Εθνικό Μετσόβιο Πολυτεχνείο
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Αναγνώριση Προτύπων

Ροή Σ: Σήματα, Έλεγχος, Ρομποτική

9ο Εξάμηνο

3η Εργαστηριακή Άσκηση: Αναγνώριση Είδους και Εξαγωγή Συναισθήματος από Μουσική

Χρήστος Δημόπουλος - 03117037
chrisdim1999@gmail.com

Ιανουάριος, 2022

Contents

Θέμα Εργαστηριακής Άσκησης

Σκοπός της άσκησης είναι η αναγνώριση του είδους και η εξαγωγή συναισθηματικών διαστάσεων από φασματογραφήματα (spectrograms) μουσικών κομματιών. Δίνονται δύο σύνολα δεδομένων, το Free Music Archive (FMA) genre με 3834 δείγματα χωρισμένα σε 20 κλάσεις (είδη μουσικής) και τη βάση δεδομένων (dataset) multitask music με 1497 δείγματα με επισημειώσεις (labels) για τις τιμές συναισθηματικών διαστάσεων όπως valence, energy και danceability. Τα δείγματα είναι φασματογραφήματα , τα οποία έχουν εξαχθεί από clips 30 δευτερολέπτων από διαφορετικά τραγούδια.
Στο προπαρασκευαστικό στάδιο, πραγματοποιείται πρώτα ανάλυση των δεδομένων και εξοικείωση με τα φασματογραφήματα και ύστερα κατασκευή ταξινομητών για το είδος της μουσικής πάνω στη βάση δεδομένων (dataset) FMA.

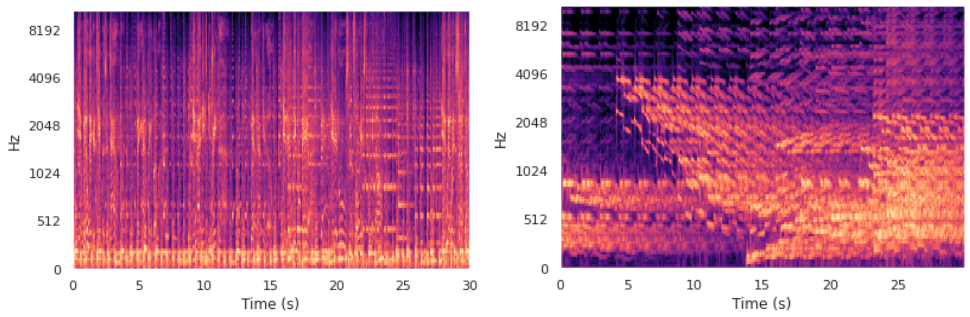
Βήματα Προπαρασκευής

1. Εξοικείωση με φασματογραφήματα στην κλίμακα mel

Αρχικά, επιλέγουμε δύο τυχαία δείγματα από ένα υποσύνολο του Free Music Archive (FMA) dataset, τα οποία έχουν διαφορετικές επισημειώσεις (labels):

- 1042.fused.full.npy.gz Blues
- 1325.fused.full.npy.gz Electronic

Στο Σχήμα 1, απεικονίζουμε τα φασματογραφήματα των δύο δειγμάτων σε κλίμακα mel, με τον οριζόντιο άξονα να είναι ο χρόνος (sec) και τον κατακόρυφο η συχνότητα (Hz).



(a) (b)

Figure 1: Φασματογραφήματα κλίμακας mel για αρχείο μουσικής (a) Blues, (b) Electronic.

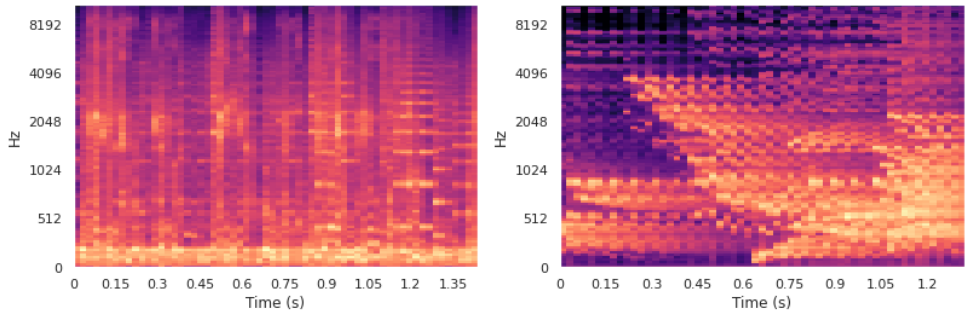
Εν γένει, τα φασματογραφήματα μας δίνουν πληροφορία για το φασματικό περιεχόμενο των μουσικών σημάτων σε ολισθαίνοντα χρονικά παράθυρα, ενώ η κλίμακα mel είναι μη γραμμική και σχετίζεται με την αντίληψη του ήχου από το ανθρώπινο αυτί. Στα δύο αυτά δείγματα, παρατηρούμε ότι το δείγμα μουσικής Blues έχει καθόλη τη διάρκεια του μουσικού αποσπάσματος σταθερό φασματικό περιεχόμενο και κυρίως χαμηλές συχνότητες. Αντιθέτως, το δεύτερο δείγμα που ανήκει στο είδος Electronic έχει απότομες μεταβολές συχνότητας καθόλη τη χρονική του διάρκεια.

2. Συγχρονισμός φασματογραφημάτων στο ρυθμό της μουσικής (beat-synced spectrograms)

Τυπώνοντας τις διαστάσεις των φασματογραφημάτων, παρατηρούμε ότι το πρώτο δείγμα (Blues) έχει **1293** χρονικά βήματα, ενώ το δεύτερο (Electonic) **1291**. Αν επιχειρούσαμε να εκπαιδεύσουμε ένα LSTM δίκτυο απευθείας πάνω σε δείγματα τέτοιων διαστάσεων, αφενός θα αργήσουμε και αφετέρου θα κινδυνεύουμε από φαινόμενα **vanishing/exploding gradient** λόγω του μεγάλου βάθους του δικτύου που θα δημιουργηθεί

από το μεγάλο πλήθος των timestamps. Ως εκ τούτου, επιβάλλεται η μείωση της διαστατικότητας των δειγμάτων.

Για τον σκοπό αυτό, συγχρονίζουμε καθένα από τα δείγματα πάνω στο ρυθμό της μουσικής, παίρνοντας τη διάμεσο (median) ανάμεσα στα σημεία που χτυπάει το beat της μουσικής. Απεικονίζοντας εκ νέου τα beat-synced φασματογραφήματα των δύο ίδιων δειγμάτων (Σχήμα 2), παρατηρεί κανείς ότι η συνολική διάρκεια των χρονικών βημάτων μειώνεται αισθητά. Ως αποτέλεσμα, η ευκρίνεια των φασματογραφημάτων μειώνεται, δίχως ωστόσο να αλλοιώνεται το χρονοσυχνοτικό περιεχόμενο τους, γεγονός που μας εξυπηρετεί στο πρόβλημα ταξινόμησης. Τα νέα χρονικά βήματα των δύο δειγμάτων γίνονται **62** και **57** αντίστοιχα.

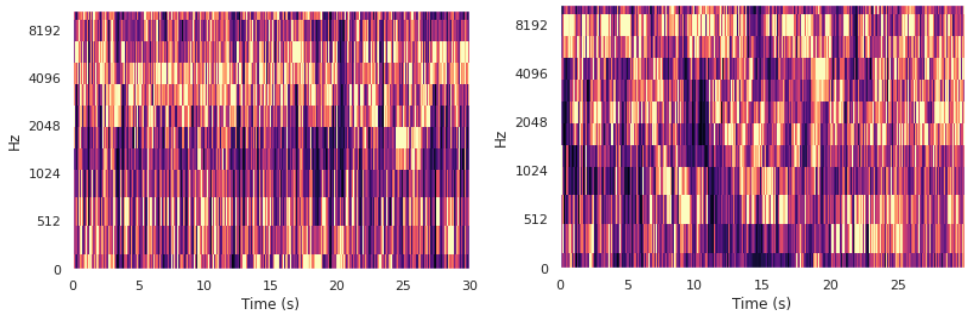


(a) (b)

Figure 2: Beat-Synced Φασματογραφήματα κλίμακας mel για αρχείο μουσικής (a) Blues, (b) Electronic.

3. Εξοικείωση με χρωμογραφήματα

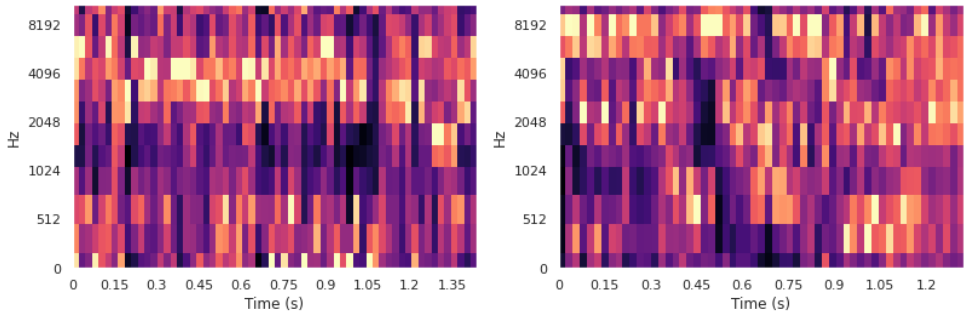
Στη συνέχεια, επαναλαμβάνουμε την ίδια διαδικασία για τα χρωμογραφήματα των δύο δειγμάτων μουσικής. Όπως και πριν, η μείωση των χρονικών βημάτων με βάση τον ρυθμό της μουσικής μειώνει την ευκρίνεια των χρωμογραφημάτων, ωστόσο διατηρεί την χρονοσυχνοτική κατανομή τους.



(a) (b)

Figure 3: Χρωμογραφήματα για αρχείο μουσικής (a) Blues, (b) Electronic.

Παρόλα αυτά, πρατηρεί κανείς ότι αν και πρόκειται για δύο αρκετά διαφορετικά μουσικά είδη, τα χρωματογραφήματα των δύο δειγμάτων δεν διαφέρουν σημαντικά, όπως δηλαδή συνέβαινε με τα φασματογραφήματα. Κάτι τέτοιο θα δυσχεραίνει το μοντέλο ταξινόμησης στο πρόβλημα κατηγοριοποίησης των δειγμάτων με βάση το μουσικό είδος.

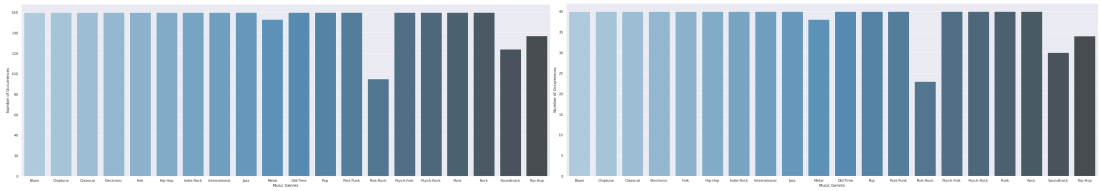


(a) (b)

Figure 4: Beat-Synced Χρωμογραφήματα για αρχείο μουσικής (a) Blues, (b) Electronic.

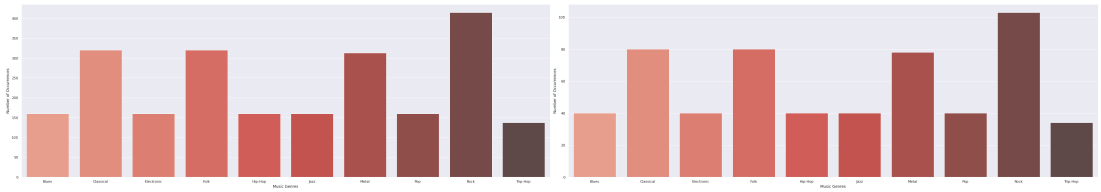
4. Φόρτωση και ανάλυση δεδομένων

Στη συνέχεια, χρησιμοποιώντας τον βοηθητικό κώδικα που παρέχεται, φορτώνονται τα δείγματα του dataset. Ως πρώτο στάδιο, πραγματοποιείται **class mapping** των δειγμάτων, δηλαδή συγχωνεύονται μουσικά είδη που είναι αρκετά παρεμφερή και διαγράφονται πλήρως άλλα είδη που είναι σπάνια και περισσότερο ασυσχέτιστα, ώστε να διευκολυνθεί περισσότερο η εκπαίδευση του ταξινομητή. Αποτέλεσμα αυτού είναι ένα πρόβλημα ταξινόμησης **10 κλάσεων**. Ως δεύτερο στάδιο, φροντίζεται ώστε όλα τα δείγματα να έχουν **ίδια διάσταση** από χαρακτηριστικά (επιλέγουμε 128 features για mel spectrograms και 12 για chromagrams). Για τον σκοπό αυτό, γίνεται zero-padding στα δείγματα με λιγότερα features και discarding σε αυτά με περισσότερα. Επιπλέον, το training set γίνεται **split σε train και validation** (με ποσοστό 1/3 το val), με το δεύτερο να χρησιμεύει στην προσαρμογή των υπερπαραμέτρων των ταξινομητών. Δημιουργούνται, τέλος, οι dataloaders με **BATCH_SIZE = 32** για το train και validation set, και **BATCH_SIZE = 16** για το test.



(a) (b)

Figure 5: Ιστόγραμμα των 20 κλάσεων πριν το Mapping στο (a) Train Set, (b) Test Set.



(a) (b)

Figure 6: Ιστόγραμμα των 10 κλάσεων μετά το Mapping στο (a) Train Set, (b) Test Set.

Παρατηρούμε ότι η κατανομή των δειγμάτων στα μουσικά είδη είναι πρακτικά ίδια ανάμεσα στο train και στο test set, τόσο πριν όσο και μετά το class mapping.

5. Αναγνώριση μουσικού είδους με LSTM

Στο σημείο αυτό, εκπαιδεύεται ένα Basic LSTM δίκτυο για την ταξινόμηση των δειγμάτων του FDA Dataset με βάση το μουσικό είδος. Στην πρώτη περίπτωση, το δίκτυο είναι **Bidirectional** με δύο κρυφά επίπεδα 64 νευρώνων το καθένα. Χρησιμοποιείται Dropout με πιθανότητα 50% και L2 Regularization. Το κριτήριο είναι Cross Entropy Loss, ενώ επιλέγεται Adam Optimizer με learning rate 10^{-3}.

Debugging Μοντέλου

Προκειμένου να επισπευθεί η διαδικασία ανάπτυξης και αποσφαλμάτωσης των μοντέλων, προστίθεται ως debugging method η επιλογή εκπαίδευσης πάνω σε ένα μικρό υποσύνολο από 4 batches για πολλές εποχές, ώστε το μοντέλο να κάνει overfit. Για παράδειγμα, υπερεκπαιδεύουμε το δίκτυο για 500 εποχές πάνω σε ένα μικρό υποσύνολο του training set από mel spectrograms, και παρατηρούμε ότι σε σχετικά σύντομο χρονικό διάστημα το training loss γίνεται μηδενικό (Σχήμα 7).

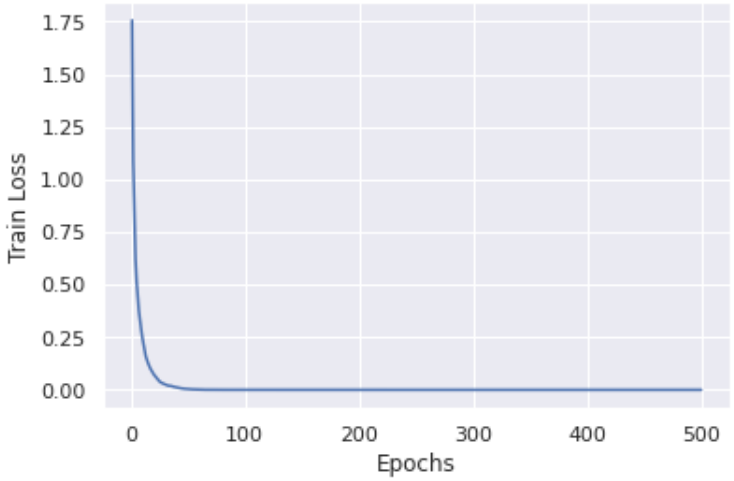


Figure 7: Train Loss μετά από 500 εποχές για Overfit Δίκτυο.

Εκπαίδευση με είσοδο mel spectrograms

Αρχικά, στην είσοδο του δικτύου τροφοδοτούμε τα mel spectrograms που εξάγαμε δίχως να μειώσουμε τη διαστατικότητα τους με βάση τον ρυθμό μουσικής. Όπως είναι φυσικό, η διαδικασία εκπαίδευσης είναι ιδιαίτερα **χρονοβόρα**, ενώ όπως φαίνεται και στο Σχήμα 8a, τόσο το training οσο και το valiation loss δεν φαίνεται να μειώνονται στο πέρασμα 50 εποχών, υποδηλώνοντας τον εγκλωβισμό σε κάποιο τοπικό ελάχιστο της συνάρτησης κόστους ή απλώς ατυχές initalization.

Εκπαίδευση με είσοδο beat-synced mel spectrograms

Επαναλαμβάνουμε τη διαδικασία εκπαίδευσης, θέτοντας στην είσοδο του δικτύου αυτή τη φορά τα beat-synced mel spectrograms αισθητά μικρότερης διαστατικότητας. Παρατηρούμε ότι η διαδικασία εκπαίδευσης επιταχύνεται αισθητά, ενώ παρατηρούμε μείωση για το train και validation loss (Σχήμα 8b). Τερματίζουμε στις 150 εποχές, ώστε να αποφύγουμε φαίνομενα overfitting, που ήδη πάνε να κάνουν την εμφάνιση τους.

Εκπαίδευση με είσοδο beat-synced chromagrams

Έπειτα, θέτουμε στην είσοδο του δικτύου τα beat-synced chromagrams των δειγμάτων. Παρατηρούμε ότι από πάρα πολύ νωρίς, το μοντέλο μας παθαίνει **overfit**, καθώς το train loss μειώνεται και το validation

loss εκτοξεύεται σε υψηλότερες τιμές από την αρχική (Σχήμα 8c). Ως εκ τούτου, η επιλογή τις εξής εισόδου φαντάζει απαγορευτική για τις συγκεκριμένες υπερπαραμέτρους.

Εκπαίδευση με είσοδο beat-synced mel spectrograms και chromagrams (fused)

Τέλος, θέτουμε στην είσοδο του δικτύου τα beat-synced mel spectrograms συνενωμένα με τα αντίστοιχα chromagrams. Η εκπαίδευση του δικτύου γίνεται κανονικά, με τις τιμές του validation loss, ωστόσο, να παρουσιάζουν υψηλές διακυμάνσεις μέχρι τη σύγκλιση (Σχήμα 8d).

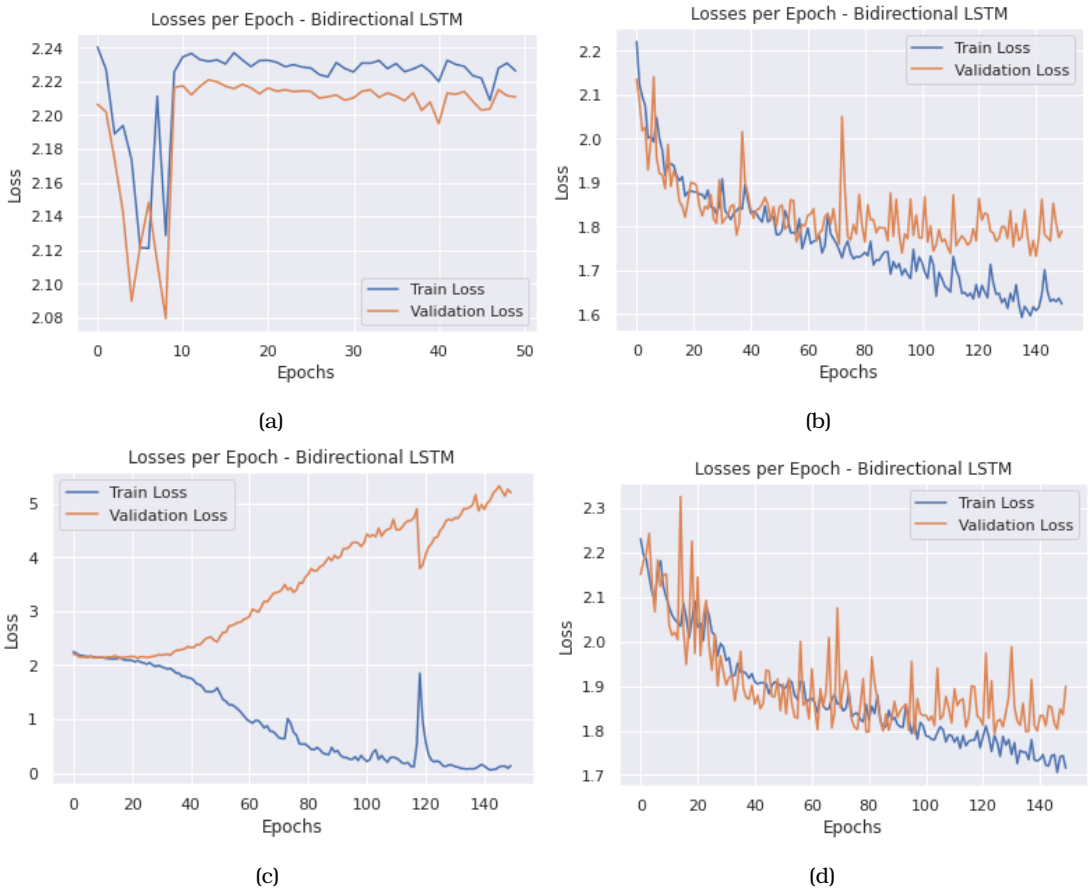


Figure 8: Εκπαίδευση LSTM Δικτύου (lr = 10^{-3}) με είσοδο (a) mel-Spectrograms, (b) Beat-Synced mel-Spectrograms, (c) Beat-Synced Chromagrams, (d) Beat-Synced Fused mel-Spectrograms Chromagrams.

Σε δεύτερο στάδιο, προκειμένου να βελτιώσουμε τις επιδόσεις του μοντέλου στο πρόβλημα ταξινόμησης και να αποφύγουμε φαινόμενα overfitting, ενσωματώνουμε τεχνικές **Early Stopping**, ενώ μειώνουμε τον ρυθμό μάθησης στην τιμή 10^{-4}, ώστε να αποφύγουμε εγκλωβισμούς σε τοπικά ελάχιστα. Τέλος, τροποποιούμε την αρχιτεκτονική του Bidirectional LSTM δικτύου αυξάνοντας τους νευρώνες κάθε κρυφού επιπέδου σε **256**. Όπως παρατηρούμε (Σχήμα 9) και για τις τέσσερεις περιπτώσεις εισόδου, το δίκτυο αποφεύγει το overfitting με πρόωρο τερματισμό της εκπαίδευσης, ενώ οι τιμές των losses συγκλίνουν σε καλύτερες τιμές από πριν.

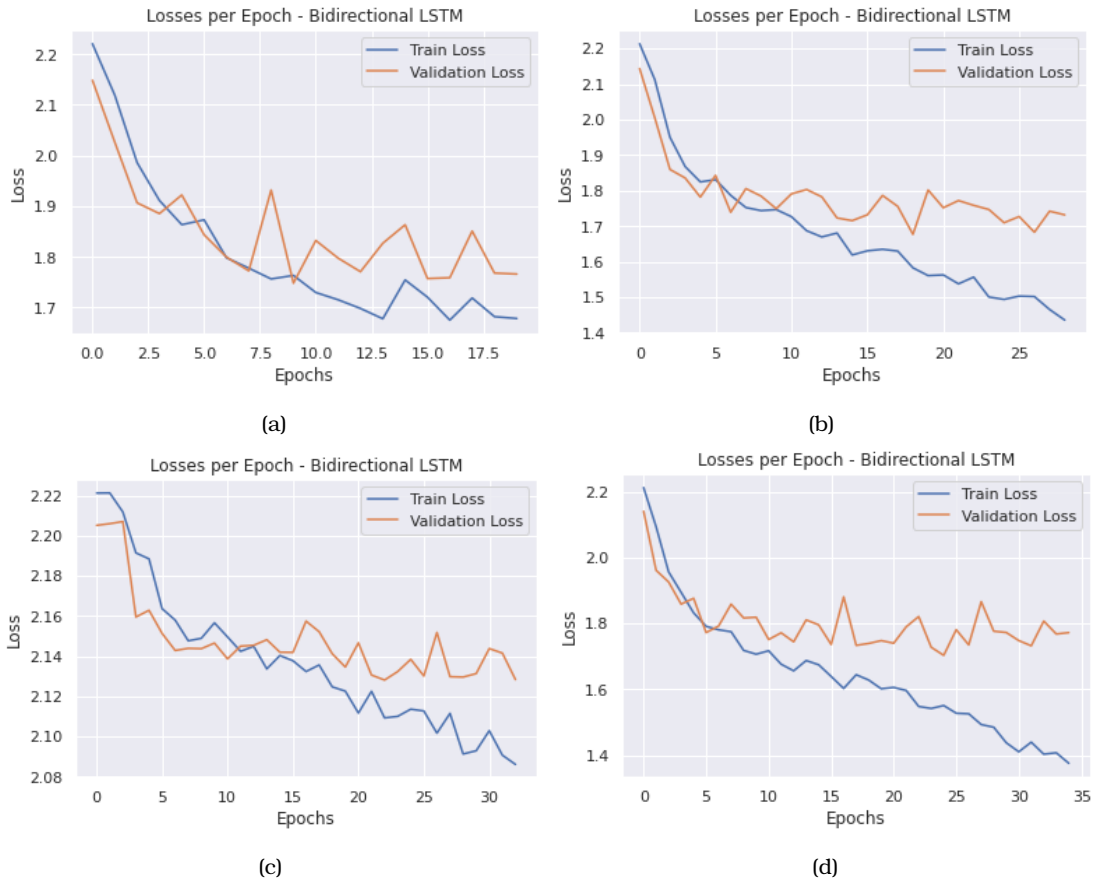


Figure 9: Early Stopping Εκπαίδευση LSTM Δικτύου (lr = 10^{-4}) με είσοδο (a) mel-Spectrograms, (b) Beat-Synced mel-Spectrograms, (c) Beat-Synced Chromagrams, (d) Beat-Synced Fused mel-Spectrograms Chromagrams.

6. Αξιολόγηση των μοντέλων

Στο τελευταίο προπαρασκευαστικό βήμα, αξιολογούμε την επίδοση του δικτύου για τις τέσσερεις πιθανές εισόδους, με βάση ορισμένες μετρικές. Η πιο συνηθισμένη εξ αυτών είναι το **Accuracy**, δηλαδή το ποσοστό των δειγμάτων που ταξινομήθηκαν ορθώς προς όλα τα δείγματα.

Εστιάζοντας σε καθεμία κλάση χωριστά, μπορούμε να ορίσουμε ως positive τα δείγματα που ανήκουν στην κλάση και ως negative αυτά που δεν ανήκουν. Ως εκ τούτου, για κάθε κλάση ορίζονται οι μετρικές:

- **Accuracy**: το ποσοστό ευστοχίας της κλάσης, δηλαδή:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**: το ποσοστό των δειγμάτων που ταξινομήθηκαν ορθώς στην κλάση προς όλα τα δείγματα που ταξινομήθηκαν σ'αυτήν:

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

- **Recall**: το ποσοστό των δειγμάτων που ταξινομήθηκαν ορθώς στην κλάση προς όλα τα δείγματα που όντως ανήκουν σ'αυτήν:

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

- **F1 Score**: ο αρμονικός μέσος μεταξύ Precision και Recall:

$$F1Score = 2\frac{Precision * Recall}{Precision + Recall}$$

Θέλοντας να ορίσουμε μετρικές που αφορούν όλες τις κλάσεις, διακρίνουμε τις εξής:

- **Macro-Averaged Metrics**: πρόκειται για το μέσο όρο των παραπάνω μετρικών (precision, recall, f1-score) πάνω σε όλες τις κλάσεις. Αν χρησιμοποιηθεί weighted macro-average, τότε πρόκειται για τον σταθμισμένο μέσο όρο, όπου τα βάρη κάθε μετρικής καθορίζονται με βάση τον αριθμό δειγμάτων που ανήκουν στην κάθε κλάση.
- **Micro-Averaged Metrics**: πρόκειται για τον υπολογισμό των παραπάνω μετρικών (precision, recall, f1-score), λαμβάνοντας υπόψη όλα τα δείγματα του dataset.

Στη γενικότερη περίπτωση, η μετρική **accuracy** χρησιμοποιείται όταν μας ενδιαφέρει κυρίως το ποσοστό των ορθώς ταξινομημένων δειγμάτων. Αντιθέτως, σε περιπτώσεις όπου δίνεται περισσότερο έμφαση στις εσφαλμένες ταξινομήσεις (πχ ιατρικά δεδομένα) προτιμώνται οι μετρικές **Precision** και **Recall**. Αν θέλουμε να αποφύγουμε False Negatives (πχ cancer detection) η μετρική Recall είναι περισσότερο ενδεικτική της απόδοσης, ενώ αν ο στόχος είναι η αποφυγή των False Positives προτιμάται το Precision. Η μετρική **F1-score**, από την άλλη, δίνει ίση σημασία στο Precision και στο Recall, ανάμεσα στα οποία υπάρχει συνήθως tradeoff, με αποτέλεσμα να μην ενδείκνυται για όλα τα προβλήματα ταξινόμησης.

Έπειτα, το accuracy είναι μια έμπιστη μετρική, εφόσον διαχειριζόμαστε balanced datasets, ειδάλλως μπορεί να δώσει παραπλανητικά αποτελέσματα που δεν λαμβάνουν υπόψη τις λιγότερο πολυπληθείς κλάσεις. Στις περιπτώσεις **unbalanced datasets**, συχνά παρατηρείται μεγάλη απόκλιση ανάμεσα σε accuracy και f1-score, με το δεύτερο να είναι περισσότερο έγκυρο.

Αντιστοίχως, οι μετρικές **macro και micro f1-scores** μπορούν να παρουσιάζουν αποκκλίσεις σε unbalanced datasets, καθώς η πρώτη αντιμετωπίζει ισάξια όλες τις κλάσεις, λαμβάνοντας έτσι σοβαρά υπόψη και αυτές με τα λιγότερα δείγματα, ενώ η δεύτερη ενδιαφέρεται μόνο για το συνολικό αριθμό ορθών και εσφαλμένων ταξινομήσεων σε ολόκληρο το dataset. [1] [2] [3]

Στο συγκεκριμένο πρόβλημα ταξινόμησης μουσικού είδους, όπως φάνηκε και στο ιστόγραμμα κλάσεων μετά το Class Mapping, το dataset που χρησιμοποιείται δεν είναι σημαντικά unbalanced. Ως εκ τούτου, η μετρική **Accuracy** είναι αρκετά φερέγγυα, ενώ επειδή δεν δίνεται έμφαση στο είδος των εσφαλμένων ταξινομήσεων (FN ή FP) η μετρική weighted average **F1-score** κρίνεται επίσης αξιόπιστο μέτρο της επίδοσης του δικτύου.

Αξιολογώντας με βάση τις άνωθι μετρικές το LSTM δίκτυο για τις τέσσερεις διαφορετικές εισόδους (Πίνακες 1 έως 4), παρατηρεί κανείς ότι τα καλύτερα αποτελέσματα accuracy δίνονται όταν στην είσοδο περιλαμβάνονται τα **beat-synced mel spectrograms**. Μάλιστα, όταν αυτά συνενώνονται με τα αντίστοιχα chromagrams βελτίωνονται οι μετρικές F1. Τα χειρότερα αποτελέσματα, από την άλλη, προκύπτουν όταν στην είσοδο του δικτύου δίνονται μόνο chromagrams, γεγονός αναμενόμενο αν θυμηθεί κανείς τις απεικονίσεις τους στα πρώτα βήματα.

Σε κάθε περίπτωση, το ποσοστό ευστοχίας του δικτύου δεν ξεπερνά το **40%**, υποδηλώνοντας μια μέτρια επίδοση στην ταξινόμηση των δειγμάτων. Κάτι τέτοιο, ωστόσο, είναι αναμενόμενο, καθώς το πρόβλημα ταξινόμησης δειγμάτων μουσικής με βάση το είδος της είναι αρκετά δύσκολο από μόνο του, και ένα απλό LSTM δίκτυο σίγουρα δεν επαρκεί.

Class	Precision	Recall	F-score	Support
0	0.0	0.0	0.0	40
1	0.4	0.65	0.5	40
2	0.31	0.7	0.43	80
3	0.32	0.55	0.41	80
4	0.25	0.05	0.08	40
5	0.0	0.0	0.0	40
6	0.5	0.53	0.51	78
7	0.0	0.0	0.0	40
8	0.35	0.35	0.35	103
9	0.33	0.03	0.05	34
accuracy			0.36	575
macro avg	0.25	0.29	0.23	575
weighted avg	0.28	0.36	0.29	575

Table 1: Classification Report για είσοδο mel-spectrograms.

Class	Precision	Recall	F-score	Support
0	0.25	0.03	0.05	40
1	0.48	0.53	0.5	40
2	0.44	0.74	0.55	80
3	0.38	0.62	0.47	80
4	0.29	0.42	0.35	40
5	0.18	0.05	0.08	40
6	0.63	0.31	0.41	78
7	0.0	0.0	0.0	40
8	0.36	0.52	0.42	103
9	0.0	0.0	0.0	34
accuracy			0.4	575
macro avg	0.3	0.32	0.28	575
weighted avg	0.35	0.4	0.34	575

Table 2: Classification Report για είσοδο beat-synced mel-spectrograms.

Class	Precision	Recall	F-score	Support
0	0.0	0.0	0.0	40
1	0.0	0.0	0.0	40
2	0.25	0.06	0.1	80
3	0.21	0.5	0.3	80
4	0.0	0.0	0.0	40
5	0.0	0.0	0.0	40
6	0.31	0.36	0.33	78
7	0.0	0.0	0.0	40
8	0.19	0.5	0.28	103
9	0.0	0.0	0.0	34
accuracy			0.22	575
macro avg	0.1	0.14	0.1	575
avg	0.14	0.22	0.15	575

Table 3: Classification Report για είσοδο beat-synced chromagrams.

Class	Precision	Recall	F-score	Support
0	0.14	0.07	0.1	40
1	0.5	0.62	0.56	40
2	0.36	0.72	0.48	80
3	0.43	0.46	0.45	80
4	0.37	0.28	0.31	40
5	0.19	0.17	0.18	40
6	0.5	0.56	0.53	78
7	0.4	0.1	0.16	40
8	0.47	0.37	0.41	103
9	0.44	0.12	0.19	34
accuracy			0.4	575
macro avg	0.38	0.35	0.34	575
avg	0.4	0.4	0.38	575

Table 4: Classification Report για είσοδο beat-synced fused mel-spectrograms & chromagrams.

Βήματα Κυρίως Μέρους

7. 2D CNN Μοντέλο

Στο δεύτερο μέρος της εργαστηριακής άσκησης, τα φασματογραφήματα των ηχητικών σημάτων αντιμετωπίζονται πλέον ως εικόνες και χρησιμοποιούνται Συνελικτικά Νευρωνικά Δίκτυα για την ταξινόμησή τους.

Εκπαίδευση στο MNIST Dataset

Ως πρώτη επαφή με τα CNNs, εξετάζεται η εσωτερική δομή και λειτουργία τους, εκπαιδεύοντας ένα CNN δύο κρυφών επιπέδων στο πρόβλημα κατηγοριοποίησης εικονογραφημένων ψηφίων του MNIST Dataset [4].

Στο Σχήμα 10 φαίνεται η εσωτερική αρχιτεκτονική του δικτύου για είσοδο μια εικόνα με το ψηφίο 6. Παρατηρεί κανείς ότι στην έξοδο του πρώτου συνελικτικού επιπέδου, προκύπτουν **high-level χαρακτηριστικά**, όπως περιγράμματα, ακμές και γωνίες του σχήματος εισόδου, τα οποία κωδικοποιούν σημασιολογική πληροφορία εύκολα αντιληπτή από το ανθρώπινο μάτι. Η ReLU activation ύστερα εισάγει μια μη γραμμική κατωφλιοποίηση, ενώ το dropout layer μειώνει τη διαστατικότητα της εικόνας. Ακολούθως, στο δεύτερο κρυφό επίπεδο του δικτύου, η διαδικασία επαναλαμβάνεται, αλλά σε εικόνες που κωδικοποιούν **low-level χαρακτηριστικά**, που γίνονται δυσκολότερα κατανοητά. Τέλος, το output Fully Connected Layer υπολογίζει τις **aposteriori πιθανότητες** για καθεμία από τις 10 κλάσεις - ψηφία και με τη βοήθεια της softmax συνάρτησης, κατηγοριοποιεί επιτυγχώς την εικόνα εισόδου ως ψηφίο 6.

Εν γένει, προκύπτει ότι καθώς προβαίνουμε σε βαθύτερα συνελικτικά επίπεδα ενός CNN, μεταβαίνουμε από υψηλής ανάλυσης high level features μιας εικόνας σε χαμηλότερης ανάλυσης low level χαρακτηριστικά, ώσπου εν τέλει να προκύψει η aposteriori πιθανότητα κατηγοριοποίησης.

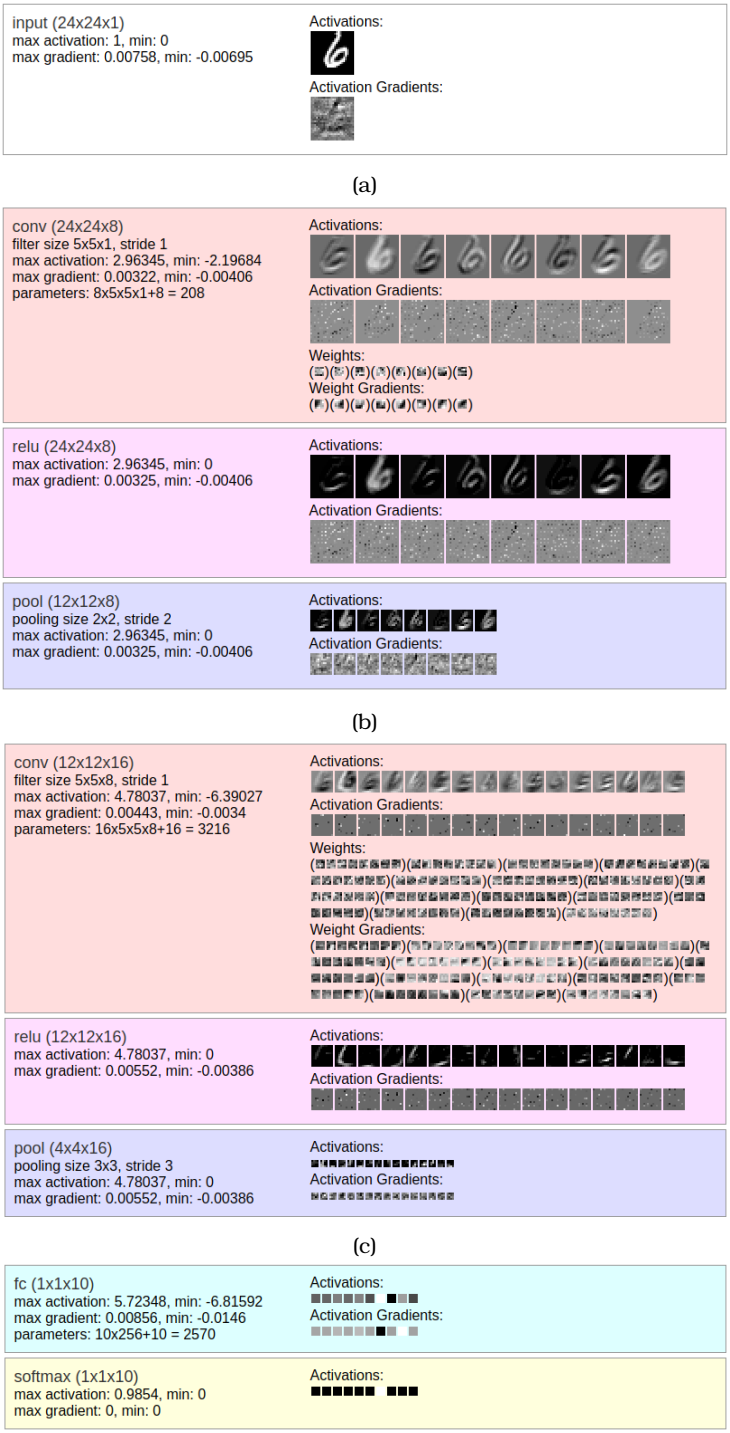


Figure 10: MNIST CNN (a) Input Layer, (b) 1st Hidden Layer, (c) 2nd Hidden Layer, (d) Output Layer.

Ταξινόμηση Φασματογραφημάτων με χρήση CNN

Για το πρόβλημα ταξινόμησης φασματογραφημάτων του dataset FMA genre, υλοποιούμε ένα 2D CNN τεσσάρων στρωμάτων με την εξής αρχιτεκτονική:

- 1ο Επίπεδο:
  - 2D Convolution: in_channels=1, out_channels=32, kernel_size=3, stride=1, padding=1
  - 2D Batch Normalization: num_feats = 32
  - ReLU Activation
  - Max-Pool Layer: kernel_size=2
- 2ο Επίπεδο:
  - 2D Convolution: in_channels=32, out_channels=64, kernel_size=3, stride=1, padding=1
  - 2D Batch Normalization: num_feats = 64
  - ReLU Activation
  - Max-Pool Layer: kernel_size=2
- 3ο Επίπεδο:
  - 2D Convolution: in_channels=64, out_channels=128, kernel_size=3, stride=1, padding=1
  - 2D Batch Normalization: num_feats = 128
  - ReLU Activation
  - Max-Pool Layer: kernel_size=4
- 4ο Επίπεδο:
  - 2D Convolution: in_channels=128, out_channels=256, kernel_size=3, stride=1, padding=1
  - 2D Batch Normalization: num_feats = 256
  - ReLU Activation
  - Max-Pool Layer: kernel_size=4
- Output Layer:
  - ReLU Activation
  - Dropout
  - Fully Connected Layer: in_features=10240, out_features=10

Convolutional Layers

Τα συνελικτικά επίπεδα αποτελούν τη βάση της αρχιτεκτονικής ενός CNN και ευθύνονται για την εκπαίδευση και κατά συνέπεια ενεργοποίηση των νευρώνων του δικτύου. Ουσιαστικά, υλοποιούν την πράξη της δισδιάστατης συνέλιξης (αθροίσματα πολλαπλασιασμών) πάνω σε τρισδιάστατα δεδομένα με χρήση ενός κυλιόμενου παράθυρου-πυρήνα. Ο πυρήνας ολισθαίνει πάνω στην εικόνα και έτσι υπολογίζεται κάθε φορά μια νέα τιμή pixel ως ένα βεβαρυμένο άθροισμα των pixels που περικλύει κάθε φορά. Τα βάρη του πυρήνα αποτελούν τις κύριες παραμέτρους του CNN, που χρειάζεται να ανανεωθούν κατά την εκπαιδευτική διαδικασία. [5] [6] [7]

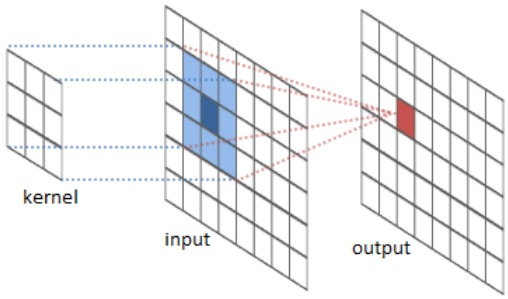


Figure 11: Οπτικοποίηση της 2D Συνέλιξης πάνω σε Εικόνα.

Batch Normalization

Η τεχνική Batch Normalization χρησιμοποιείται για την εκπαίδευση Νευρωνικών Δικτύων με μεγάλο βάθος, ώστε να σταθεροποιείται η διαδικασία μάθησης και να μειώνεται δραματικά ο συνολικός αριθμός εποχών που απαιτείται για την εκπαίδευση του δικτύου. Στην πράξη, τα δεδομένα **κανονικοποιούνται** σε κάθε επίπεδο του δικτύου και για κάθε mini-batch, με τρόπο τέτοιον ώστε να έχουν μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση. Βασικό πλεονέκτημα της εν λόγω τεχνικής είναι ότι αυξάνει το generalization capacity του δικτύου, διότι αφενός αντιμετωπίζει το "**internal covariate shift**" (δηλαδή τις μεταβολές στην κατανομή των δεδομένων εισόδου στους εσωτερικούς κόμβους του δικτύου κατά την εκπαίδευση του) και αφετέρου ομαλοποιεί τη συνάρτηση βελτιστοποίησης που χρησιμοποιείται.

ReLU Activation Function

Η έννοια της μη-γραμμικότητας - που εγγενώς στερείται η πράξη της συνέλιξης - εισάγεται στο CNN μέσω της μη γραμμικής συνάρτησης ενεργοποίησης ReLU (Rectified Linear Unit), η οποία ορίζεται ως $ReLU(x) = max(0, x)$.

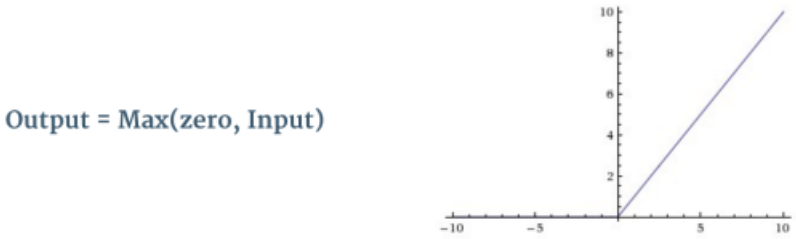


Figure 12: Η μη γραμμική συνάρτηση ενεργοποίησης ReLU.

Ο κύριος λόγος που η ReLU προτιμάται κατά κόρον στα CNNs έναντι άλλων μη-γραμμικών συναρτήσεων, είναι η ικανότητα της να αποτρέπει την εκθετική αύξηση στους υπολογισμούς που απαιτούνται για τη λειτουργία του δικτύου. Αυξάνοντας το βάθος ενός CNN, το υπολογιστικό κόστος με τη χρήση ReLUs αυξάνεται γραμμικά, καθώς πολύ νευρώνες δεν ενεργοποιούνται κατά την εκπαίδευση του.

Max Pooling

Τέλος, η τεχνική του Spatial Pooling χρησιμοποιείται για τη μείωση της διαστατικότητας κάθε feature map, διατηρώντας συγχρόνως τη σημαντική πληροφορία. Στην περίπτωση, του Max Pooling, χρησιμοποιείται κυλιόμενο παράθυρο-πυρήνας που κάθε φορά διατηρεί τη μέγιστη τιμή των pixels που περικλείει. Εν γένει, η τεχνική αυτή καθιστά τις αναπαραστάσεις εισόδου μικρότερες και πιο διαχειρίσιμες, μειώνοντας έτσι το πλήθος των προς εκμάθηση παραμέτρων και το συνολικό υπολογιστικό κόστος του δικτύου. Συγχρόνως,

αποτρέπει την εκδήλωση φαινομένων overfitting και παρέχει scale invariant, αναλλοίωτες αναπαραστάσεις που επιτρέπουν τον εντοπισμό αντικειμένων σε εικόνες οπουδήποτε στον χώρο. [5] [6] [7]

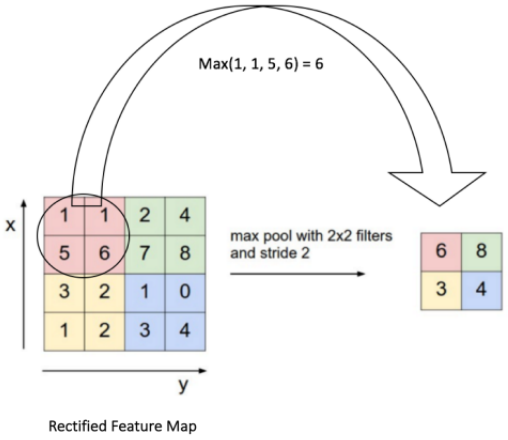


Figure 13: Παράδειγμα εφαρμογής ενός πυρήνα max-pooling.

Στην προκειμένη περίπτωση, εκπαιδεύουμε το CNN που δημιουργήσαμε πάνω στα non-beat synced φασματογραφήματα, χρησιμοποιώντας Cross Entropy Loss, Adam Optimizer (learning rate = 0.001, weight decay = 0.0001), για 30 εποχές και batch size 32. Χρησιμοποιούμε τεχνικές Early Stopping, ώστε να αποφύγουμε την υπερεκπαίδευση του μοντέλου και εν τέλει τερματιζόυμε στις 16 εποχές.

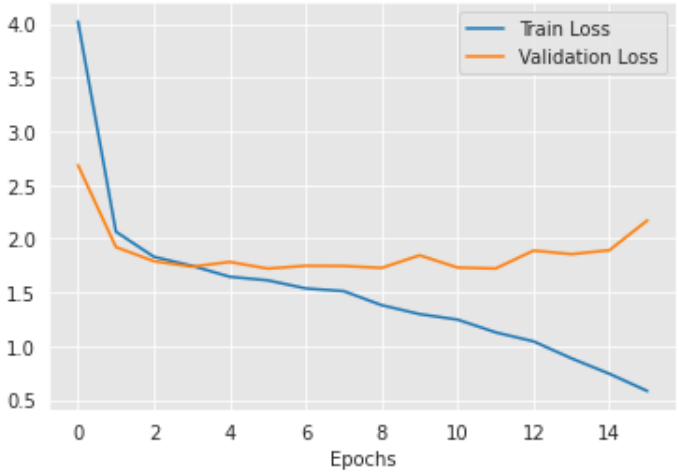


Figure 14: Learning Curve του CNN πάνω στο multitask dataset.

Αξιολογώντας το CNN με βάση τις μετρικές της προπαρασκευής (Πίνακας 5) παρατηρούμε ποσοστό ευστοχίας **41%**, το οποίο είναι μεγαλύτερο από κάθε μοντέλο LSTM που δοκιμάστηκε στο Βήμα 5, υπερνικώντας μάλιστα κατά πολύ το αντίστοιχο LSTM στα non-beat-synced data. Ως εκ τούτου, το CNN φαίνεται να υπερτερεί για το εν λόγω task.

Class	Precision	Recall	F-score	Support
0	0.24	0.17	0.2	40
1	0.58	0.55	0.56	40
2	0.54	0.54	0.54	80
3	0.43	0.5	0.46	80
4	0.55	0.53	0.54	40
5	0.17	0.33	0.22	40
6	0.63	0.4	0.49	78
7	0.0	0.0	0.0	40
8	0.39	0.35	0.37	103
9	0.28	0.62	0.39	34
accuracy			0.41	575
macro avg	0.38	0.4	0.38	575
avg	0.41	0.41	0.4	575

Table 5: CNN Classification Report για είσοδο mel-spectrograms.

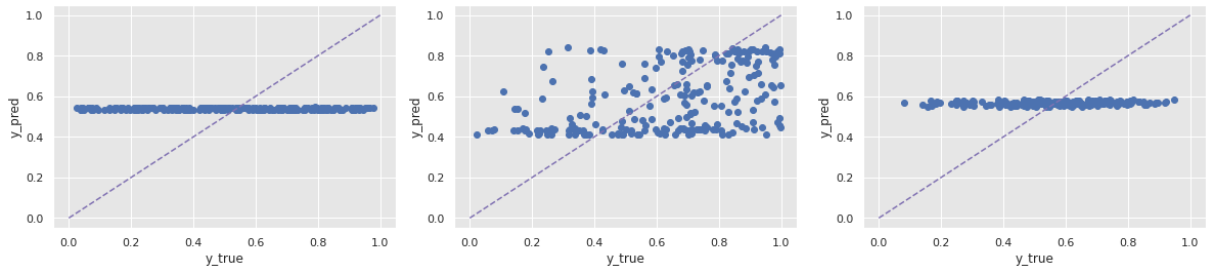
8. Εκτίμηση συναισθήματος - συμπεριφοράς με παλινδρόμηση

Στη συνέχεια, εστιάζουμε στο multitask dataset, το οποίο εμπεριέχει φασματογραφήματα, καθώς και επισημειώσεις σε 3 άξονες που αφορούν το συναίσθημα του τραγουδιού. Οι επισημειώσεις είναι πραγματικοί αριθμοί μεταξύ 0 και 1 (valence, energy, danceability).

Φορτώνοντας το CNN του προηγούμενου βήματος και το LSTM της προπαρασκευής που έδωσε τα καλύτερα αποτελέσματα (bidirectional LSTM, hidden_size = 128, 4 layers) προβαίνουμε στην εκπαίδευση τους πάνω στο εν λόγω dataset. Χρησιμοποιούμε τεχνικές Early Stopping, ενώ ακριβώς επειδή δεν διαθέτουμε labelled test data, κρατάμε ένα ποσοστό δεδομένων ως validation set, το οποίο χρησιμοποιείται επίσης για την αξιολόγηση των μοντέλων μας.
Η αξιολόγηση της επίδοσης γίνεται με βάση τη μετρική **Spearman Correlation**, η οποία ορίζεται ως:

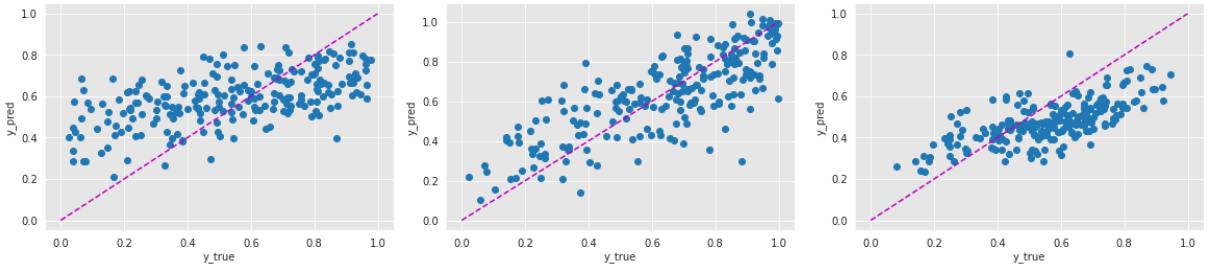
$$\rho = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}$$

και εκφράζει κατά πόσο η σχέση δύο μεταβλητών μπορεί να προσεγγιστεί με μια μονοτονική συνάρτηση (λαμβάνοντας τιμές στο διάστημα [-1,1]).



(a) (b) (c)

Figure 15: Scatter Plots των y_gold, y_pred για το LSTM με labels (a) valence, (b) energy, (c) danceability.



(a) (b) (c)

Figure 16: Scatter Plots των y_gold, y_pred για το CNN με labels (a) valence, (b) energy, (c) danceability.

Model	Valence	Energy	Danceability	Mean Value
LSTM	0.0710	0.4507	0.1673	0.230
CNN	0.5682	0.7639	0.6719	0.668

Table 6: Spearman Correlations ανά label για τα δύο μοντέλα.

Παρατηρώντας τις τιμές των Spearman συχετίσεων στον Πίνακα 6, είναι φανερό ότι **το CNN μοντέλο φαίνεται να υπερνικά κατά πολύ το LSTM**, το οποίο αδυνατεί να προσαρμοστεί στο πρόβλημα regression και ταξινομεί τα πάντα κοντά στο 0.5. Οι τιμές της μετρικής περισσότερο κοντά στη μονάδα, μάλιστα, επιτυγχάνονται για το μέγεθος energy, ξεπερνώντας την τιμή του 70% για το CNN.

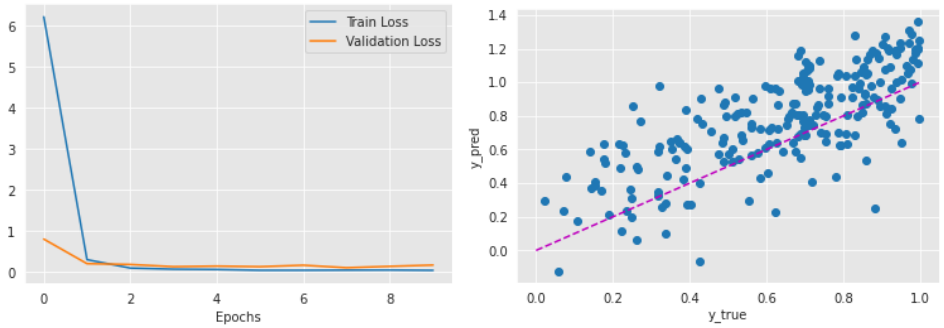
9. Μεταφορά Μάθησης (Transfer Learning)

Ακολούθως, πειραματιζόμαστε με τεχνικές Transfer Learning για την αξιοποίηση προεκπαιδευμένων μοντέλων στο πρόβλημα ταξινόμησης του multitask dataset. Όπως αναφέρεται στο paper των **Yosinski et al.** [8], αξιοποιώντας το γεγονός ότι καθώς διασχίζουμε τα επίπεδα ενός DNN μαθαίνονται χαρακτηριστικά αρχικά πιο γενικά - υπό την έννοια ότι βρίσκουν εφαρμογή σε διαφορετικά datasets και διαφορετικά προβλήματα - κατ'αναλογία με τα Gabor φίλτρα ή τα color blobs, και τελικώς περισσότερο task specific, προτείνεται η μεταφορά βαρών προεκπαιδευμένων δικτύων. Η μεταφερσιμότητα - εφόσον εκτελείται κυρίως στα πρώτα layers και συνοδευόμενη με κατάλληλο fine tuning - συμβάλλει στην επιτάχυνση της διαδικασίας μάθησης και στην ενίσχυση του generalization ability του δικτύου, υπερνικώντας την τυχαία αρχικοποίηση των βαρών.

Στην περίπτωση του multitask dataset, επιλέγουμε το CNN μοντέλο που έδωσε τα καλύτερα αποτελέσματα. Αποφεύγουμε να χρησιμοποιήσουμε το LSTM δίκτυο, διότι αφενός έχει εκ φύσεως έχει λίγα στρώματα και αφετέρου η μεταφερσιμότητα σε LSTM Δίκτυα δεν είθισται, λόγω της αναδρομικής τους φύσης.

Αφού αποθηκεύσουμε το CNN μοντέλο που εκπαιδεύτηκε στα fma genre non beatsynced φασματογραφήματα και έδωσε τα καλύτερα αποτελέσματα με βάση το accuracy, παγώνουμε τα βάρη των κρυφών επιπέδων του και αλλάζουμε την κεφαλή του σε ένα Fully Connected Layer με μία έξοδο. Ο λόγος που επιλέχθηκαν τα non beat synced φασματογραφήματα είναι επειδή τα CNNs δεν κινδυνεύουν από φαινόμενα vanishing/exploding gradients όσο τα LSTMS, οπότε δεν έχει νόημα να πετάξουμε την πλεονάζουσα πληροφορία. Η μετρική accuracy δε είναι ενδεικτική της συνολικής επίδοσης του μοντέλου στο εν λόγω πρόβλημα κατηγοριοποίησης.

Εκπαιδεύοντας, λοιπόν, το CNN μόνο για 10 εποχές (fine tuning) για τον συναισθηματικό άξονα του energy, επιτυγχάνουμε Spearman Correlation **0.7365**, τιμή ελάχιστα χαμηλότερη από την αντίστοιχη του προηγούμενου βήματος. Συνεπώς, με την τεχνική Transfer Learning επιτυγχάνουμε πολύ ικανοποιητικά αποτελέσματα, δίχως να χρειαστεί να εκπαιδεύσουμε το μοντέλο για πάρα πολλές εποχές, καθώς αξιοποιούμε προεκπαιδευμένα βάρη.



(a) (b)

Figure 17: Transfer Learning με CNN (a) Learning Curve, (b) Spearman Correlation.

10. Multitask Learning

Στο τελικό βήμα, εκμεταλλευόμαστε το γεγονός ότι μας δίνονται πολλές επισημειώσεις (valence, energy, danceability) και καταφεύγουμε σε τεχνικές multitask learning. Στο paper των Kaiser et al. [9], προκύπτει ότι η ταυτόχρονη εκπαίδευση ενός μόνο μοντέλου πάνω σε διαφορετικές εργασίες (tasks) αποδίδει εξίσου καλά με άλλες state-of-the-art αρχιτεκτονικές. Μάλιστα, προκύπτουν τα εξής συμπεράσματα:

• Η ταυτόχρονη εκπαίδευση των tasks δίνει ίδια ή καλύτερα ποσοστά ευστοχίας από την ξεχωριστή εκπαίδευση πολλών δικτύων ανά task, ιδίως σε περιπτώσεις με μικρά datasets. Η διαδικασία του Multitask Learning δε παρομοιάζεται με μεθόδους Transfer Learning από tasks με πληθώρα δεδομένων σε άλλα με πιο μικρά datasets.

• Ορισμένα blocks της multitask αρχιτεκτονικής που φαίνεται να είναι άσχετα ή ελάχιστης σημασίας για κάποιο συγκεκριμένο task, δεν επηρεάζουν την επίδοση του μοντέλου, ενώ συχνά παρέχουν μια μικρή βελτίωση.

Στην περίπτωση του CNN μας πάνω στο πρόβλημα του multitask dataset, αρχικά υλοποιούμε μια **συνάρτηση κόστους** (σαν nn.Module), η οποία λαμβάνει σαν είσοδο τα logits (outputs του μοντέλου) και τα targets (πραγματικές τιμές) για το valence, energy και danceability και επιστρέφει το άθροισμα των επιμέρους losses για κάθε task. Επιπλέον, τροποποιούμε την κεφαλή του δικτύου, ώστε στην έξοδο να προκύπτουν τρεις τιμές ως logits, μία για κάθε συναισθηματικό άξονα.
Εκπαιδεύοντας το συνολικό μοντέλο για 30 εποχές και διατηρώντας τις ίδιες υπερπαραμέτρους με πριν, προκύπτουν οι μετρικές του Πίνακα 7. Όπως είναι φανερό, η επίδοση του Multitask Trained δικτύου είναι **καλύτερη** συγκριτικά με τα επιμέρους δίκτυα του βήματος 8 και για τους τρεις συναισθηματικούς άξονες. Κάτι τέτοιο υποδηλώνεται και από τη μέση τιμή του Spearman Correlation, που αγγίζει την τιμή 72% και επιβεβαιώνει τα κύρια συμπεράσματα του άνωθι paper.

Model	Valence	Energy	Danceability	Mean Value
Multitask CNN	0.6526	0.8047	0.7125	0.7233

Table 7: Spearman Correlations ανά label με την τεχνική Multitask Learning.

11. Υποβολή στο Kaggle

Αποθηκεύοντας το CNN που έδωσε τα καλύτερα αποτελέσματα κατά τη φάση του Multitask Learning, προβαίνουμε σε ορισμένες τροποποιήσεις με στόχο τη βέλτιστη κατηγοριοποίηση των test δεδομένων. Αρχικά, προσθέτουμε ένα επιπλέον κρυφό επίπεδο στην αρχιτεκτονική του δικτύου και πειραματιζόμαστε με το batch size και το ποσοστό του train-validation split. Εκπαιδεύοντας εκ νέου το δίκτυο, υποβάλλουμε τις προβλέψεις για τις τιμές valence, energy και danceability κάθε δείγματος στον διαγωνισμό του Kaggle και επιτυγχάνουμε επίδοση **0.69407** στο σύνολο όλων των test data.

Το ποσοστό αυτό είναι ελάχιστα μικρότερο από το αντίστοιχο Spearman Correlation που καταφέραμε στο προηγούμενο βήμα, γεγονός που οφείλεται αφενός στην τυχαιότητα της εκπαιδευτικής διαδικασίας και αφετέρου στο ότι τα test data είναι εντελώς νέα. Σε κάθε περίπτωση, το generalization capacity του δικτύου μας είναι αρκετά ικανοποιητικό.

References

[1] B. Shmueli. (2019) Multi-class metrics made simple, part i: Precision and recall. [Online]. Available: https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

[2] ——. (2019) Multi-class metrics made simple, part ii: the f1-score. [Online]. Available: https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1

[3] I. Kuznetsov. (2019) Metrics for imbalanced classification. [Online]. Available: https://towardsdatascience.com/metrics-for-imbalanced-classification-41c71549bbb5

[4] Stanford. Convnetjs: Deep learning in your browser. [Online]. Available: https://cs.stanford.edu/people/karpathy/convnetjs/

[5] colah's blog. Understanding convolutions. [Online]. Available: https://colah.github.io/posts/2014-07-Understanding-Convolutions/

[6] ujjwalkarn. An intuitive explanation of convolutional neural networks. [Online]. Available: https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

[7] A. Saxena. Convolutional neural networks (cnns): An illustrated explanation. [Online]. Available: https://blog.xrds.acm.org/2016/06/convolutional-neural-networks-cnns-illustrated-explanation/

[8] Y. B. H. L. Jason Yosinski, Jeff Clune, "How transferable are features in deep neural networks?"

[9] N. S. A. V. N. P. L. J. J. U. Łukasz Kaiser, Aidan N. Gomez, "One model to learn them all."