό, πραγματοποιήθηκε η υλοποίηση (Βήμα 10) που αναπαρίσταται στην Εικόνα 4.5.



Εικόνα 4.5: Απεικόνιση δομής του μοντέλου CNN για multitask learning.

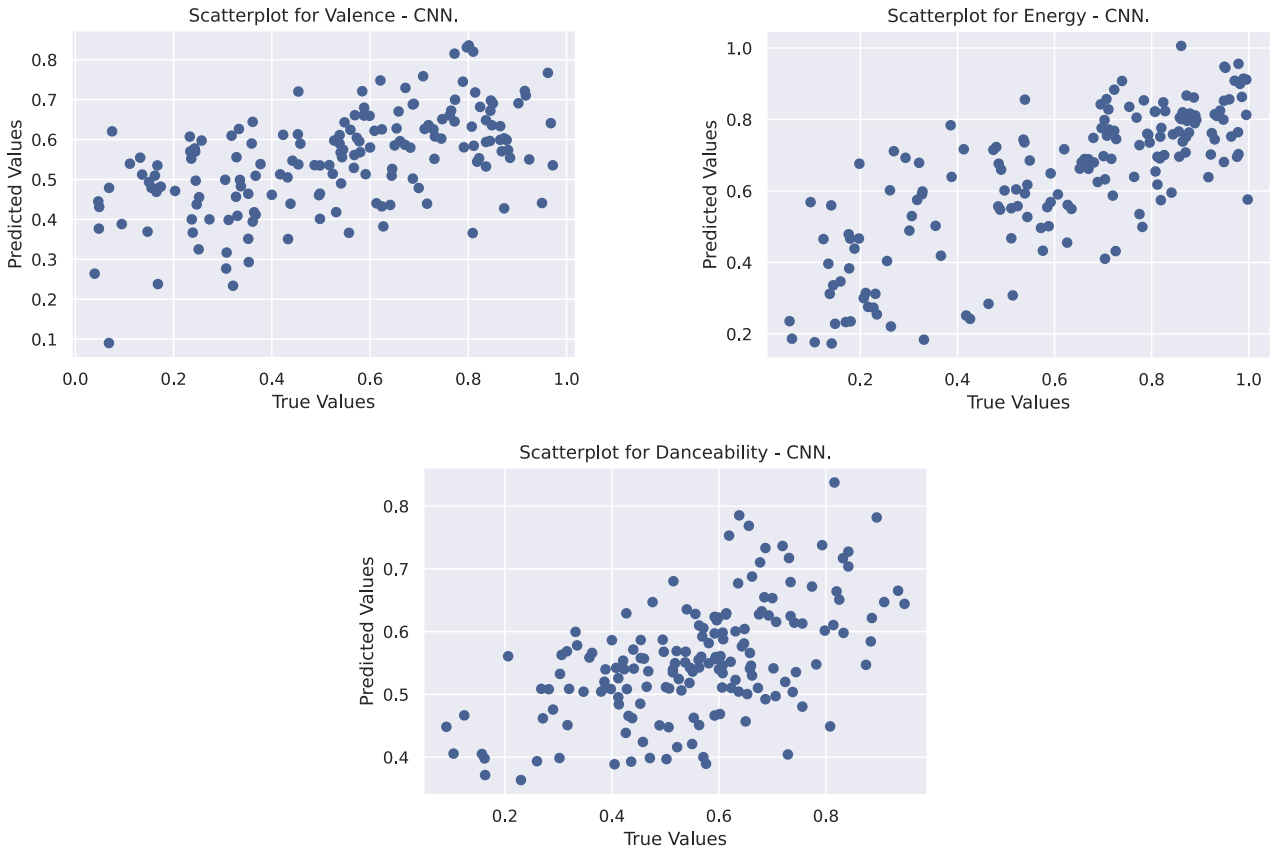
Συγκεκριμένα, το μοντέλο που κατασκευάστηκε αποτελούταν από ένα βασικό κορμό με την αρχιτεκτονική και τις παραμέτρους που είχαν τα επί μέρους CNN του Βήματος 8. Με το πέρας των συνελίξεων και των max pooling, πραγματοποιούταν κατ' ουσίαν μια διακλάδωση, σε τρία κλαδιά πανομοιότυπης αρχιτεκτονικής: 3 επίπεδα FC, με το πρώτο να έχει έξοδο 128, το δεύτερο να έχει έξοδο 64 και το τρίτο να έχει έξοδο 1. Ανάμεσα στα επίπεδα αυτά παρεμβάλλονταν μη γραμμικότητες τύπου ReLU, καθώς και ενδεχόμενα dropout με πιθανότητα 15%. Συνολικά, το δίκτυο είχε έξοδο διάστασης 3, δηλαδή 1 για κάθε συναισθηματικό δείκτη προς προσδιορισμό.

Η θεμελιώδης διαφορά με τα τρία επί μέρους δίκτυα του Βήματος 8 ήταν πως το συγκεκριμένο δίκτυο εκπαιδευόταν ως ένα ενιαίο σύνολο, με αποτέλεσμα η εκπαίδευσή του ως προς κάθε δείκτη να επηρεάζει την εκπαίδευσή του ως προς τους άλλους δύο. Για την εν λόγω εκπαίδευση κατέστη απαραίτητη η ανάπτυξη μιας νέας συνάρτησης κόστους, ούτως ώστε η προαναφερθείσα επιρροή να πραγματοποιείται μέσω του βήματος οπισθοδιάδοσης. Η συνάρτηση κόστους που ορίστηκε δεν ήταν παρά ένας σταθμισμένος μέσος όρος τριών επί μέρους MSELoss - μία για κάθε συναισθηματικό δείκτη. Τα βάρη επιλέχθηκαν ίσα με 2/3, 1/6 και 1/6 για τους δείκτες valence, energy και danceability, αντίστοιχα. Ο λόγος για αυτό ήταν πως σε όλα τα προηγούμενα πειράματα η χαμηλότερη επίδοση σημειωνόταν στον προσδιορισμό του δείκτη valence, επομένως κρίθηκε σκόπιμο εκεί να δοθεί μεγαλύτερο βάρος.

Αφότου το συγκεκριμένο μοντέλο εκπαιδεύτηκε στο σύνολο των δεδομένων που είχαν αξιοποιηθεί και στις προηγούμενες περιπτώσεις, αξιολογήθηκε στο αντίστοιχο σύνολο αξιολόγησης σε δύο άξονες: αφενός βάσει της συσχέτισης Spearman κάθε συναισθηματικού δείκτη ξεχωριστά και αφετέρου βάσει της μέσης συσχέτισης Spearman και για τους τρεις συναισθηματικούς δείκτες. Στην Εικόνα 4.6 απεικονίζεται το διάγραμμα διασποράς με τις προβλέψεις του δικτύου για τα δεδομένα αξιολόγησης, ενώ στον παρακάτω πίνακα (Πίνακας 4.2) συνοψίζονται τα αποτελέσματα αξιολόγησης που αντιστοιχούν στα διαγράμματα αυτά (όλα με p-value < 0.001).

	Valence	Energy	Danceability	Total
Spearman Correlation	56.71%	76.14%	54.16%	62.33%

Πίνακας 4.2: Αποτελέσματα για την επίδοση του CNN που εκπαιδεύτηκε με multitask learning.



Εικόνα 4.6: Διαγράμματα διασποράς για τις προβλέψεις του CNN που εκπαιδεύτηκε με multitask learning.

Παρότι η επίδοση για την πρόβλεψη του danceability μειώθηκε ελαφρώς σε σχέση με το αντίστοιχο μοντέλο του Βήματος 8, το multitask learning πράγματι απέφερε τους αναμενόμενους καρπούς, αφού η σημαντική αύξηση στην επίδοση για την πρόβλεψη των valence και energy οδήγησε σε μια συνολική αύξηση $\sim 8.5\%$ της μέσης συσχέτισης Spearman.

Αξίζει να σημειωθεί πως τα παραπάνω αποτελέσματα δεν αποτελούν προϊόν τυχαιότητας, αλλά είναι αντιπροσωπευτικά ως προς την πραγματική κατάσταση. Το γεγονός αυτό αποδείχθηκε μέσω υποβολής του εν λόγω μοντέλου (Βήμα 11) στον διαγωνισμό του Kaggle, όπου η αναμενόμενη επίδοση (στο 30% των πραγματικών δεδομένων αξιολόγησης) εκτιμήθηκε περίπου ίση με 61.32% σε ό,τι αφορά τη μέση συσχέτιση Spearman.

Αναφορες

[Cal20] O. Calin, *Deep Learning Architectures*, Springer, 2020. ISBN: 978-3-030-36721-3 Cited on p. 12

[KT08] K. Koutroumbas, S. Theodoridis, *Pattern Recognition*, Academic Press, 2008. ISBN: 978-1-59749-272-0

[LAB2] N. Stamatis, S. Rigas, *Αναγνώριση Φωνής με Κρυφά Μαρκοβιανά Μοντέλα και Αναδρομικά Νευρωνικά Δίκτυα*, DSML, 2021. 4

[YCBL] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, *How transferable are features in deep neural networks?*, Advances in Neural Information Processing Systems 27, 2014. 21

[MTSK] L. Kaiser, A. N. Gomez, N. Shazeer, A. Vaswani, N. Parmar, L. Jones, J. Uszkoreit, *One Model To Learn Them All*, 2017. 23