



National Technical University of Athens  
Master of Data Science and Machine Learning

**[3208] Pattern Recognition  
2nd Laboratory Assignment**

**Submitted by:**

Evangelos Chaniadakis 03400279  
Alexios Filippakopoulos 03400278

**Submission Date:**  
December 23, 2024

Στην δεύτερη εργαστηριακή άσκηση αξιοποιήσαμε δύο datasets: ένα υποσύνολο του **Free Music Archive (FMA)** για ταξινόμηση μουσικών κομματιών σε είδη μουσικής και το **Multitask Dataset** για να προβλέψουμε συναισθηματικές διαστάσεις μουσικών κομματιών, όπως το *valence* (θετικότητα/αρνητικότητα συναισθήματος), *energy* (ένταση συναισθήματος) και *danceability* (χορευτική δυνατότητα), με χρήση παλινδρόμησης. Για την επεξεργασία των δεδομένων, χρησιμοποιήθηκαν φασματογραφήματα και χρωμογραφήματα ως αναπαραστάσεις των μουσικών σημάτων. Για την ταξινόμηση των μουσικών κομματιών του FMA dataset, εφαρμόστηκαν οι αρχιτεκτονικές: LSTM για ανάλυση χρονικών εξαρτήσεων, 2D CNN για την εκμετάλλευση των φασματογραφημάτων ως εικόνες και Audio Spectrogram Transformer. Στη συνέχεια, για την πρόβλεψη των συναισθηματικών διαστάσεων (*valence*, *energy*, *danceability*), υλοποιήθηκαν μοντέλα παλινδρόμησης, με αντίστοιχο backbone όπως νωρίτερα ((CNN, LSTM & AST), που προσαρμόστηκαν στις απαιτήσεις του Multitask Dataset. Για την αξιολόγηση των ταξινομητών, χρησιμοποιήθηκαν μετρικές όπως accuracy, precision, recall & F1-score, ενώ για την αξιολόγηση των παλινδρομικών μοντέλων (ως προς τις συναισθηματικές διαστάσεις), αξιοποιήθηκαν τα Spearman & Pearson correlation, MSE και αρκετά ακόμη.

Ταυτόχρονη με την ταξινόμηση των κομματιών, εφαρμόστηκε η τεχνική του multitask learning για την εκπαίδευση ως προς όλες τις συναισθηματικές διαστάσεις, αλλά και το transfer learning για τη βελτίωση της απόδοσης των Transformers λόγω περιορισμένων δεδομένων για εκπαίδευση. Τέλος, έγινε οπτικοποίηση των κρυφών αναπαραστάσεων κάποιων μουσικών κομματιών μέσω πολλαπλών τεχνικών μείωσης διαστασιμότητας, για την αξιολόγηση της ποιότητας διαχωρισμού των κλάσεων.

## Βήματα 1, 2 & 3

Ξεκινάμε την μελέτη μας με το FMA dataset, το οποίο έχει εξαχθεί από το πραγματικό FMA dataset που υπάρχει διαθέσιμο στο GitHub - FMA Dataset και περιέχει μουσικά αποσπάσματα, μεταδεδομένα και διάφορα χαρακτηριστικά αυτών.

Το συγκεκριμένο υποσύνολο που χρησιμοποιείται εδώ περιλαμβάνει 3.834 φασματογραφήματα και χρωμογραφήματα, κατανεμημένα σε 20 διαφορετικές κατηγορίες μουσικών ειδών. Χάριν απλότητας, ζητούμαστε να τις εκφυλίσουμε σε 10, ενοποιώντας παρεμφερή genres και αγνοώντας άλλα. Σε αυτό το σημείο να τονίσουμε ότι διατήρουμε και καλούμαστε να διερευνήσουμε δύο διαφορετικές εκδοχές του παραπάνω συνόλου δεδομένων. Οι εκδοχές αυτές διαχωρίζονται στα original spectrograms & chromagrams των μουσικών αποσπασμάτων και στα beat-synced spectrograms & chromagrams. Η beat-synced εκδοχή είναι χρήσιμη στη μελέτη μας καθώς παρουσιάζει μεγαλύτερη συνοχή με τη μουσική δομή του κάθε ξεχωριστού κομματιού. Είναι επίσης ένας αξιόλογος τρόπος να κάνουμε dimensionality reduction ενισχύοντας τα χαρακτηριστικά του κάθε κομματιού και μειώνοντας θόρυβο όπως πάυσεις κλπ.

Τα φασματογραφήματα είναι οπτικές αναπαραστάσεις του ηχητικού σήματος που δείχνουν τον τρόπο που η ενέργεια του ήχου κατανέμεται σε διαφορετικές συχνότητες με την πάροδο του χρόνου. Τα χρωμογραφήματα είναι ένας τρόπος αναπαράστασης της μουσικής που δείχνει την ενέργεια του ήχου για κάθε μία από τις 12 νότες της μουσικής κλίμακας, ανεξάρτητα από την οχτάβα.

Στο δια ταύτα λοιπόν, η διερεύνηση αυτή υλοποιείται στο *step\_1\_2\_3.ipynb*. Ο κώδικας ιδιαίτερα απλοικός χωρίς κάποια ιδιαίτερη εξάρτηση σε εξωτερικές βιβλιοθήκες ή utilities πέραν του librosa που μας βοηθάει στην επεξεργασία/οπτικοποίηση των δεδομένων μας. Διαβάζουμε τα δεδομένα από το δωθέν archive και τα ομαδοποιούμε με βάση το genre. Τυχαία επιλέγουμε δύο κομμάτια ξεχωριστού genre και τα οπτικοποιούμε.

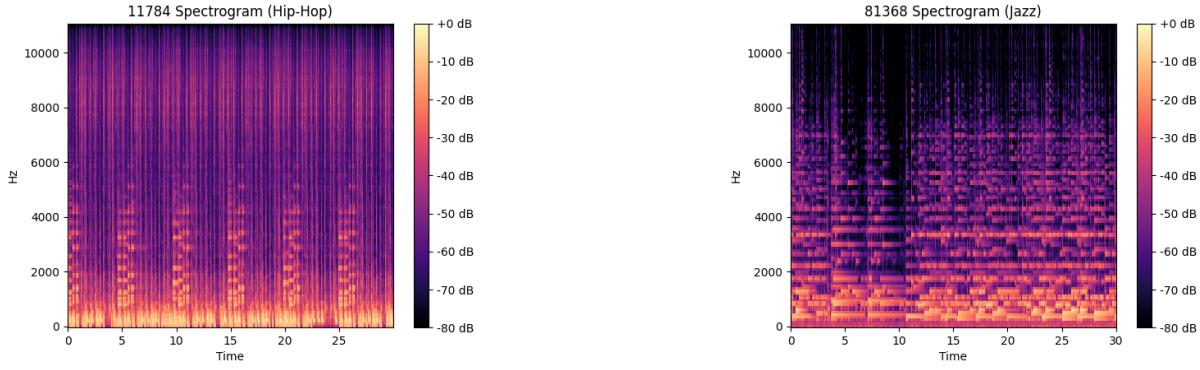


Figure 1: Spectrograms

Τα φασματογραφήματα που παρουσιάζονται παραπάνω αποτυπώνουν τη χρονική εξέλιξη της ενέργειας του ήχου σε διαφορετικές συχνότητες και παρέχουν πληροφορίες για το ηχόχρωμα, τη δυναμική και τη συχνοτική κατανομή των δειγμάτων.

Στο φασματογράφημα του hip-hop κομματιού 11784 παρατηρούμε υψηλή συγκέντρωση στις χαμηλές συχνότητες, βασικά κάτω από 500Hz, και μικρή συγκέντρωση στις μεσαίες, μεταξύ δηλαδή 500Hz και 2000Hz, αλλά και έντονα, περιοδικά peaks που επαναλαμβάνονται ανά περίπου 5 δεύτερα, τα οποία απλώνονται και στο φάσμα των υψηλών. Από τα παραπάνω καταλαβαίνουμε την υπάρξη ρυθμικού μπάσου. Το γεγονός ότι οι υψηλές συχνότητες είναι περιορισμένες, υποδηλώνει έλλειψη ιδιαίτερα μελωδικών στοιχείων.

Στο φασματογράφημα του jazz κομματιού 81368, σε αντίθεση με το προηγούμενο δείγμα, παρατηρούμε πιο έντονη δραστηριότητα στις μεσαίες, όχι ιδιαίτερα στις χαμηλές και διάσπαρτη συγκέντρωση στις υψηλές συχνότητες. Η ενέργεια είναι κατανεμημένη σε ευρύτερο φάσμα συχνοτήτων, κάτι που φανερώνει μεγαλύτερη μουσική ποικιλία και πιθανόν την ύπαρξη διαφορετικών οργάνων. Δεν υπάρχει έντονη ρυθμική επανάληψη όπως στο hip-hop κομμάτι. Αντίθετα, παρατηρείται ακανόνιστη δομή στο χρόνο, κάτι το οποίο αντανακλά και τον αυτοσχεδιαστικό χαρακτήρα της jazz μουσικής.

Η κλίμακα Mel είναι μια κλίμακα συχνοτήτων σχεδιασμένη έτσι ώστε να αντανακλά την ανθρώπινη ακουστική αντίληψη των συχνοτήτων. Δημιουργήθηκε για να προσεγγίσει το γεγονός ότι το ανθρώπινο αυτί δεν αντιλαμβάνεται τις συχνότητες γραμμικά, αλλά βασικά δίνει μεγαλύτερη έμφαση στις χαμηλές συχνότητες. Δηλαδή, σε χαμηλές συχνότητες είναι πιο ευαίσθητο σε μικρές διαφορές, ενώ όσον αφορά υψηλές συχνότητες χρειάζονται μεγαλύτερες διαφορές για να γίνουν αντιληπτές. Αφορά την σχετική αντίληψη της απόστασης μεταξύ δύο συχνοτήτων, από τον άνθρωπο, η οποία αναλύθηκε μέσω ψυχοακουστικών πειραμάτων.

Η κλίμακα Mel χρησιμοποιείται στην επεξεργασία μουσικών σημάτων γιατί, όπως προείπαμε, προσομοιώνει την ανθρώπινη ακουστική αντίληψη. Συμπυκνώνει το φάσμα σε λιγότερες ζώνες, άρα είναι και τεχνική dimensionality reduction, διατηρώντας βέβαια τις σημαντικές πληροφορίες. Με αυτόν τον τρόπο, εξάγονται αποδοτικά χαρακτηριστικά όπως τα Mel-spectrograms και τα MFCCs, που αποφαίνονται ιδιαίτερα χρήσιμα στην ανάλυση και ταξινόμηση μουσικών σημάτων, αλλά και σημάτων φωνής.

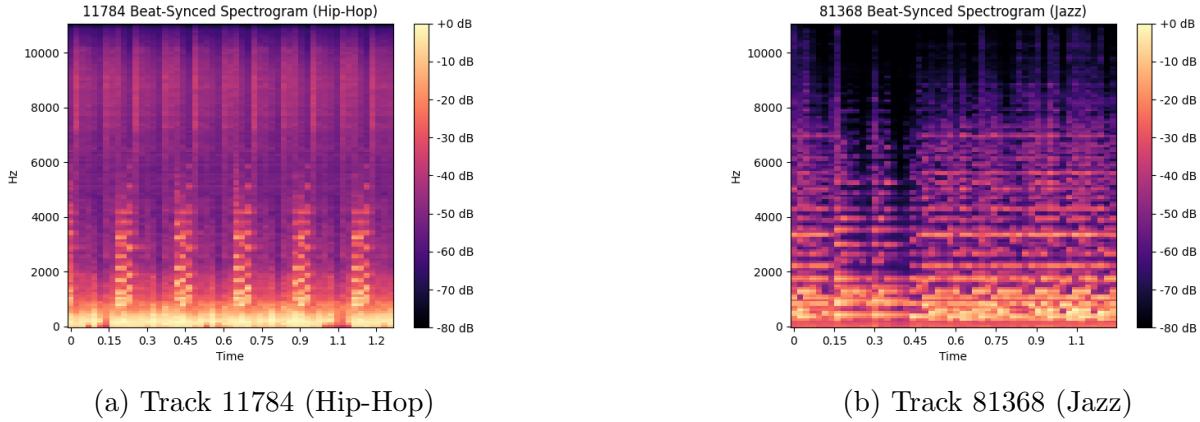


Figure 2: Beat-Synced Spectrograms

Για να βρούμε πόσα χρονικά βήματα έχει ένα φασματογράφημα εξάγουμε ένα batch από τα δεδομένα εκπαίδευσης. Εξάγοντας το πρώτο βατζη, παρατηρώ ότι κάθε φασματογράφημα του dataset μου δεν έχει όμοια χρονικά βήματα αλλά κυμαίνονται γύρω στα 1291-1293. Έχουμε επίσης 128 χαρακτηριστικά (συχνότητες) που εξάγονται για κάθε χρονικό βήμα, τα οποία αντιστοιχούν στην κλίμακα Mel. Αναλογιζόμενος τα παραπάνω, δηλαδή την τάξη μεγέθους των χρονικών βημάτων των φασματογραφημάτων και την φύση-αρχιτεκτονική του LSTM που ζητούμαστε να εκπαιδεύσουμε πάνω τους, γρήγορα οδηγούμαι στο συμπέρασμα ότι η προσέγγιση αυτή δεν είναι αποδοτική και βέλτιστη. Η εκπαίδευση ενός LSTM σε δεδομένα με μεγάλο αριθμό χρονικών βημάτων παρουσιάζει σημαντικά μειονεκτήματα. Αφενός, η υπολογιστική πολυπλοκότητα αυξάνεται γραμμικά με το μήκος των χρονικών βημάτων, καθιστώντας την εκπαίδευση αργή και απαιτητική σε πόρους. Αφετέρου, το φαινόμενο των vanishing gradients περιορίζει την ικανότητα του δικτύου να μάθει μακροπρόθεσμες εξαρτήσεις, καθώς οι πληροφορίες που βρίσκονται σε απομακρυσμένα χρονικά σημεία χάνονται κατά τη διαδικασία του backpropagation. Σαν να μην έφτανε αυτό, η ασυμμετρία στον αριθμό των χρονικών βημάτων, κατα μήκος του συνόλου δεδομένων, προσθέτει ακόμη περισσότερη πολυπλοκότητα καθώς απαιτεί ιδιαίτερο χειρισμό, όπως το padding.

Η beat-synced προσέγγιση συγχρονίζει τα φασματογραφήματα στα σημεία που χτυπάει το beat. Αντί δηλαδή να χρησιμοποιούμε όλα τα χρονικά βήματα, επιλέγουμε τα beat-aligned σημεία. Αυτό μειώνει τα χρονικά βήματα, διατηρώντας τα σημαντικά σημεία που σχετίζονται με το ρυθμό της μουσικής. Όπως ήδη προείπαμε, είναι μια αξιόλογη τεχνική dimensionality reduction που μας βοηθάει να κρατήσουμε τις πιο χρήσιμες πληροφορίες ανάλογα με την φύση του κομματιού.

Παρατηρώντας τα beat-synced spectrograms βλέπω αρχικά ότι έχει υποστεί συμπίεση ο χρονικός άξονας. Διατηρούν τα σημαντικά ρυθμικά μοτίβα, αποτυπώνοντας τον ρυθμό πιο ξεκάθαρα χωρίς να αποτυπώνουν διάχυτη ολόκληρη την, πιθανόν περιττή σε σημεία, πληροφορία. Το πρόβλημα των vanishing gradients μειώνεται σημαντικά πλέον αφού το μέγιστο πλήθος χρονικών βημάτων είναι 129, δηλαδή το υποδεκαπλάσιο. Άρα μειώνεται η πολυπλοκότητα, και το μοντέλο είναι πιο αποδοτικό ως προς τον στόχο μας.

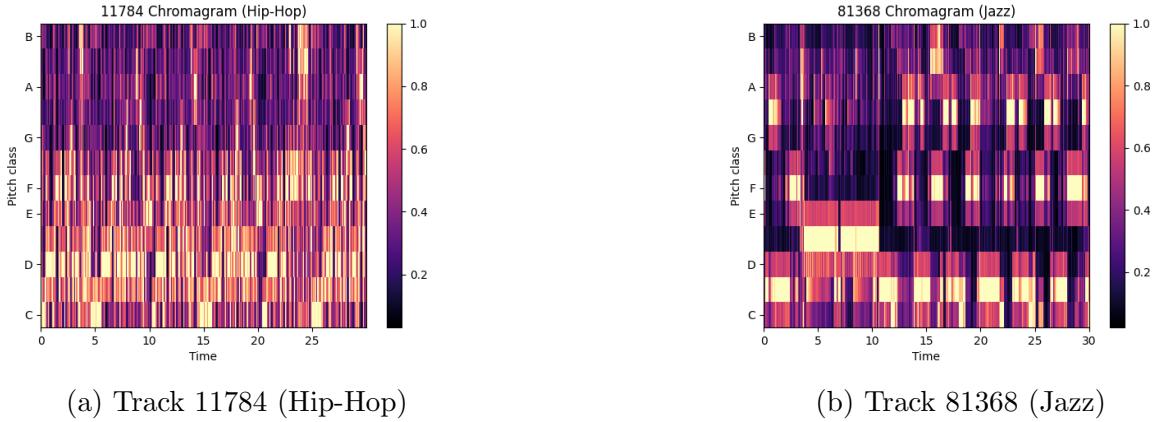


Figure 3: Chromograms

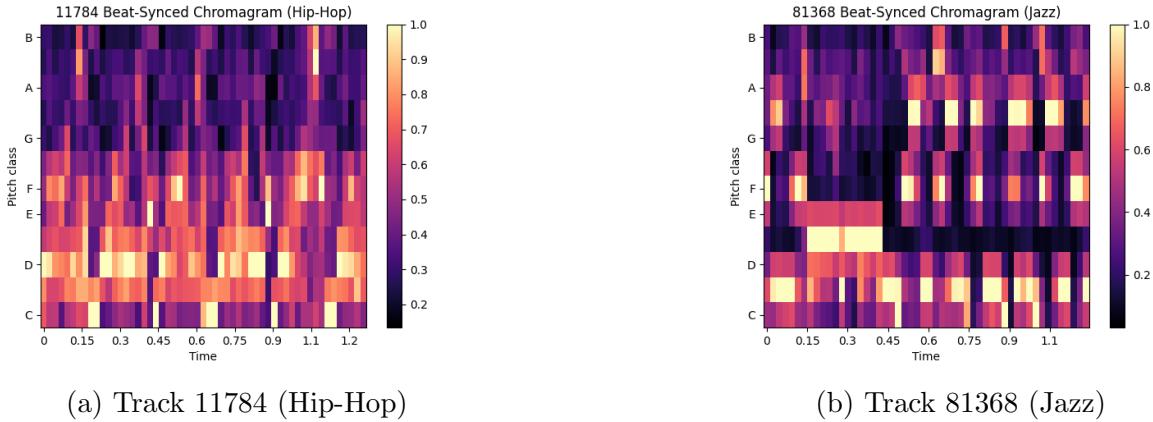


Figure 4: Beat-Synced Chromograms

Τα χρωμογραφήματα διαφέρουν ουσιαστικά από τα φασματογραφήματα ως προς τον τύπο της πληροφορίας που αναπαριστούν. Τα χρωμογραφήματα επικεντρώνονται στην αρμονική πληροφορία, καταγράφοντας την ένταση των 12 μουσικών νοτών, ανεξαρτήτως οκτάβας. Αυτή η απλοποίηση καθιστά τα χρωμογραφήματα ιδανικά για την ανάλυση μελωδίας και αρμονίας, καθώς προσφέρουν πιο συμπαγή και περιοδικά μοτίβα που αποτυπώνουν την τονικότητα και τις αρμονικές δομές του μουσικού κομματιού. Προσφέρουν μια απλουστευμένη αναπαράσταση, σε αντίθεση με τα φασματογραφήματα που αποτυπώνουν πιο λεπτομερώς τη ενέργεια των συχνοτήτων και το ρυθμικό περιεχόμενο της μουσικής.

Για αυτό εκτιμώ πως δεν θα είναι τόσο αποδοτικά όπως τα mel spectrograms στην ταξινόμηση ειδών μουσικής, καθώς το συγκεκριμένο task δεν έχει να κάνει μόνο με τις 12 μουσικές νότες αλλά και με άλλα χαρακτηριστικά τα οποία πιθανόν να μην αποτυπώνονται τόσο καλά πλέον, με την μείωση της διαστατικότητας που έχει επέλθει.

## Βήμα 4

Για την δημιουργία ενός αντικειμένου *Dataset*, χρησιμοποιούμε την κλάση *SpectrogramDataset*, η οποία δέχεται ως ορίσματα ένα μονοπάτι που βρίσκονται τα αντίστοιχα αρχεία, τον τύπο των χαρακτηριστικών {*chroma*, *mel*, *fused*}, το μέγιστο μήκος ακολουθίας των τύπο των ετικετών, δηλαδή αν το πρόβλημα είναι παλινδρόμηση ή ταξινόμηση, και αν το σύνολο δεδομένων προς επιστροφή είναι προς εκπαίδευση ή όχι. Χρησιμοποιεί ένα λεξικό *CLASS\_MAPPING* για να μειώσει το σύνολο των ετικετών με το να συγχωνεύει υποείδη μουσικής στα αντίστοιχα ευρύτερα αλλά και να βγάζει είδη για τα οποία δεν υπάρχουν πολλά δεδομένα. Έτσι το τελικό σύνολο δεδομένων είναι μικρότερο όσο αφορά τις διάφορες κλάσεις και κάθεμία από αυτές να έχει περισσότερα δείγματα. Χρησιμοποιεί την συνάρτηση *read\_spectrogram* για να διαβάσει τα αρχεία και να εξάγει τον κατάλληλο τύπο χαρακτηριστικών, εφαρμόζει το *mapping* των ετικετών, εφαρμόζει *zero-padding* για να φέρει όλα τα χαρακτηριστικά στο ίδιο μήκος και χρησιμοποιεί τον *LabelTransformer* για να κωδικοποιήσει της ετικέτες ανάλογα το πρόβλημα. Τέλος επιστρέψει ένα σύνολο του οποίου τα στοιχεία είναι τριάδες που αποτελούνται από το αντίστοιχο δείγμα τύπου *nd.array* και σχήματος (*max\_length*, *num\_features*, τη αντίστοιχη ετικέτα και το αρχικό μήκος του δείγματος πριν από το *padding*.

Ακολουθούν τα ιστογράμματα για τα σύνολα δεδομένων με και χωρίς συγχώνευση/αφαίρεση κλάσεων ώστε να οπτικοποιηθεί η διαφορά τους

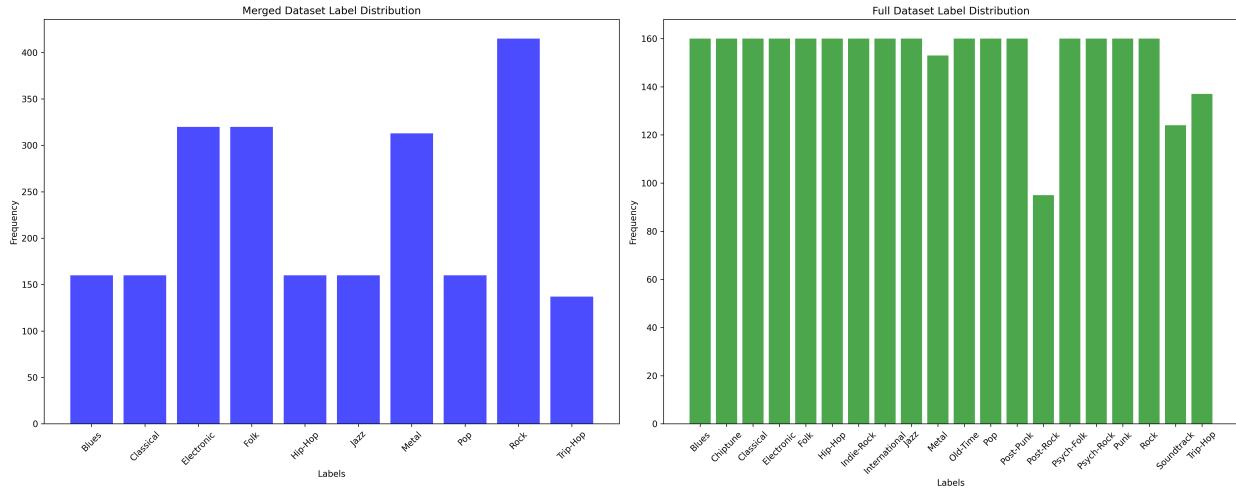
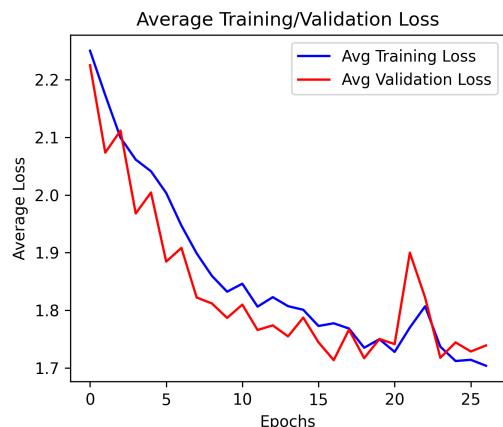


Figure 5: Dataset before and after applying the class mapping.

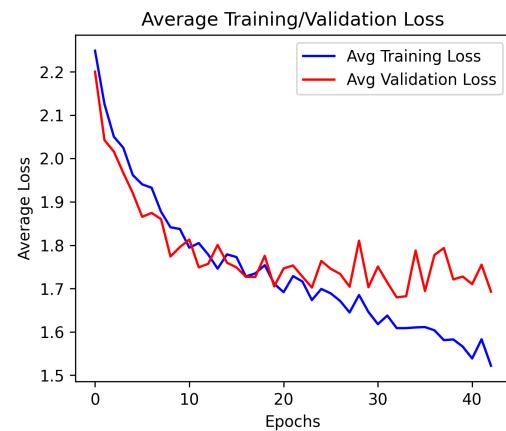
Παρατηρούμε ότι μετά το *mapping* οι κλάσεις πράγματι μειώθηκαν από 20 σε 10 ενώ τα δείγματα από περίπου 100 σε κάθε κλάση, αυξήθηκαν κατ' ελάχιστο σε περίπου 150, με μερικές να έχουν λίγο πάνω από 300 και την μεγαλύτερη να έχει λίγο πάνω από 400 δείγματα. Η συγχώνευση των κλάσεων έκανε το σύνολο δεδομένων λιγότερο ισορροπημένο σε σχέση με πριν, παρόλο που μειώθηκε ο αριθμός των κλάσεων.

## Βήμα 5

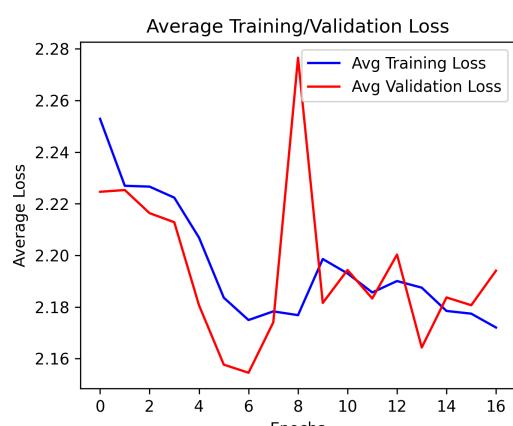
Για την αναγνώριση του μουσικού είδους χρησιμοποιώντας δίκτυα *LSTM*, χρησιμοποιήσαμε και προσαρμόσαμε την χλάση *LSTM*. Επιπλέον φτιάξαμε μία χλάση *Classifier* που χρησιμοποιεί το δίκτυο *LSTM* για εξαγωγή χαρακτηριστικών και κάνει προβλέψεις χρησιμοποιώντας ένα γραφικό επίπεδο μετά τα επίπεδα του *LSTM*. Για να εκπαιδεύσουμε τα δίκτυα δημιουργήσαμε την μέθοδο *train(model, train\_loader, val\_loader, optimizer, epochs = 400, save\_path ='checkpoint.pth', device = "cuda", overfit\_batch = False, regression\_flag = False, patience = 5)* που μπορεί να χρησιμοποιηθεί και για *debugging* με την χρήση του *overfit\_batch = True*. Η συγχεκριμένη συνάρτηση μπορεί να χρησιμοποιήσει *EarlyStopping* ως μέθοδο αποφυγής υπερπροσαρμογής του μοντέλου στα δεδομένα εκπαίδευσης, ενώ αποθηκεύει τα καλύτερη βάρη για μελλοντική χρήση. Τέλος η συνάρτηση επιστρέφει τα *losses* που καταγράφηκαν κατά την εκπαίδευση για καθένα από τα σύνολα δοκιμής και επικύρωσης.



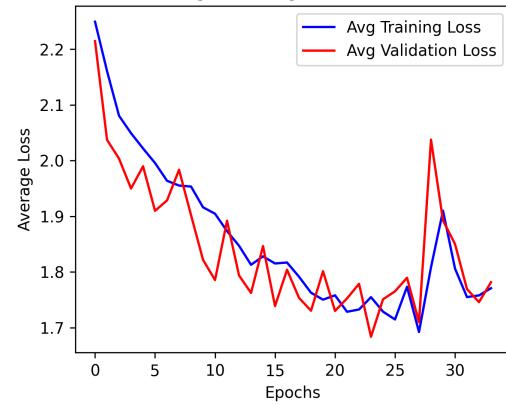
(α') LSTM on Mel Spectograms



(β') LSTM on Beat-Synced Mel Spectograms



(γ') LSTM on Chromagrams



(δ') LSTM on Fused Chromagrams-Spectograms

Σχήμα 6: Training and validation losses the LSTM trained on chromagrams and fused chromagrams and spectrograms.

Κάθε σύνολο δεδομένων χωρίστηκε σε σύνολο εκπαίδευσης και επικύρωσης με αναλογία 80% και 20% αντίστοιχα. Χρησιμοποιήσαμε 8 επίπεδα *LSTM*, με διάσταση 256, ρυθμό μάθησης 0.0001 και *patience* = 10. Παρατηρώντας τα γραφήματα των *losses* εκπαίδευσης και επικύρωσης, το Beat-Synced μοντέλο πετυχαίνει το χαμηλότερο validation loss με τα μοντέλα Fused και Spectrogram να το προσεγγίζουν. Το μοντέλο των Chromagrams έχει την πιο ασταθής και κακή σύγκλιση, κάτι που μας προϊδεάζει πως δεν θα έχει πολύ καλή απόδοση. Φαίνεται πως το καλύτερο μοντέλο θα είναι ανάμεσα στα Beat-Synced, Fused και Spectrogram εκπαιδευμένα μοντέλα.

## Bήμα 6

Η μετρικές που μας ενδιαφέρουν στα προβλήματα ταξινόμησης είναι οι εξής:

- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ : Πρόκειται για το ποσοστό των σωστών προβλέψεων του μοντέλου ως προς όλες τις προβλέψεις που έγιναν.
- $Precision = \frac{TP}{TP+FP}$ : Πρόκειται για το ποσοστό των σωστών προβλέψεων για μία κλάση ως προς όλες τις προβλέψεις που είχαν αυτή την κλάση. Έτσι αντιλαμβανόμαστε την ικανότητα του μοντέλου να προβλέπει την συγκεκριμένη κλάση μόνο όταν βλέπει δείγματα που ανήκουν πραγματικά σε αυτή την κλάση. Μία υψηλή τιμή δείχνει ότι το μοντέλο προβλέπει σωστά τα περισσότερα δείγματα αυτής της κλάσης, ενώ μία χαμηλή τιμή δείχνει ότι το μοντέλο προβλέπει συχνά δείγματα άλλων κλάσεων ως δείγματα αυτής της κλάσης.
- $Recall = \frac{TP}{TP+FN}$ : Πρόκειται για το ποσοστό των σωστών προβλέψεων για μία κλάση ως προς όλα τα δείγματα που ανήκουν πραγματικά στην συγκεκριμένη κλάση. Μία υψηλή τιμή δείχνει ότι το μοντέλο μπορεί να προβλέψει σωστά τα περισσότερα δείγματα της κλάσης ενώ μία χαμηλή τιμή δείχνει το αντίθετο, ότι δηλαδή το μοντέλο θέτει λάθος ετικέτες σε αρκετά δείγματα της συγκεκριμένης κλάσης.
- $F1 - Score = 2 \frac{Recall * Precision}{Recall + Precision}$ : Πρόκειται για τον αρμονικό μέσο μεταξύ των *Precision* και *Recall*. Είναι μία μετρική που ενθύλιακώνει ισορροπημένα και τις δύο προαναφερθέντες μετρικές. Είναι χρήσιμη ως γενική μετρική, όπως το *accuracy*, για να ποσοτικοποιήσουμε τις γενικές ικανότητες του μοντέλου.

Σε προβλήματα ταξινόμησης που δεν είναι δυαδικά, δηλαδή υπάρχουν παραπάνω από δύο κλάσεις, πέρα από τα *class-specific precision, recall* και *f1-score*, μπορούμε να βγάλουμε και τις αντίστοιχες γενικές μετρικές, δηλαδή για όλες τις κλάσεις, με το να βρούμε τις αντίστοιχες μέσες τιμές τους. Υπάρχουν διάφοροι τρόποι βρούμε τον μέσο όρο, όπως *micro* και *macro averaging*. Με την μέθοδο του *macro-averaging*, ουσιαστικά υπολογίζουμε τις *task-specific* μετρικές (*precision, recall, f1-score*) και έπειτα παίρνουμε τον μη βεβαρημένο μέσο όρο της κάθεμιας. Δηλαδή για να βρούμε την οποιαδήποτε *macro-averaged* μετρική απλά διαιρούμε το άθροισμα των συγκεκριμένων *task-specific* μετρικών με τον αριθμό των κλάσεων. Όσο αφορά τις *micro-averages* μετρικές, αυτές προκύπτουν με το *global aggregation* των *TP, FP* και *FN* και στην συνέχεια τον υπολογισμό τις κάθε *micro-averaged* μετρικής. Δηλαδή δεν υπολογίζουμε *task-specific* μετρικές, αντάυτού βρίσκουμε όλα τα *TP, FP, FN* για όλες τις

κλάσεις και υπολογίζουμε τις αντίστοιχες μετρικές.

Ενδεικτικά, με  $C$  να είναι ο αριθμός των κλάσεων:

$$\text{Macro-Averaged Precision} = \frac{\sum_i^C \text{Precision}_i}{C}$$

$$\text{Micro-Averaged Precision} = \frac{\sum_i^C TP_i}{\sum_i^C TP_i + FP_i}$$

Τυάρχει περίπτωση σε κάποια προβλήματα να παρατηρήσουμε μεγάλη απόκλιση μεταξύ του Accuracy και του F1-Score. Αυτό είναι σύνηθες σε ανισοροπή σύνολα δεδομένων όπου κάποιες κατηγορίες έχουν πολύ λίγα δείγματα (minority classes) σε σχέση με με άλλες (majority classes). Για παράδειγμα, έστω ένα δυαδικό πρόβλημα ταξινόμησης με 10 δείγματα για την κλάση 1 και 90 για την κλάση 0 και έστω ότι το μοντέλο μας σημειώνει την εξής επίδοση  $TP = 5, FP = 5, TN = 85, FN = 5$ . Βάση αυτών έχουμε  $Accuracy = 0.9$  και  $F1\text{-Score} = 0.5$ . Αυτή η μεγάλη διαφορά οφείλεται στο γεγονός ότι το accuracy μετράει την γενική ακρίβεια που εδώ επισκιάζεται από την κλάση 0 που έχει το 90% των δειγμάτων ενώ το f1-score που χρησιμοποιεί το precision και το recall δίνει μία πιο σωστή εικόνα για την πραγματική επίδοση του μοντέλου. Αντίστοιχα μπορεί να υπάρχει σημαντική διαφορά ανάμεσα σε micro και macro f1-score, πάλι λόγω ανισορροπίας στο σύνολο δεδομένων. Έστω ένα πρόβλημα ταξινόμησης με την κλάση A να έχει το 90% των δεδομένων και τις κλάσεις B και Γ να έχουν από 5%. Στον υπολογισμό του micro F1-Score θα κυριαρχήσει η κλάση A, λόγω των πολλών δειγμάτων, και δεδομένου ότι τα πάμε καλά σε αυτή θα έχουμε υψηλό micro F1-Score ανεξάρτητα αν δεν τα πάμε καλά για τις κλάσεις B και Γ. Αντίθετα το macro F1-Score, λόγω του τρόπου υπολογισμού του θα δώσει ίση βαρύτητα στην επίδοση της κάθε κλάσης και κατά συνέπεια μια πιο αμερόληπτη μέτρηση της απόδοσης.

Βέβαια υπάρχουν και προβλήματα που μπορεί να μας ενδιαφέρει μία συγκεκριμένη μετρική σε σχέση με τις άλλες. Το precision μας ενδιαφέρει περισσότερο κυρίως σε προβλήματα όπου θέλουμε να ελαχιστοποιήσουμε τα False Positives όπως spam email detection, recommender systems etc. Αντίστοιχα δίνουνε περισσότερη βάση στο recall σε προβλήματα όπου θέλουμε να αποφύγουμε τα false negatives όπως medical diagnosis, alert systems etc. Γενικά τα accuracy και f1-score είναι περισσότερο μετρικές γενικού σκοπού για να αξιολογήσουμε τα μοντέλας μας σε υπό ένα γενικότερο πρίσμα. Αν θέλουμε πιο εξειδικευμένες μετρικές μπορούμε να δούμε precision, recall, κάποιο weighted f1-score ή ακομά και να μελετήσουμε την καμπύλη precision-recall. Πάντα επιλέγουμε μετρικές που είναι στενά συνυφασμένες με τις απαιτήσεις μας για το εκάστοτε πρόβλημα.

Στο δικό μας πρόβλημα, σίγουρα εστιάζοντας μόνο στο accuracy δεν θα έχουμε μία αντιπροσωπευτική εικόνα για τις ικανότητες των μοντέλων μας. Αυτό γιατί το σύνολο δεδομένων μας δεν είναι ισορροπημένο, έτσι μία καλή απόδοση σε κατηγορίες όπως το Rock, που έχουν πολλά δείγματα, θα επισκιάζε τυχών κακές αποδόσεις σε κλάσεις με λίγα δείγματα. Συνεπώς θα χρησιμοποιήσουμε την macro-averaged F1-Score ως την γενική μας μετρική ενώ θα παρατηρήσουμε και τα class-specific precision, recall και f1-score για να αξιολογήσουμε την απόδοση σε επίπεδο κατηγορίας. Για να πάρουμε αυτές τις μετρικές έχουμε φτιάξει την συνάρτηση `get_classification_report(y_pred, y_true)` η οποία επεκτείνει την συνάρτηση `sklearn.metrics.classification_report` με το να δίνει επιπλέον τα micro-averaged precision,

recall και f1-score.

Class	Precision				Recall				F1-Score				Support
	Mel Spec	Beat-Synced	Chroma	Fused	Mel Spec	Beat-Synced	Chroma	Fused	Mel Spec	Beat-Synced	Chroma	Fused	
0	nan	nan	nan	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	40
1	0.50	0.45	nan	0.70	0.55	0.62	0.00	0.74	0.52	0.52	0.00	0.72	40
2	0.34	0.44	0.14	0.35	0.56	0.71	0.07	0.78	0.42	0.54	0.10	0.48	80
3	0.29	0.36	0.25	0.37	0.65	0.61	0.57	0.57	0.40	0.46	0.35	0.45	80
4	0.00	0.20	nan	0.23	0.00	0.30	0.00	0.09	0.00	0.24	0.00	0.13	40
5	nan	0.21	nan	0.11	0.00	0.10	0.00	0.01	0.00	0.14	0.00	0.01	40
6	0.45	0.53	0.26	0.46	0.41	0.62	0.37	0.58	0.43	0.57	0.31	0.51	78
7	nan	nan	nan	nan	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	40
8	0.31	0.41	0.21	0.33	0.40	0.29	0.50	0.31	0.35	0.34	0.30	0.32	103
9	0.20	0.20	nan	0.32	0.06	0.06	0.00	0.09	0.09	0.09	0.00	0.14	34
Macro Avg	0.30	0.35	0.22	0.36	0.26	0.33	0.15	0.32	0.22	0.29	0.11	0.28	575
Micro Avg	0.34	0.39	0.23	0.39	0.34	0.39	0.23	0.39	0.34	0.39	0.23	0.39	575
Weighted Avg	0.32	0.38	0.22	0.36	0.34	0.39	0.23	0.39	0.28	0.34	0.16	0.32	575
Accuracy	0.34	0.39	0.23	0.39									

Table 1: LSTM Evaluation on Spectograms, Beat-Synced Spectograms, Chromagrams and Fused Spectrograms + Chromagrams.

Παρατηρούμε πως το μοντέλο με το υψηλότερο macro-average F1-Score είναι το LSTM που εκπαιδεύτηκε στα Beat-Synced Spectograms, ξεπερνώντας το μοντέλο με τα fused δεδομένα κατά 1%. Το ίδιο συμβαίνει και για το macro-average Recall, ενώ στο macro-averaged precision το μοντέλο με τα fused δεδομένα ξεπερνάει το Beat-Synced μοντέλο κατά 1%. Τα δύο αυτά μοντέλα έχουν ίσες τις micro-average μετρικές τους και ίσο accuracy, αλλά με weighted-averaging το Beat-Synced μοντέλο σημειώνει καλύτερο precision, recall και ίσο f1-score με το fused μοντέλο. Το γεγονός ότι το καλύτερο μοντέλο είναι αυτό που εκπαιδεύτηκε στα Beat-Synced δεδομένα βγάζει νόημα καθώς αυτά τα δεδομένα έχουν το μικρότερο μήκος ακολουθίας. Το χειρότερο μοντέλο είναι με διαφορά αυτό που εκπαιδεύτηκε στα Chromagrams και κάπως καλύτερο αλλά υστερώντας σε σχέση με τα Beat-Synced και Fused μοντέλα είναι το μοντέλο που εκπαιδεύτηκε στα Mel Spectograms. Τέλος, να πούμε ότι τα nan στο precision σημαίνουν πως το μοντέλο δεν έκανε κανένα prediction με την συγκεκριμένη κλάση, το precision = 0 σημαίνει ότι κανένα από τα predictions για την συγκεκριμένη κλάση δεν ήταν σωστά και το recall = 0 σημαίνει πως το μοντέλο απέτυχε να κάνει σωστή πρόβλεψη για την συγκεκριμένη κλάση. Το chromagram μοντέλο είχε τα περισσότερα 0 και nan.

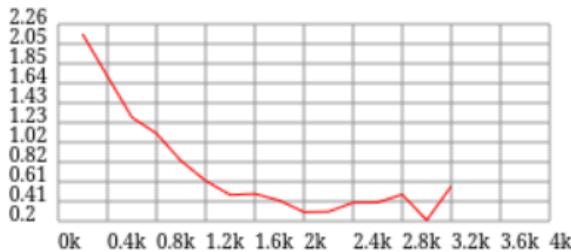
## Βήμα 7.1

1. Convolution (Συνέλιξη) : Η συνέλιξη, στο δικό μας πλαίσιο, πρόκειται για την εφαρμογή ενός φίλτρου (πυρήνα) πάνω σε ένα χάρτη χαρακτηριστικών προκειμένου να εξάγουμε ένα καινούργιο χάρτη χαρακτηριστικών. Πρόκειται για έναν αλγόριθμο κυλιόμενου παραθύρου το οποίο 'περνάει' σταδιακά την είσοδο και υπολογίζει το εσωτερικό γινόμενο των εκάστοτε στοιχείων του παραθύρου. Οι παράμετροι της συνέλιξης είναι το μέγεθος του φίλτρου, το stride που δείχνει πόσο κινείται το φίλτρο ανά βήμα και το padding που προσθέτει συνήθως 0 γύρω από τον χάρτη εισόδου για να ελέγξουμε τις διαστάσεις του χάρτη εξόδου. Γενικά επίπεδα συνέλιξης στην αρχή του δικτύου μαθαίνουν απλά χαρακτηριστικά όπως edges και corners, ενώ τέτοια επίπεδα που βρίσκονται πιο βαθιά στην αρχιτεκτονική μπορούν να αιχμαλωτίσουν πιο σύνθετα χωρικά χαρακτηριστικά που προκύπτουν ως συνδυασμό των πολλών απλών χαρακτηριστικών που έμαθαν τα πρότερα τέτοια επίπεδα
2. Batch Normalization (Ομαλοποίηση Δέσμης): Πρόκειται για μία τεχνική που δρα κατά την εκπαίδευση και στοχεύει στην σταθεροποίηση της αλλά και στην βελτίωση της επίδοσης του μοντέλου. Ουσιαστικά ομαλοποιεί την είσοδο ενός λαφερώστε να έχει  $\mu = 0$  και  $\sigma = 1$ . Έτσι κάθε χαρακτηριστικό  $x_i$  ενός mini-batch γίνεται  $\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$ . Επιπλέον μετά την ομαλοποίηση εφαρμόζεται και ο εξής γραμμικός μετασχηματισμός  $y_i = \gamma \hat{x}_i + \beta$  όπου τα  $\gamma$  (scale) και  $\beta$  (shift) είναι learnable parameters. Γενικά το Batch Normalization μειώνει το covariate shift ανά τις είσοδους των επιπέδων του μοντέλου, επιταχγύνει την σύγκλιση και μειώνει το overfitting.
3. ReLU (Rectified Linear Unit): Πρόκειται για μία συνάρτηση ενεργοποίησης,  $ReLU(x) = \max(0, x)$ , η οποία θέτει στις αρνητικές τιμές το 0 ενώ δεν αλλάζει τις θετικές. Είναι μία από τις πιο συνήθης επιλογές καθώς είναι υπολογιστικά φυληνή, εισάγει μη γραμμικότητα βιοηθώντας στην αναπαράσταση περίπλοκων χαρακτηριστικών καθώς και με τα 0 που εισάγει, οδηγεί σε αραιές ενεργοποιήσεις sparse activations που κάνουν το μοντέλο πιο efficient και βελτιώνουν την γενίκευση της γνώσης.
4. Max Pooling: Πρόκειται για μία τεχνική μείωσης διαστατικότητας (down-sampling method) με κύρια εφαρμογή σε χάρτες χαρακτηριστικών. Είναι ένας αλγόριθμος κυλιόμενου παραθύρου που για κάθε παράθυρο επιλέγει το στοιχείο με την μέγιστη τιμή. Έστι πέρα από την μείωση της διαστατικότητας των χαρτών χαρακτηριστικών, που μειώνει το υπολογιστικό κόστος και ελαττώνει την υπερπροσαρμογή, βοηθά και στην εξαγωγή χαρακτηριστικών καθώς επιλέγει το πιο prominent χαρακτηριστικό κάθε περιοχής.

Αναφορικά με το μοντέλο που εκπαιδεύσαμε στον σύνδεσμο, πρόκειται για ένα συνελικτικό δίκτυο με την εξής αρχιτεκτονική:

```
layer_defs.push({type:'input', out_sx:24, out_sy:24, out_depth:1});
layer_defs.push({type:'conv', sx:5, filters:8, stride:1, pad:2, activation:'relu'});
layer_defs.push({type:'pool', sx:2, stride:2});
layer_defs.push({type:'conv', sx:5, filters:16, stride:1, pad:2, activation:'relu'});
layer_defs.push({type:'pool', sx:3, stride:3});
layer_defs.push({type:'softmax', num_classes:10})
```

Loss:



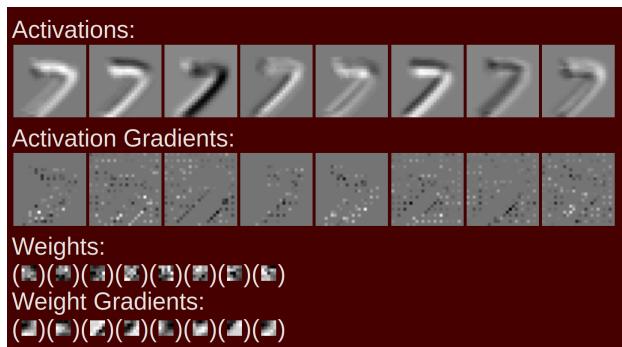
Σχήμα 7: Training Loss Over Training Steps.

Παρατηρώντας το γράφημα βλέπουμε πως το μοντελό έχει συγκλίνει σε πολύ καλό τοπικό ελάχιστο, με το training loss να είναι περίπου 0.2. Για να σχολιάσουμε το μοντέλο πρέπει πρώτα να εξηγήσουμε τι σημαίνουν τα εξής:

1. Activations: Αυτές είναι οι έξοδοι του πρώτου επιπέδου συνελικτικής επεξεργασίας μετά την εφαρμογή της συνάρτησης ενεργοποίησης ReLU στην εικόνα εισόδου.
2. Activation Gradients: Αυτές αντιπροσωπεύουν τις κλίσεις (gradients) της απώλειας (loss) σε σχέση με τις ενεργοποιήσεις, δείχνοντας πώς οι αλλαγές στις ενεργοποιήσεις θα επηρεάσουν την απώλεια.
3. Weights: Τα βάρη αντιπροσωπεύουν τα μαθημένα φίλτρα του εκάστοτε επιπέδου.
4. Weight Gradient: Οι κλίσεις βάρους αντιπροσωπεύουν τον τρόπο με τον οποίο αλλάζει η απώλεια σε σχέση με τα βάρη του φίλτρου κατά την οπισθοδιάδοση. Αυτές οι κλίσεις παρέχουν πληροφορίες σχετικά με το ποια βάρη ενημερώνονται πιο πολύ. Τα φίλτρα με πιο έντονες κλίσεις υποδεικνύουν ότι τα εν λόγω χαρακτηριστικά είναι επί του παρόντος πιο σημαντικά για τη μείωση της απώλειας.

Αναφορικά με το τι μαθαίνει το μοντέλο, ακολουθούν τα παρακάτω γραφήματα.

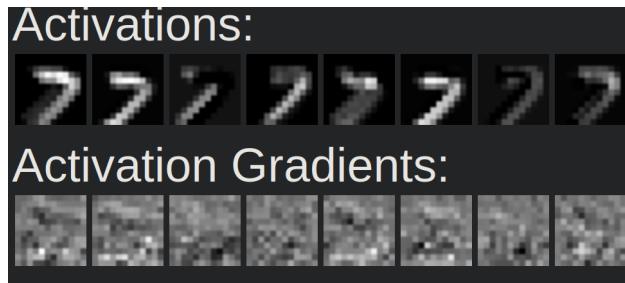
Πρώτο Συνελικτικό Επίπεδο (sx:5, filters:8, stride:1, pad:2, activation:'relu'):



Σχήμα 8: Visualization of representations.

- Activations: Οι ενεργοποιήσεις καταγράφουν διάφορες ακμές edges και απλά σχήματα στην εικόνα εισόδου. Για παράδειγμα, οι φωτεινές περιοχές αντιστοιχούν σε υψηλές τιμές ενεργοποίησης, υποδεικνύοντας ότι τα συγκεκριμένα φίλτρα ανιχνεύουν ακμές, γωνίες, ή άλλα μοτίβα σε αυτές τις περιοχές.
- Activation Gradients: Οι κλίσεις είναι αραιές, υποδεικνύοντας ότι δεν συμβάλλουν εξίσου στην απώλεια όλες οι περιοχές των ενεργοποιήσεων. Αυτή η αραιότητα συχνά αντανακλά αποτελεσματική μάθηση, καθώς το μοντέλο εστιάζει στα πιο σημαντικά χαρακτηριστικά.
- Weights: Από την οπτικοποίηση, τα βάρη δείχνουν ανιχνευτές ακμών και μοτίβα κλίσης, τα οποία είναι τυπικά για το πρώτο στρώμα των NN που εργάζονται σε δεδομένα εικόνας.
- Weight Gradients: Παρατηρούμε πως τα βάρη ενημερώνονται ώστε να γίνουν αποτελεσματικότεροι ανιχνευτές ακμών και μοτίβων κλίσης.

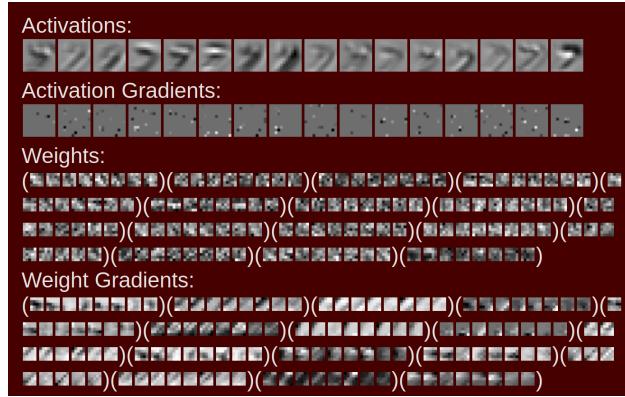
Πρώτο Pool Επίπεδο (type:'pool', sx:2, stride:2):



Σχήμα 9: Visualization of representations.

- Activations: Στην εικόνα, οι ενεργοποιήσεις είναι εμφανώς μειωμένες σε σύγκριση με τις αρχικές συνελίξεις. Τα πιο σημαντικά τμήματα του ψηφίου «7» παραμένουν εμφανή, τονίζοντας τη διατήρηση των χαρακτηριστικών υψηλού επιπέδου, όπως οι ακμές και τα σχήματα.
- Activation Gradients: Οι κλίσεις εμφανίζονται λιγότερο αραιές, με εστιασμένες περιοχές έντασης. Αυτό υποδηλώνει ότι το μοντέλο δίνει έμφαση σε συγκεκριμένα χαρακτηριστικά, ενώ απορρίπτει λιγότερο σημαντικά μέρη, γεγονός που ευθυγραφιζεται με τον στόχο της μείωσης της διαστατικότητας της λειτουργίας συγκέντρωσης.

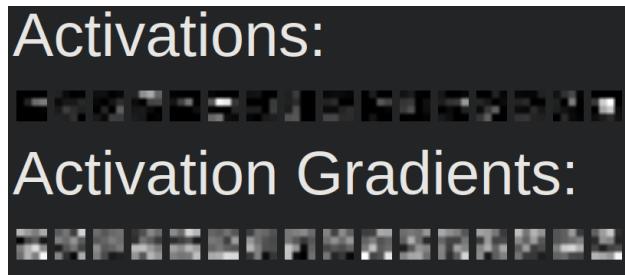
Δεύτερο Συνελικτικό Επίπεδο (sx:5, filters:8, stride:1, pad:2, activation:'relu'):



Σχήμα 10: Visualization of representations.

- **Activations:** Αυτές οι απεικονίσεις αποκαλύπτουν πιο σύνθετα και αφηρημένα χαρακτηριστικά σε σύγκριση με το πρώτο στρώμα συνελίξεων. Τα φίλτρα σε αυτό το στρώμα καταγράφουν μοτίβα που αποτελούν συνδυασμούς των χαρακτηριστικών χαμηλότερου επιπέδου (π.χ. ακμές και γραμμές) που ανιχνεύονται στο πρώτο στρώμα. Για παράδειγμα, οι ενεργοποιήσεις αναδεικνύουν τμήματα του ψηφίου «7» με πιο εντοπισμένους και περίπλοκους τρόπους, εστιάζοντας σε υποπεριοχές όπως καμπύλες ακμές ή διασταυρώσεις.
- **Activation Gradients:** Οι κλίσεις είναι αφαιές και εστιασμένες, γεγονός που υποδηλώνει ότι μόνο συγκεκριμένες ενεργοποιήσεις επηρεάζουν σημαντικά την απώλεια. Αυτή η αφαιότητα υποδηλώνει αποτελεσματική οπισθοδιάδοση, όπου το μοντέλο δίνει έμφαση στα πιο κρίσιμα χαρακτηριστικά, ενώ καταστέλλει τις άσχετες ή περιττές πληροφορίες. Τα μοτίβα των κλίσεων αντιστοιχούν σε λεπτές προσαρμογές, που βελτιώνουν τους χάρτες χαρακτηριστικών που μαθαίνονται από τις ενεργοποιήσεις.
- **Weights:** Οι απεικονίσεις των βαρών εμφανίζουν ποικίλα μοτίβα, όπως καμπύλες, διαγώνιες και σταυροειδείς δομές, τα οποία είναι απαραίτητα για την αποτύπωση των περίπλοκων λεπτομερειών των σχημάτων των ψηφίων. Αυτά τα φίλτρα λειτουργούν σε συνδυασμό με τις εξόδους του πρώτου στρώματος συνελικτικής ανάλυσης για να σχηματίσουν μια πιο ισχυρή ιεραρχία χαρακτηριστικών.
- **Weight Gradients:** Ορισμένα φίλτρα παρουσιάζουν ισχυρότερες κλίσεις, γεγονός που υποδηλώνει ότι είναι πιο κρίσιμα για τη μείωση της απώλειας ή ότι τα αντίστοιχα χαρακτηριστικά τους υποεκπροσωπούνται και απαιτούν περαιτέρω συντονισμό.

Δεύτερο Pool Επίπεδο (sx:3, stride:3):

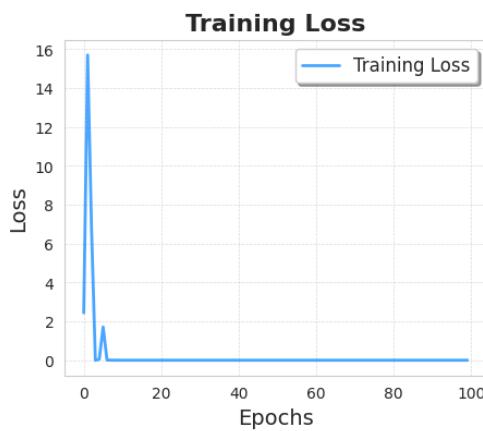


Σχήμα 11: Visualization of representations.

- Activations: Στην εικόνα, οι ενεργοποιήσεις είναι πάλι εμφανώς μειωμένες σε σύγκριση με πριν συνελίξεις. Διατηρούνται πάλι αφηρημένα αλλά υψηλότερου επιπέδου χαρακτηριστικά.
- Activation Gradients: Οι κλίσεις εμφανίζονται λιγότερο αραιές, με εστιασμένες περιοχές έντασης. Αυτό υποδηλώνει ότι το μοντέλο δίνει έμφαση σε συγκεκριμένα χαρακτηριστικά, ενώ απορρίπτει λιγότερο σημαντικά μέρη.

### Εκπαίδευση CNN στο Genre Classification Task

Αξιοποιώντας την έτοιμη υλοποίηση του CNNBackbone που μας παρέχεται και ορίζοντας τις κατάλληλες παραμέτρους στην συγκεκριμένη κλάση, εκπαίδευσα πειραματικά το μοντέλο μου σε ένα μοναδικό batch ώστε να προκαλέσω overfitting, όπως υποδεικνύεται, για να ελέγξω ότι όντως λειτουργεί με τον αναμενόμενο τρόπο και εκπαίδευται.

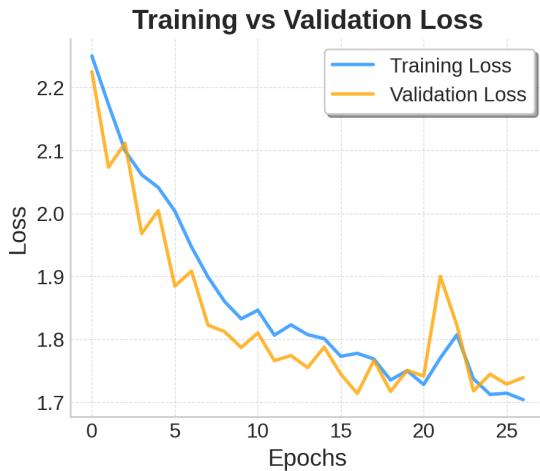


Σχήμα 12: Training loss when overfitting CNN on a single batch

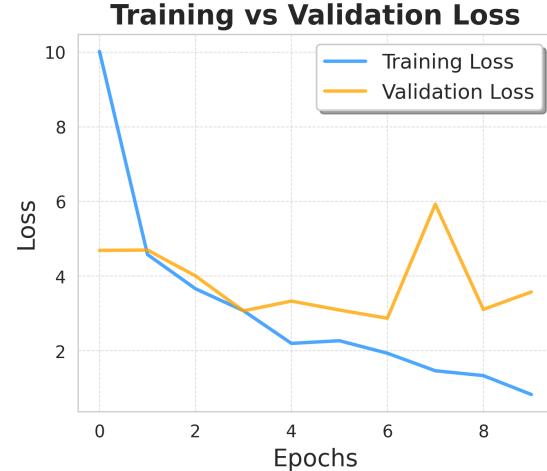
Πράγματι, παρατηρούμε ότι το μοντέλο υπερπροσαρμόζεται όπως έπρεπε, άρα είμαστε έτοιμοι να προχωρήσουμε στην εκπαίδευση πάνω στο fma\_genre\_spectrograms dataset.

Να επισημάνουμε σε αυτό το σημείο τις παραμέτρους με τις οποίες τρέχουμε το CNN Backbone μοντέλο, οι οποίες έχουν ως εξής:

- input\_shape=128, αριθμός frequency bins φασματογραφημάτων
- in\_channels=1, αριθμός καναλιών εισόδου του μοντέλου
- filters=[32, 64, 128, 256], αριθμός κελιών ανά επίπεδο
- feature\_size=256, μέγεθος του διανύσματος που βγαίνει από το τελευταίο hidden layer



(a) LSTM on Mel Spectograms



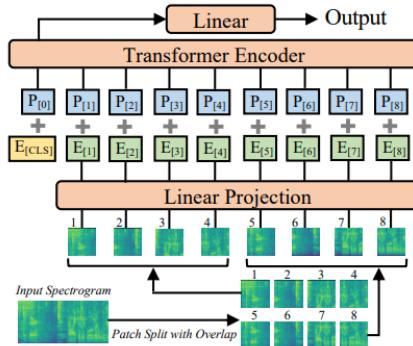
(b) CNN on Mel Spectograms

Ξεκινάω συγχρίνοντας τις αρχιτεκτονικές των δικτύων LSTM & CNN και επεξηγώντας γιατί το τελευταίο είναι καταδικασμένο να πετύχει καλύτερα αποτελέσματα από τη φύση του. Τα LSTM είναι RNN που σχεδιάστηκαν για να μοντελοποιούν χρονικές ακολουθίες ή δεδομένα με εξάρτηση από τη χρονική σειρά. Είναι ιδιαίτερα χρήσιμα για δεδομένα όπου οι χρονικές συσχετίσεις είναι σημαντικές. Ωστόσο, τα φασματογραφήματα έχουν δύο διαστάσεις, τον χρόνο και τη συχνότητα. Η αρχιτεκτονική αυτή δεν είναι βέλτιστη για την επεξεργασία τέτοιων δισδιάστατων δεδομένων, καθώς η δομή τους βασίζεται στην αλληλουχία και όχι στην ανάλυση εικόνας. Αντιθέτως, τα CNN είναι ειδικά σχεδιασμένα για δισδιάστατα δεδομένα, όπως εικόνες ή, στην περίπτωσή μας, φασματογραφήματα. Τα συνελικτικά φίλτρα που χρησιμοποιούν επιτρέπουν την εξαγωγή τοπικών χαρακτηριστικών, αξιοποιώντας τις χωρικές συσχετίσεις που υπάρχουν σε δεδομένα όπως τα φασματογραφήματά μας. Είναι περισσότερο αποδοτικά σε δεδομένα με ισχυρές χωρικές πληροφορίες, καθώς επιτρέπουν την ανάλυση διαφορετικών επιπέδων χαρακτηριστικών.

Πρακτικά, παρατηρούμε ότι το CNN ξεκινάει από πιο υψηλό loss αλλά σύντομα καταφέρνει να το μειώσει. Στο τέλος συγκαλίνουν σε παρόμοια μεγέθη με αυτά του LSTM, μόνο που το CNN είναι πιο αποδοτικό από το προηγούμενο. Έχουμε ήδη προηγουμένως εκτιμήσει ότι το βέλτιστο metric για να μετρήσουμε την απόδοση των μοντέλων στο dataset μας είναι οι μετρικές macro. Το CNN υπερέχει του LSTM, παρουσιάζοντας κατά 5% υψηλότερο macro-avg precision, 3% καλύτερο macro-avg recall και 5% υψηλότερο macro-avg F1-score, καταδεικνύοντας καλύτερη γενική απόδοση σε όλες τις κλάσεις. Άρα οι θεωρητικές μας εκτιμήσεις επιβεβαιώνονται από την πρακτική μας ενασχόληση.

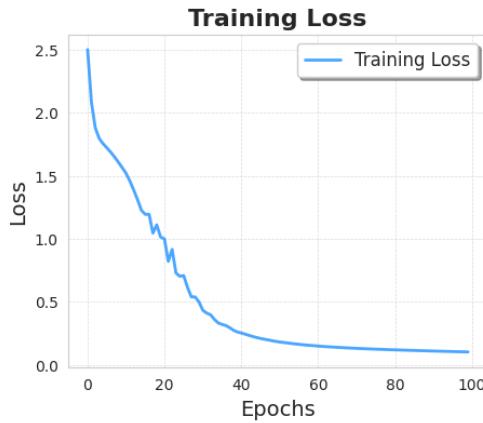
## Βήμα 7.2

**α)** Το μοντέλο που αξιοποιούμε σε αυτό το βήμα είναι ο Audio Spectrogram Transformer. Από την δημοσίευση που μας παρέχει η εκφώνηση, αντλούμε χρήσιμες πληροφορίες όσον αφορά, την μέχρι τώρα προσέγγιση των επιστημόνων στο πρόβλημα που αντιμετωπίζουμε, τον τρόπο λειτουργίας και τα πολλά υποσχόμενα πρώτα αποτελέσματα του συγκεκριμένου μοντέλου. Όπως φαίνεται, τα CNN (pure ή hybrid με attention mechanism) είχαν την μερίδα του λέοντος μέχρι πρότινος, ως κύρια μέθοδος για την ταξινόμηση ήχου, εκπαιδευόμενα με supervised learning πάνω σε φασματογραφήματα ήχου. Το AST παρουσιάστηκε ως το πρώτο μη συνελικτικό μοντέλο που αξιοποιήθηκε για ταξινόμηση ήχου. Η προσπάθεια αυτή στέφθηκε με επιτυχία αφού τα κατάφερε αποτελέσματα που μέχρι τότε δεν είχαμε δει. Το μοντέλο AST δέχεται ως είσοδο Mel-Spectrograms (128 χαρακτηριστικών) που έχουν εξαχθεί από τα αρχεία ήχου του συνόλου δεδομένων. Τα φασματογραφήματα αρχικά διαιρούνται σε αλληλοεπικαλυπτόμενα patches, όπου το καθένα μετατρέπεται σε ένα μονοδιάστατο embedding (μεγέθους 768) μέσω ενός επιπέδου linear projection. Έχουμε επίσης ένα ακόμη embedding (ίδιου μεγέθους) για να διατηρήσουμε την χωρική πληροφορία του input και ένα [CLS] token στην αρχή της ακολουθίας διότι επιχειρούμε supervised learning. Η ακολουθία αυτή διέρχεται από τον Transformer (συγκεκριμένα χρησιμοποιείται η αρχιτεκτονική του original Transformer που προτάθηκε στην γνωστή δημοσίευση Attention Is All You Need), όπου αξιοποιούνται μόνο τα encoder layers αφού έχουμε να κάνουμε με ταξινόμηση. Πιο συγκεκριμένα, ο Transformer μας αποτελείται από 12 encoder layers, με κάθε layer να περιέχει 12 attention heads, με embeddings διαστάσεων 768. Το output representation του encoder περνάει από ένα γραμμικό επίπεδο με σιγμοειδή συνάρτηση ενεργοποίησης το οποίο υλοποιεί την ταξινόμηση σε κλάσεις.



Σχήμα 14: Audio Spectrogram Transformer Architecture

β) Αξιοποιώντας την έτοιμη υλοποίηση του ASTBackbone που μας παρέχεται και ορίζοντας τις κατάλληλες παραμέτρους στην συγκεκριμένη κλάση, εκπαίδευσα πειραματικά το μοντέλο μου σε ένα μοναδικό batch ώστε να προκαλέσω overfitting, όπως υποδεικνύεται, για να ελέγξω ότι όντως λειτουργεί με τον αναμενόμενο τρόπο και εκπαιδεύεται.

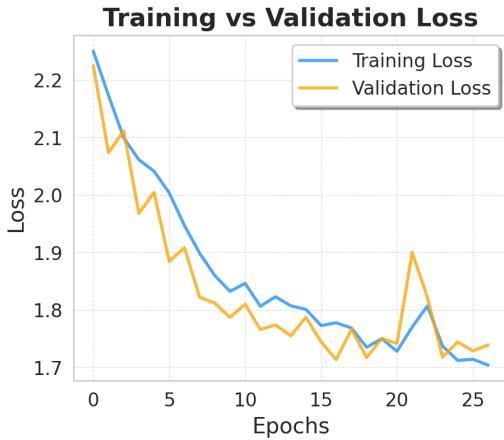


Σχήμα 15: Training loss when overfitting on a single batch

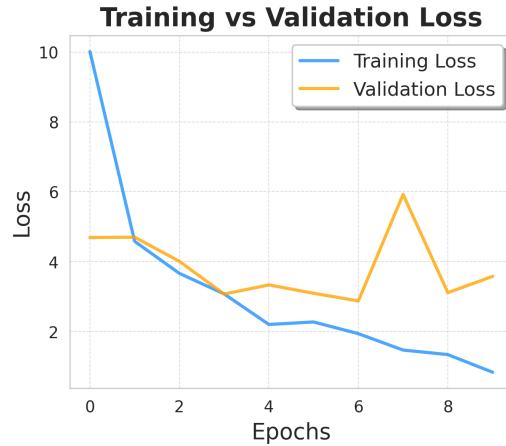
Πράγματι, παρατηρούμε ότι το μοντέλο υπερπροσαρμόζεται όπως θα έπρεπε, άρα είμαστε έτοιμοι να προχωρήσουμε στην εκπαίδευση πάνω στο σύνολο δεδομένων. Να επισημάνουμε σε αυτό το σημείο τις παραμέτρους με τις οποίες τρέχουμε το ASTBackbone μοντέλο, οι οποίες έχουν ως εξής:

- fstride=10 (by default), με τη βοήθεια του ορίζεται η επικάλυψη στον άξονα συχνοτήτων
- tstride=10 (by default), με τη βοήθεια του ορίζεται η επικάλυψη στον άξονα χρόνου
- input\_fdim=128, αριθμός frequency bins φασματογραφημάτων
- input\_tdim=1293, αριθμός χρονικών βημάτων φασματογραφημάτων
- imagenet\_pretrain=True, χρήση ImageNet weights γιατί επιτυγχάνουμε καλύτερα αποτελέσματα, κάτι το οποίο θα δουμε εκτενέστερα στο βήμα που αφορά το transfer learning
- model\_size='tiny224', επιλογή μικρού μοντέλου λόγω περιορισμένων πόρων
- feature\_size=10, αριθμός κλάσεων του dataset target

Προτού παρουσιάσουμε τα αποτελέσματα εκπαίδευσης του AST, να συνοψίσουμε λίγο με βάση τη θεωρία. Ο Audio Spectrogram Transformer αντικειτωπίζει τα φασματογραφήματα ως εικόνες όπως και το CNN, αλλά πέραν άλλων χρησιμοποιεί και τον μηχανισμό self-attention για να μοντελοποιήσει παγκόσμιες εξαρτήσεις και σχέσεις μεταξύ όλων των περιοχών του φασματογραφήματος. Αυτό του επιτρέπει να συλλαβεί πιο σύνθετες χρονικές και συχνοτικές δομές, κάτι που το CNN δυσκολεύεται να επιτύχει. Σε σύγκριση με το LSTM, που επεξεργάζεται δεδομένα διαδοχικά με περιορισμό σε μακροπρόθεσμες εξαρτήσεις, ο AST προσφέρει παράλληλη επεξεργασία και καλύτερη αποδοτικότητα εκπαίδευσης.



(a) Long Short-Term Memory Network  
(Batch Size: 32)



(b) Convolutional Neural Network  
(Batch Size: 8)

Ο AST επιτυγχάνει καλύτερη γενίκευση σε μεγάλα σύνολα δεδομένων χάρη στις πλουσιότερες αναπαραστάσεις που εξάγει, αλλά οι υψηλές υπολογιστικές απαιτήσεις τον καθιστούν λιγότερο ευέλικτο σε μικρότερα datasets. Η χρήση του pretrained on ImageNet AST ήταν καθοριστική για την επίτευξη των αποτελεσμάτων που βλέπουμε παρακάτω.

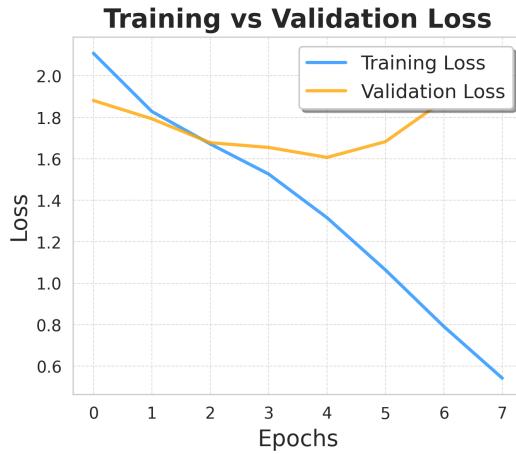


Figure 17: Audio Spectrogram Transformer  
(Batch Size: 8)

Τα διαγράμματα μας δημιουργούν την ψευδαίσθηση ότι το LSTM και ο AST παρουσιάζουν παρόμοια σύγκλιση σε διαφορετική ταχύτητα. Ωστόσο, είναι σημαντικό να τονίσουμε εδώ ότι ο AST εκπαιδεύτηκε με μέγεθος παρτίδας 8, ενώ το LSTM με 32. Αυτό δίνει στον LSTM φαινομενικό προβάδισμα σε σταθερότητα εκπαίδευσης, καθώς το μεγαλύτερο μέγεθος παρτίδας εξομαλύνει τις ενημερώσεις των παραμέτρων και μειώνει την πιθανότητα υπερπροσαρμογής. Αντίθετα, ο Transformer, λόγω της αυξημένης υπολογιστικής πολυπλοκότητας, απαιτεί μικρότερες παρτίδες για να είναι δυνατόν να εκπαιδευθεί σε ένα κοινό μηχάνημα. Παρά αυτήν την αναγκαστική ρύθμιση, ο AST επιδεικνύει καλύτερη απόδοση και σταθερότητα αφού αξιοποιεί το self-attention.

## Βήμα 8

Στο συγκεκριμένο ερώτημα θα εκπαιδεύσουμε 3 μοντέλα σε 3 προβλήματα παλινδρόμησης. Δεδομένων των φασματογραφημάτων κάθε τραγουδιού θα πρέπει να προβλέψουμε μια τιμή για τις εξής τρεις συναίσθηματικές διαστάσεις:

- Valence (πόσο θετικό ή αρνητικό είναι το συναίσθημα), όπου αρνητικό κοντά στο 0, θετικό κοντά στο 1.
- Energy (πόσο ισχυρό είναι το συναίσθημα), όπου ασθενές κοντά στο 0, ισχυρό κοντά στο 1.
- Danceability (πόσο χορευτικό είναι το τραγούδι), όπου μη χορευτικό κοντά στο 0, χορευτικό κοντά στο 1.

Θα συγκρίνουμε τρία μοντέλα με διαφορετικό backbone (CNN, LSTM, AST) με τελική μετρική το μέσο spearman corellation coefficient ανάμεσα στις πραγματικές και στις προβλεπόμενες τιμές για όλους τους άξονες. Κατά την εκπαίδευση χρησιμοποιούμε το μέσο τετραγωνικό σφάλμα ως συνάρτηση κόστους, τον βελτιστοποιητή Adam με ρυθμό μάθησης 0.0001, batch\_size = 8, 80/20 split για τα σύνολα εκπαίδευσης και επικύρωσης και patience = 10. Το LSTM μοντέλο έχει 2 χρυφά layers μεγέθους 64, το CNN έχει 4 συνελικτικά επίπεδα με διαστάσεις εξόδου 32, 64, 128 και 256 και το AST μοντέλο είναι το pre-trained tiny224. Ακολουθούν τα διαγράμματα των losses των συνόλων επικύρωσης και εκπαίδευσης κατά την εκπαίδευση.

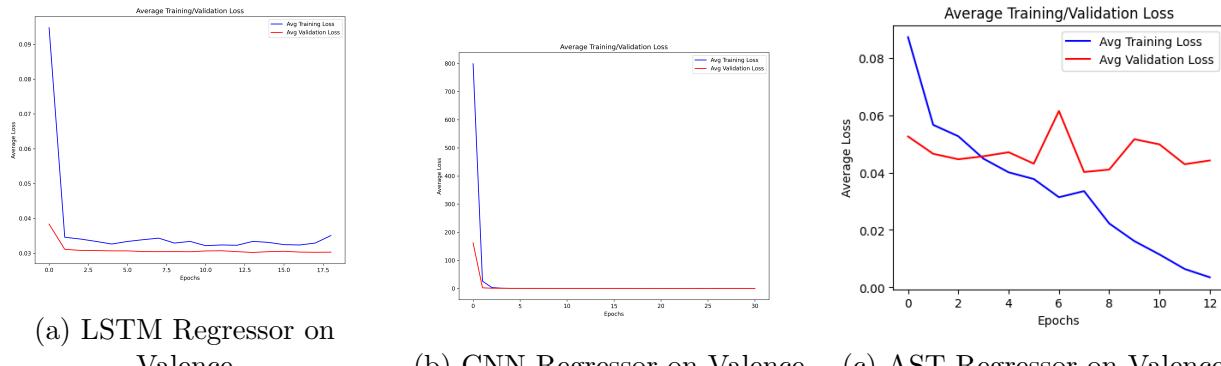
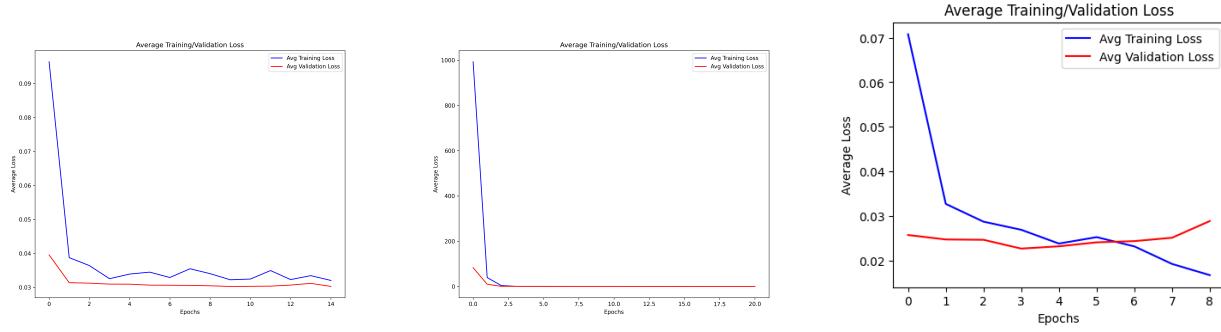


Figure 18: Training and Validation losses for Valence Regression using different backbones.

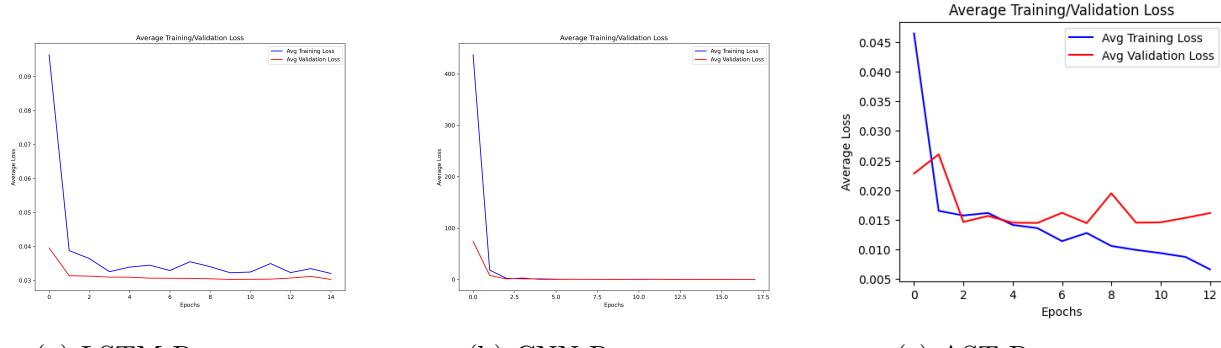
Παρατηρώντας τα γραφήματα για το Valence, βλέπουμε πως όλα τα μοντέλα είχαν πολύ καλή σύγκλιση (σε σημεία πολύ κοντά στο 0), με το AST μοντέλο να χρειάζεται τις λιγότερες εποχές μιας και είναι προεκπαιδευμένο. Το CNN στην αρχή πρέπει να ξεκίνησε από πολύ κακό σημείο, λόγω του υπερβολικά υψηλού πρώτου loss, ωστόσο αμέσως σταθεροποιήθηκε. Γενικά είμαστε πολύ ευχαριστημένοι από την σύγκλιση.



(a) LSTM Regressor on Energy    (b) CNN Regressor on Energy    (c) AST Regressor on Energy

Figure 19: Training and Validation losses for Energy Regression using different backbones.

Παρατηρώντας τα γραφήματα για το Energy, έχουμε πολύ παρόμοια εικόνα με τα γραφήματα του Valence. Δηλαδή, βλέπουμε πως όλα τα μοντέλα είχαν πολύ καλή σύγκλιση (σε σημεία πολύ κοντά στο 0), με το AST μοντέλο να χρειάζεται τις λιγότερες εποχές μιας και είναι προεκπαιδευμένο. Το CNN στην αρχή πρέπει να ξεκίνησε από πολύ κακό σημείο, λόγω του υπερβολικά υψηλού πρώτου loss, ωστόσο αμέσως σταθεροποιήθηκε. Γενικά είμαστε πολύ ευχαριστημένοι από την σύγκλιση και εδώ.



(a) LSTM Regressor on Danceability

(b) CNN Regressor on Danceability

(c) AST Regressor on Danceability

Figure 20: Training and Validation losses for Danceability Regression using different backbones.

Παρατηρώντας τα γραφήματα για το Danceability, έχουμε πολύ παρόμοια εικόνα με τα προηγούμενα γραφήματα. Δηλαδή, βλέπουμε πως όλα τα μοντέλα είχαν πολύ καλή σύγκλιση (σε σημεία πολύ κοντά στο 0), με το AST μοντέλο να χρειάζεται τις λιγότερες εποχές μιας και είναι προεκπαιδευμένο. Το CNN στην αρχή πρέπει να ξεκίνησε από πολύ κακό σημείο, λόγω του υπερβολικά υψηλού πρώτου loss, ωστόσο αμέσως σταθεροποιήθηκε. Γενικά είμαστε πολύ ευχαριστημένοι από την σύγκλιση και εδώ.

Αναφορικά με τον συντελεστή Spearman, ουσιαστικά ποσοτικοί είναι την μονοτονική σχέση μεταξύ δύο μεταβλητών. Αυτό σημαίνει πόσο ανάλογη είναι η αύξηση ή η μείωση των δύο μεταβλητών. Εμείς θέλουμε έναν υψηλό, κατά απόλυτη τιμή κοντά στο 1, τέτοιο συντελεστή κανόνας αυτό σημαίνει πως ακόμα και αν οι προβλέψεις δεν είναι αριθμητικά τέλειες, δηλαδή ίδιες

Metric	Valence			Energy			Danceability		
	LSTM	CNN	AST	LSTM	CNN	AST	LSTM	CNN	AST
<b>Spearman</b>	0.123	0.191	<b>0.526</b>	0.453	-0.038	<b>0.728</b>	0.264	0.207	<b>0.624</b>
<b>MSE</b>	0.067	0.425	<b>0.044</b>	0.046	4.356	<b>0.029</b>	0.029	1.439	<b>0.016</b>
<b>MAE</b>	0.220	0.518	<b>0.170</b>	0.169	1.973	<b>0.132</b>	0.142	0.978	<b>0.097</b>
<b>RMSE</b>	0.259	0.652	<b>0.210</b>	0.214	2.087	<b>0.170</b>	0.171	1.199	<b>0.127</b>

Table 2: Performance Metrics for Valence, Energy, and Danceability Across Models.

με την επισημείωση, η σχετική τους κατάταξη (ordering) είναι συνεπής με την πραγματική τιμή, κάτι που συχνά αποτελεί χρίσιμο στοιχείο της πρόβλεψης.

Παρατηρώντας τον πίνακα το καλύτερο μοντέλο και στους τρεις συναισθηματικούς άξονες είναι αυτό που χρησιμοποιεί τον AST ως backbone καθώς σημειώνει τον υψηλότερο συντελεστή Spearman άλλα και τις μικρότερες τιμές MSE, MAE, RMSE κάθε φορά. Ειδικά όσο αναφορά τον συντελεστή συσχέτισης η διαφορά ανάμεσα στο AST και τα υπόλοιπα μοντέλα είναι αρκετά μεγάλη και στους τρεις άξονες. Το LSTM μοντέλο είναι δεύτερο καθώς σημειώνει τα δεύτερα χαμηλότερα MSE, MAE, RMSE κάθε φορά και το δεύτερο μεγαλύτερο συντελεστή Spearman, εκτός από το Valence. Αυτό όμως δεν σημαίνει πως για το Valence δεν είναι το δεύτερο καλύτερο μοντέλο αφού η διαφορά των MSE, MAE, RMSE που σημείωσε σε σχέση με το CNN είναι αρκετά μεγάλη. Τέλος το CNN μοντέλο είναι το χειρότερο με τα μεγαλύτερα MSE, MAE, RMSE και μικρότερο συντελεστή Spearman κάθε φορά. Ειδικά στο Energy Regression βλέπουμε πως το CNN έχει συντελεστή Spearman οριακά 0 σε σχέση με τα άλλα δύο μοντέλα. Δεν αποτελεί έκπληξη πως το AST είναι το καλύτερο μοντέλο μιας και χρησιμοποιεί μηχανισμούς Attention και είναι και προεκπαιδευμένο.

## Bήμα 9

**α)** Η δημοσίευση καταλήγει στο συμπέρασμα ότι τα εκπαιδευμένα βάρη βαθιών νευρωνικών δικτύων, είναι πιο γενικά στα πρώτα επίπεδα, δηλαδή έχουν την δυνατότητα να εφαρμοστούν σε διάφορετα tasks, ενώ τα υψηλότερα παρουσιάζουν μεγαλύτερη εξειδίκευση στο εν λόγω σύνολο δεδομένων. Το transferability των βαρών μειώνεται όσο αυξάνεται η σχετική απόσταση μεταξύ των tasks, αλλά η αξιόποιηση transfer learning, ακόμα και μεταξύ μη σχετικών tasks, υπερτερεί της τυχαίας αρχικοποίησης βαρών. Τέλος, οι συγγραφείς υποστηρίζουν ότι η αρχικοποίηση ενός νευρωνικού δικτύου με προεκπαιδευμένα βάρη προσδίδει σταθερά καλύτερη γενίκευση.

**β)** Το μοντέλο που επιλέγω για να εφαρμόσω Transfer Learning είναι, χωρίς δεύτερη σκέψη, ο AST λόγω της ιδιαίτερης αρχιτεκτονικής του Transformer. Οι Transformers βασίζονται στον self-attention μηχανισμό, ο οποίος τους βοηθάει να ανιχνεύουν σύνθετες συσχετίσεις μεταξύ των δεδομένων. Αυτό έχει ως αποτέλεσμα την εξαγωγή γενικών και επαναχρησιμοποιήσιμων χαρακτηριστικών, τα οποία είναι περισσότερο μεταβιβάσιμα σε διαφορετικά datasets σε σύγκριση με τα τοπικά χωρικά μοτίβα που μαθαίνουν τα CNN ή τις χρονικές συσχετίσεις που μαθαίνουν

τα LSTM. Η ικανότητά του να κατανοεί περίπλοκες σχέσεις στα δεδομένα τον καθιστά ιδανικό για προβλήματα με πολυδιάστατη φύση, όπως η εκτίμηση συναισθημάτων. Γενικότερα, ακόμη και διαισθητικά να το προσεγγίσουμε, το transfer-learning ταιριάζει άρρητα στην φύση/αρχιτεκτονική των Transformers, ενώ θα είχε αμφιβόλου μέτρου επιτυχία στους άλλους δύο τύπους δικτύων.

**γ & δ)** Στο προηγούμενο βήμα εκπαιδεύσαμε τα τρία μοντέλα μας (LSTM, CNN, AST) στα τρία διαφορετικά tasks εκτίμησης συναισθήματος-συμπεριφοράς (valence, energy, danceability) με χρήση παλινδρόμησης. Εκεί χρησιμοποιήσαμε τον pretrained on ImageNet AST κάτι το οποίο μας έδωσε ένα σημαντικό performance boost σε σχέση με τα υπόλοιπα μοντέλα. Μπορεί έτσι κάποιος δικαιολογημένα να θεωρήσει ότι έχουμε ήδη εφαρμόσει transfer learning. Ωστόσο, αυτό το κάναμε βασικά γιατί η εκπαίδευση του Transformer απαιτεί μεγάλο σύνολο δεδομένων και αρκετούς υπολογιστικούς πόρους για να γίνει εκτενής εκπαίδευση. Επειδή δεν έχουμε τα παραπάνω στην διάθεση μας, αξιοποιούμε τα pretrained βάρη για να επιτύχουμε μια καλύτερη αρχικοποίηση τους. Σε αυτό το βήμα λοιπόν, εκπαιδεύουμε αρχικά τον pretrained on ImageNet AST στο Genre Classification Task. Εδώ θα υλοποιήσουμε την λειτουργικότητα του transfer-learning / fine-tuning, όπου εκπαιδεύουμε το μοντέλο μας, στο task εκτίμησης συναισθήματος-συμπεριφοράς, παγώνουμε τις παραμέτρους όλων των encoder layers του μοντέλου, πλην των τελευταίων δύο, ως εξής:

```

1 # for each layer we want to freeze
2 for param in layer.parameters():
3     param.requires_grad = False
4
5 # update only the trainable parameters of the model
6 optimizer = optim.Adam(
7     filter(lambda p: p.requires_grad, energy_model.parameters()),
8     lr=LR
9 )

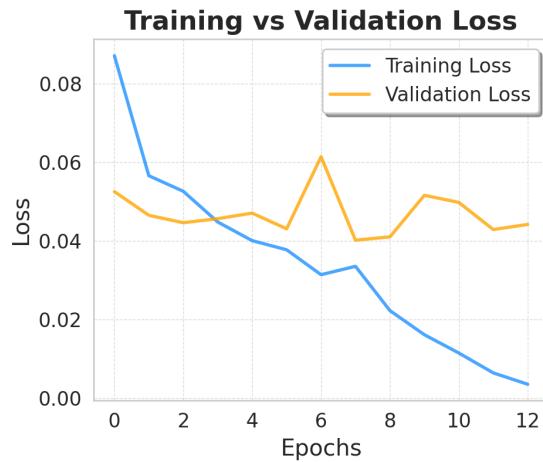
```

Εκπαιδεύω τον AST ως προς valence, energy & danceability για σύγκριση. Επιλέγω το danceability για τον σχολιασμό μου, αφού είναι το πιο αποδοτικό όπως φαίνεται παρακάτω.

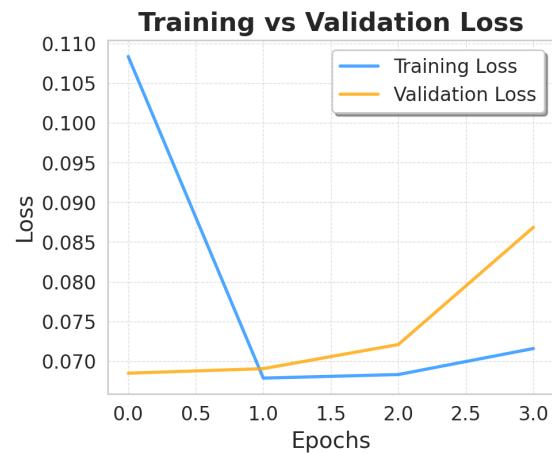
**ε)** Θα κάνω τον σχολιασμό μου στην απόδοση του fine-tuned AST ως προς το danceability task, που φαίνεται να είναι το πιο αποδοτικό γενικότερα, πιθανόν λόγω απλοϊκότητας της φύσεως του συγκεκριμένου προβλήματος.

Στην αριστερή στήλη, μεσαία γραμμή της παραπάνω σελίδας, παρουσιάζεται το διάγραμμα των losses για το AST που εκπαιδεύτηκε αποκλειστικά στο danceability dataset, με αρχικοποίηση βαρών από το ImageNet pretrained μοντέλο. Στη δεξιά στήλη, εμφανίζεται το διάγραμμα των losses για το AST που υποβλήθηκε σε fine-tuning στο danceability dataset, αφού προηγουμένως είχε εκπαιδευτεί (pre-trained) στο FMA genre classification task (επίσης με βάρη από το ImageNet pretrained μοντέλο).

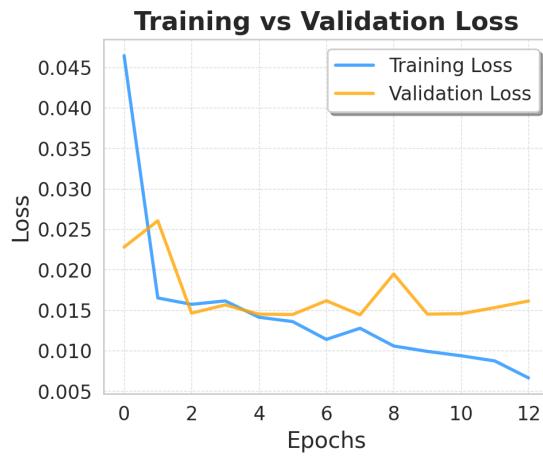
Παρατηρείται ότι και τα δύο μοντέλα παρουσιάζουν παρόμοια σύγκλιση και απόδοση, κάτι που εξηγείται από την κοινή αρχικοποίηση βαρών από το ImageNet pretrained μοντέλο. Τα βάρη αυτά προκύπτουν από εκπαίδευση σε ένα πολύ μεγάλο και πλούσιο σύνολο δεδομένων (ImageNet), το οποίο υπερβαίνει κατά πολύ το μέγεθος και την ποικιλία του danceability dataset ή του FMA genre dataset.



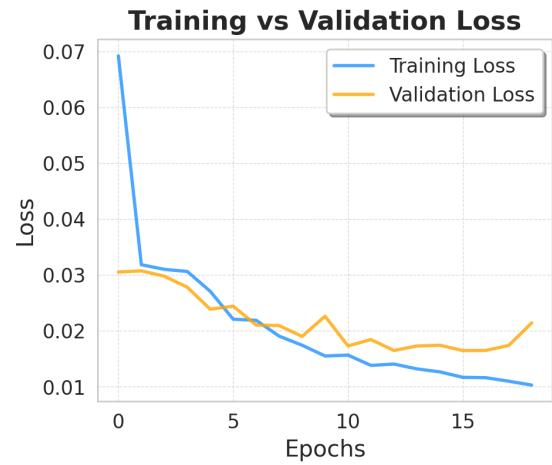
(a) Trained on Valence  
AST with ImageNet pretraining



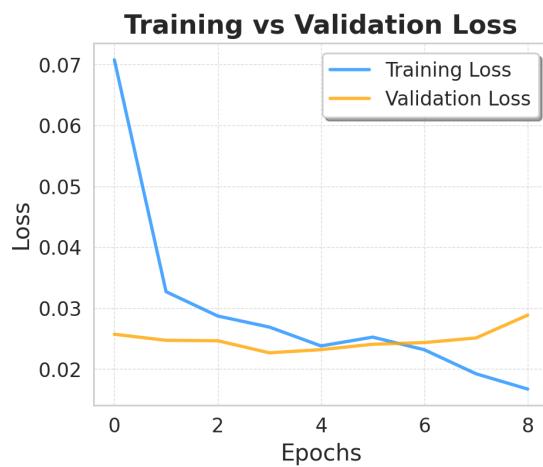
(b) Finetuned on Valence  
Pretrained AST on FMA Genre



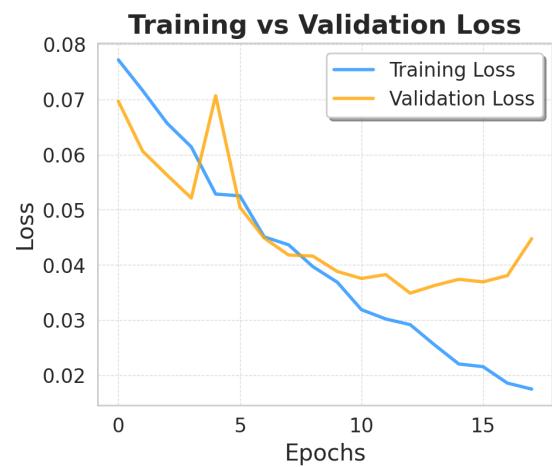
(c) Trained on Danceability  
AST with ImageNet pretraining



(d) Finetuned on Danceability  
Pretrained AST on FMA Genre



(e) Trained on Energy  
AST with ImageNet pretraining



(f) Finetuned on Energy  
Pretrained AST on FMA Genre

Θα περίμενε κανείς καλύτερη απόδοση από το δεξιό μοντέλο λόγω του επιπλέον pre-training στο FMA genre. Ωστόσο, δεδομένου ότι η εκπαίδευση στο FMA genre πραγματοποιήθηκε με μικρότερο όγκο δεδομένων, λιγότερες εποχές και μικρότερο batch size, η επίδραση του είναι περιορισμένη και μάλλον η φάση αυτή συγκαταλέγεται στο fine-tuning παρά στο pretraining. Στην ουσία, η κύρια συνεισφορά στην επιλογή των βαρών προέρχεται από το transfer learning μέσω των ImageNet pretrained βαρών, ενώ η περαιτέρω εκπαίδευση στο FMA genre, δεν έχει ζωτικής σημασίας ρόλο. Μάλλον αρνητική επίδραση παρατηρούμε ότι έχει στην συνολική απόδοση του μοντέλου, και αυτό επειδή αυξήσαμε την πολυπλοκότητα της εκπαίδευσης μας, ζητώντας από το μοντέλο να ερμηνεύσει τρία διαφορετικά datasets.

Βέβαια, σκοπός του ερωτήματος ήταν να δούμε την βελτίωση της απόδοσης που μας αποδίδει το transfer-learning, κάτι το οποίο εμείς είδαμε ήδη αξιοποιώντας τα ImageNet pretrained βάρη, και το διερευνήσαμε περισσότερο με το περαιτέρω 'pre-training' FMA Genre, όπου καταλήξαμε ότι καλό είναι να μην επιτελούμε πάνω από μια pre-training φάσεις, όταν τα tasks αυτών δεν ομοιάζουν και τα σύνολα δεδομένων τους έχουν διαφορετική πληθυσμότητα.

## Bήμα 10 (Multitask Learning)

Η δημοσίευση (One Model To Learn Them All) ουσιαστικά εισάγει μία νέα αρχιτεκτονική, ονόματι MultiModel, που πρόκειται για μια αρχιτεκτονική βαθιάς μάθησης που είναι σχεδιασμένη έτσι ώστε να μπορεί το μοντέλο να αποδίδει σε διάφορα tasks ανεξάρτητου domain όπως vision, nlp και speech recognition. Ενωποιεί τα διάφορα domain-specific architectural components, όπως convolutional layers (vision), attention mechanisms (nlp) μέσω sparsely-gated mixture-of-experts layers και χρησιμοποιεί modality nets για να διαχειριστεί δεδομένα διαφορετικών modalities.

Για να εκπαιδεύσουμε, αυτή την φορά, ένα μοντέλο ταυτόχρονα και στα 3 προβλήματα παλινδρόμησης αντί για ένα μοντέλο για κάθε πρόβλημα, επιλέγουμε ως συνάρτηση κόστους το άθροισμα των μέσων τετραγωνικών σφαλμάτων που σημείωσε το μοντέλο στα επιμέρους προβλήματα. Θα μπορούσε να γίνει και βεβαρημένο το άθροισμα, ωστόσο παρατηρήσαμε ότι το μοντέλο συγκλίνει ικανοποιητικά και σε αυτή την απλούστερη μορφή της συνάρτησης κόστους. Ως backbone χρησιμοποιήσαμε την tiny εκδοχή της transformer αρχιτεκτονικής AST και από πάνω προσθέσαμε 3 γραμμικά επίπεδα, ένα για κάθε πρόβλημα, σύμφωνα με τις αρχές του multitask learning. Ως βελτιστοποιητή χρησιμοποιήσαμε τον Adam με ρυθμό μάθησης  $\eta = 0.01$  καθώς και έναν scheduler συνημιτόνου για τον ρυθμό μάθησης. Τέλος αποφασίσαμε να δοκιμάσουμε τόσο ένα pretrained backbone όσο και ένα randomly initialized backbone.

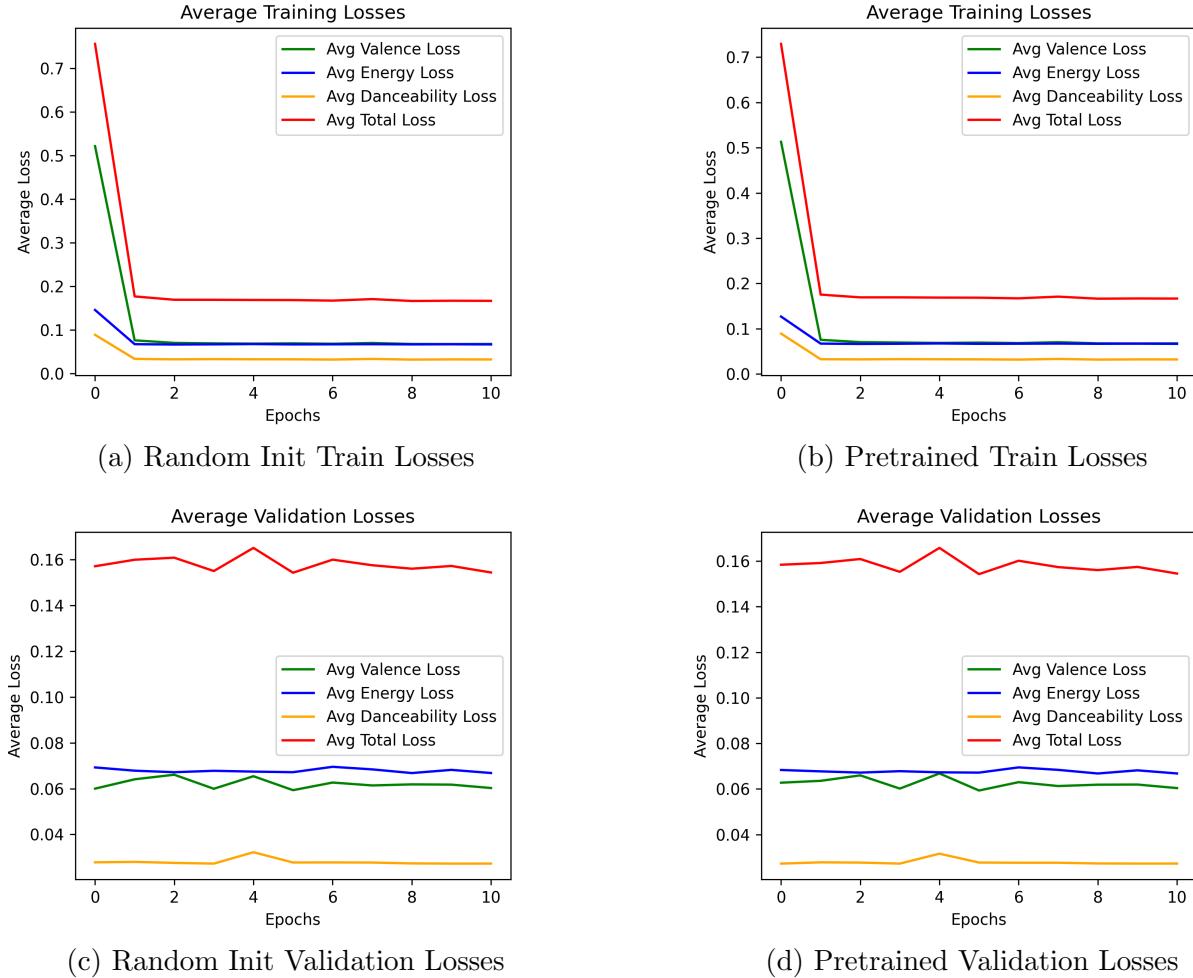


Figure 22: Training and Validation losses for randomly initialized and pretrained backbones.

Παρατηρώντας τα γραφήματα και τα δύο μοντέλα συγκλίνουν πολύ γρήγορα σε τιμές πολύ κοντά στο 0. Παρατηρούμε ότι το πρόβλημα πρόβλεψης του Danceability είναι το ευκολότερο για τα μοντέλα μιας και έχει υποδιπλάσιο validation loss σε σχέση με τα Energy και Valence, τα οποία φαίνονται να έχουν τον ίδιο βαθμό δυσκολίας χρίνοντας από το πόσο κοντά είναι τα αντίστοιχα training και αλιδατιον λοσσες κατά την εκπαίδευση και των δύο μοντέλων. Αναφορικά με το καλύτερο και τα δύο μοντέλα παρουσιάζουν σχεδόν πανομοιότυπη συμπεριφορά οπότε θα πρέπει να δούμε την αξιολόγησή τους βάση των μετρικών. Για να συγκρίνουμε τα ξεχωριστά AST μοντέλα για Valence, Energy και Danceability που εκπαιδεύσαμε στο βήμα 8, θα χρισμοποιήσουμε την μέση τιμή κάθε μετρικής που σημείωσαν τα 3 αυτά μοντέλα στο Βήμα 8.

*Ακολουθεί ο πίνακας με τα αποτελέσματα.*

Metric	AST (Step 8)	AST-Multi	AST-Multi (Pre)
Spearman	0.626	0.153	0.229
MSE	0.030	0.003	0.092
MAE	0.133	0.051	0.304
RMSE	0.169	0.051	0.304

Table 3: Comparaison of Single-Task and Multi-Task Trained Models

Παρατηρώντας τον πίνακα με τα αποτελέσματα βλέπουμε πως τα μοντέλα του βήματος 8 έχουν σημειώσει συντελεστή Spearman κατά πολύ μεγαλύτερο σε σχέση με τα Multi-Task μοντέλα. Παρόλα αυτά το τυχαία αρχικοποιημένο Multi-Task μοντέλο σημειώνει τις μικρότερες τιμές MSE, MAE, RMSE. Κατά την γνώμη μας, το καλύτερο μοντέλο είναι το τυχαία αρχικοποιημένο Multi-Task μοντέλο μιας και οι τιμές MSE, MAE, RMSE που σημείωσε είναι πολύ χαμηλότερες από τις αντίστοιχες μέσες τιμές των μοντέλων του βήματος 8. Παρόλο που έχει χαμηλότερο συντελεστή Pearson, οι μετρικές MSE, MAE, RMSE είναι στενότερα συνδεδεμένες με την απόχλιση των προβλέψεων σε σχέση με την πραγματική τιμή και τις θεωρούμε περισσότερο σημαντικές. Τέλος το γεγονός ότι θα έχουμε ένα μοντέλο ικανό και για τα 3 προβλήματα αντί για ένα μοντέλο για κάθε πρόβλημα, κάνει την επιλογή αυτή ακόμα περισσότερο ελκυστική.

## Βήμα 11: Οπτικοποίηση κρυφών αναπαραστάσεων

Εδώ μας ζητήθηκε να οπτικοποιήσουμε τις κρυφές αναπαραστάσεις που παράγουν οι καλύτερες εκδοχές του καινούργιου από τα τρία μοντέλα μας για κάθε τραγούδι του test set, το οποίο εξηγάγει από το fma\_genre\_spectrograms dataset.

Από την εκφώνηση, μας προτείνεται να δοκιμάσουμε και να πειραματιστούμε με την αποδοτικότητα διάφορων αλγορίθμων στην εφαρμογή dimensionality reduction για βέλτιστη οπτικοποίηση του latent representation χώρου του κάθε μοντέλου σε δύο διαστάσεις. Παρακάτω παραθέτω τις τεχνικές με τις οποίες πειραματίστηκα για να επιτύχω τον στόχο μου:

- **PCA (Principal Component Analysis):** Είναι μια γραμμική μέθοδος που βασίζεται στη διατήρηση της μέγιστης διασποράς στα δεδομένα. Υπολογίζει τα (principal components), μέσω των ιδιοτιμών (eigenvalues) και ιδιοδιανυσμάτων (eigenvectors) του πίνακα συνδιακύμανσης. Θεωρείται μια απλή, αποδοτική τεχνική, αλλά περιορίζεται από τη γραμμική της φύση και κάποιες φορές αποδεικνύεται ανεπαρκής για δεδομένα με περίπλοκες σχέσεις όπως τα φασματογραφήματα.
- **t-SNE (t-Distributed Stochastic Neighbor Embedding):** Είναι μια μη γραμμική μέθοδος που διατηρεί τοπικές σχέσεις μεταξύ των δεδομένων, μετατρέποντας αποστάσεις υψηλών διαστάσεων σε πιθανότητες. Στόχος του είναι να διατηρήσει την εγγύτητα των γειτονικών σημείων, δημιουργώντας σαφή (clusters). Είναι βέλτιστο για την κατανόηση της τοπικής δομής των δεδομένων, αλλά δεν διατηρεί τη γενική δομή και ίσως υπολογιστικά απαιτητικό.
- **UMAP (Uniform Manifold Approximation and Projection):** Βασίζεται στο (manifold theory), διατηρώντας τόσο τις τοπικές όσο και τις γενικές δομές των δεδομένων. Φαίνεται να έχει μικρή υπολογιστική πολυπλοκότητα και λειτουργεί καλά για μεγάλα και σύνθετα δεδομένα. Η ικανότητά του να κατανοεί τη συνολική γεωμετρία του χώρου το καθιστά ιδιαίτερα κατάλληλο για φασματογραφήματα, όπου τόσο οι τοπικές σχέσεις όσο και οι συνολικές κατηγορίες είναι κρίσιμες.
- **Spectral Embedding:** Χρησιμοποιεί θεωρία γραφημάτων για να αναλύσει τη συνδεσιμότητα των δεδομένων. Υπολογίζει τις ιδιοτιμές του γραφήματος γειτνίασης για να εντοπίσει μη γραμμικά μοτίβα. Αν και θεωρητικά ισχυρό, είναι πιο περιορισμένο σε datasets με υψηλή διαστασιμότητα και απαιτεί καλά επιλεγμένες παραμέτρους.
- **MDS (Multidimensional Scaling):** Διατηρεί τις αποστάσεις μεταξύ των σημείων, προβάλλοντας τα δεδομένα σε έναν χώρο χαμηλών διαστάσεων, προσφέροντας μια συνολική και σφαιρική εικόνα των σχέσεών τους. Ωστόσο, η αποδοτικότητά του μειώνεται σε datasets υψηλής διαστασιμότητας και δεν προσαρμόζεται καλά σε μη γραμμικές σχέσεις.
- **ISOMAP (Isometric Mapping):** Είναι μια επέκταση του MDS (Multidimensional Scaling), σχεδιασμένο να διατηρεί geodesic distances σε μη γραμμικούς χώρους. Είναι κατάλληλο για δεδομένα με ισχυρή μη γραμμικότητα, αλλά τείνει να χάνει τη συνολική δομή όταν το dataset είναι μεγάλο ή θορυβώδες.

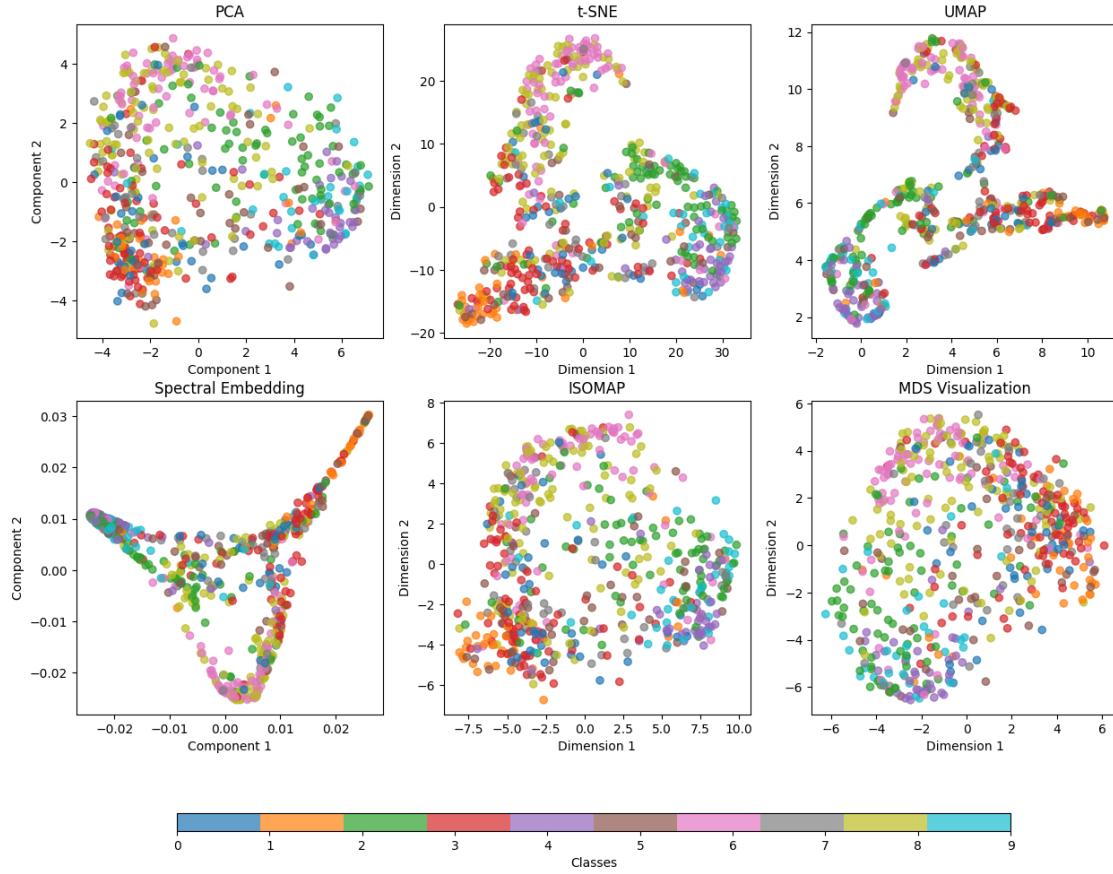
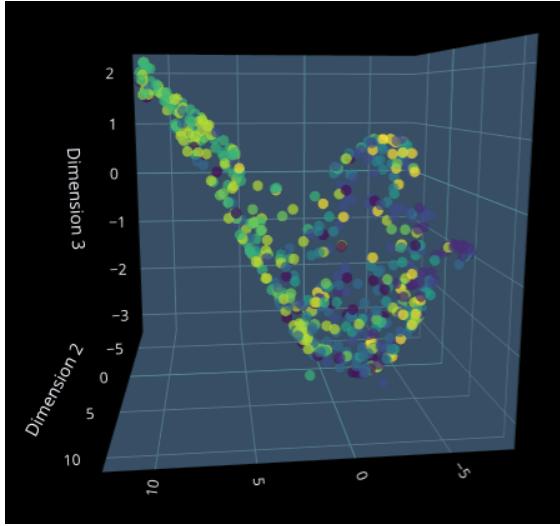


Figure 23: AST Latent Representations' Visualisation with different Dimensionality Reduction techniques

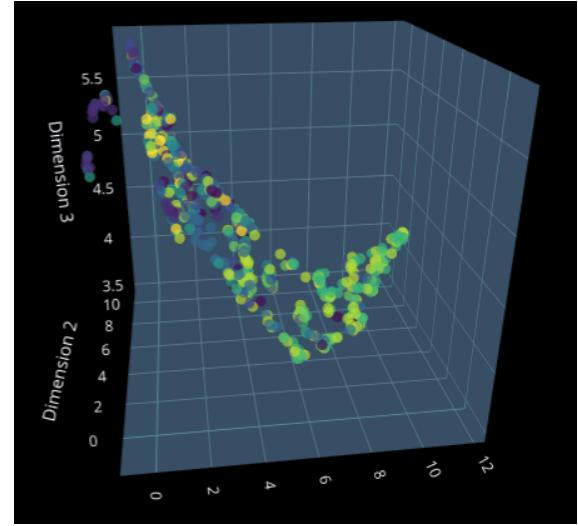
Παραπάνω βλέπουμε τα αποτελέσματα/οπτικοποιήσεις των δειγμάτων του *fma\_genre\_spectrograms* test dataset, αξιοποιώντας τις διαφορετικές τεχνικές που μελετήσαμε νωρίτερα. Με βάση τη θεωρία, οι τεχνικές PCA, MDS και Spectral Embedding αποτυγχάνουν να αναλύσουν επαρχώς και να προβάλλουν αποδοτικά σε χαμηλότερη διάσταση μη-μονοτονικά δεδομένα, όπως τα φασματογραφήματα, καθώς αδυνατούν να κατανοήσουν, σε ικανοποιητικό βαθμό, τις μη γραμμικές σχέσεις και την περίπλοκη τοπική γεωμετρία των δεδομένων. Επίσης, όσον αφορά το ISOMAP, παρόλο που καταφέρνει να διατηρήσει αποστάσεις σε μη γραμμικούς χώρους, παρουσιάζει δυσκολίες στη διατήρηση της συνολικής δομής όταν τα δεδομένα είναι μεγάλα ή υφρυβώδη, κάτι που πράγματι ισχύει στα φασματογραφήματα.

Αντιθέτως, οι τεχνικές t-SNE και UMAP, βάσει της θεωρίας, υπερέχουν στην ανάλυση μη-μονοτονικών δεδομένων, όπως τα φασματογραφήματα, καθώς διατηρούν αποτελεσματικά τις τοπικές δομές και τις σχέσεις γειτνίασης. Η t-SNE εστιάζει στη διατήρηση της τοπικής εγγύτητας μεταξύ σημείων, διασφαλίζοντας ότι οι τοπικές σχέσεις στον αρχικό χώρο αποτυπώνονται με ακρίβεια στον χαμηλής διάστασης χώρο. Από την άλλη, η UMAP, βασιζόμενη σε αρχές γραφημάτων και αλγεβρικής τοπολογίας, προσφέρει καλύτερη ισορροπία μεταξύ τοπικών και παγκόσμιων σχέσεων, γεγονός που την καθιστά πιο κατάλληλη για πολύπλοκα και πολυδιάστατα δεδομένα.

Με βάση τα προλεχθέντα, επιχειρώ να οπτικοποιήσω τα latent representations στον τρισδιάστατο χώρο, χρησιμοποιώντας τις πιο πολλά υποσχόμενες τεχνικές, δηλαδή την t-SNE και την UMAP.



(a) LSTM t-SNE 3D Visualisation



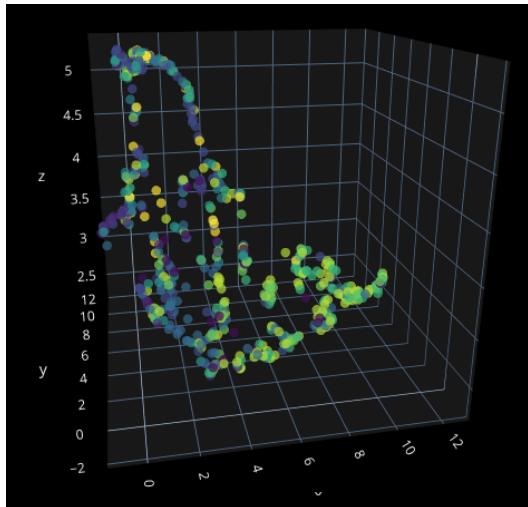
(b) LSTM UMAP 3D Visualisation

Figure 24: Comparison of t-SNE and UMAP 3D Visualisations of the best trained LSTM.

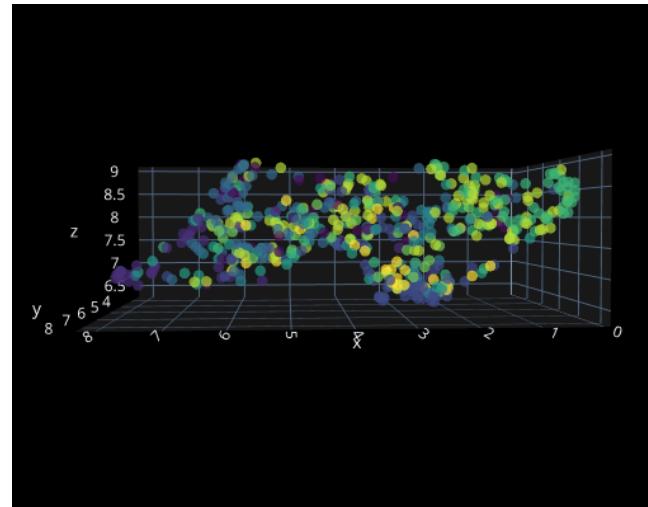
Στην παραπάνω απεικόνιση, παρατηρούμε δύο τρισδιάστατες οπτικοποιήσεις, του (αποθηκευμένου καλύτερου μοντέλου) LSTM, χρησιμοποιώντας τις δύο βέλτιστες τεχνικές μείωσης διαστάσεων. Η αριστερή οπτικοποίηση η οποία δημιουργήθηκε με χρήση του t-SNE είναι ένα παραβολοειδές ή αλιώς convex surface, bowl-shaped function. Είναι αστείο και ενδιαφέρον, γιατί αυτό το γράφημα συχνά συναντάται στην αναπαράσταση επιφανειών κόστους σε προβλήματα βελτιστοποίησης, όπως το gradient descent. Βέβαια, στην περίπτωση μας δεν έχει απολύτως κανένα αντίκρισμα ή ουσία αυτή η παρατήρηση, παρα μόνο θεωρητικό ενδιαφέρον. Στα δεξιά, στην οπτικοποίηση του UMAP, βλέπουμε μια πιο ομαλή τοπολογία. Παρότι παρέχει παρόμοιες πληροφορίες με το αριστερό γράφημα, η δομή της είναι διαφορετική. Αντί για παραβολοειδές, διωρίζονται δύο πιο επίπεδες επιφάνειες, μια μεγαλύτερη και μια σχετικά μικρή, οι οποίες συγκλίνουν σε ένα κοινό ολικό ελάχιστο, αναδεικνύοντας τη συνολική δομή των δεδομένων με τρόπο πιο συνεκτικό και ισορροπημένο.

Βάσει της θεωρίας, αλλά και των οπτικών μου συμπερασμάτων, καταλήγω στο ότι η τεχνική UMAP (Uniform Manifold Approximation and Projection) είναι η πιο αποδοτική και απεικονίζει με τον πλέον βέλτιστο τρόπο τα latent representations των δεδομένων μου συγκριτικά με τις υπόλοιπες, για τους λόγους που έχω επεκταθεί επαρκώς παραπάνω, με γραπτό λόγο αλλά και οπτικές αναπαραστάσεις. Θα ήθελα μόνο να παραθέσω έναν ακόμη λόγο που επιλέγω την τεχνική UMAP έναντι του t-SNE. Η t-SNE είναι πιο ευαίσθητη μεθόδος, κάτι που τονίζω με αρνητική χροιά, διότι εξαρτάται σε μεγάλο βαθμό από την παράμετρο *perplexity* (η οποία καθορίζει την τοπική δομή που θα αναδειχθεί) και από τις στοχαστικές διαδικασίες της οι οποίες μπορούν να δώσουν διαφορετικά αποτελέσματα κάθε φορά. Αυτό σημαίνει ότι χρειάζεται περισσότερη προσοχή στη ρύθμιση των παραμέτρων για να παράγει καλή οπτικοποίηση. Αντίθετα, η UMAP είναι πιο σταθερή και απαιτεί λιγότερη ρύθμιση.

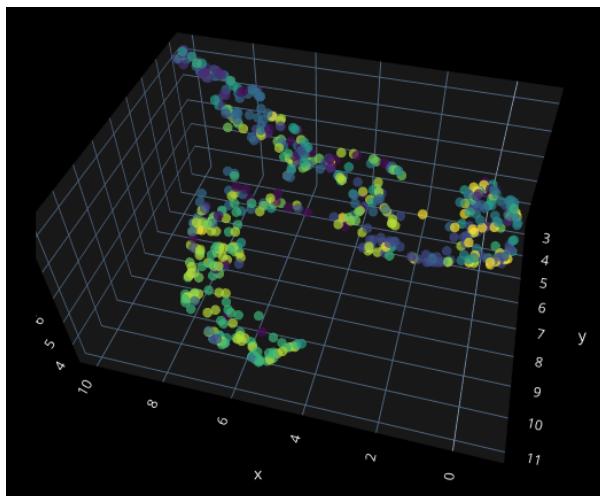
Έχοντας καταλήξει ότι η UMAP αντιπροσωπεύει την καλύτερη μου ελπίδα να οπτικοποιήσω αποδοτικά τα latent representations του test dataset, συνεχίζω πράττοντας ακριβώς αυτό για τα τρία πιο αποδοτικά αποθηκευμένα μοντέλα που έχω, δηλαδή τις καλύτερες εκδοχές των LSTM, CNN, AST στο Genre Classification Task. Επίσης, οπτικοποιώ και το χειρότερο μοντέλο στο συγκεκριμένο task το οποίο είναι το LSTM που εκπαιδεύθηκε επάνω σε chromagrams.



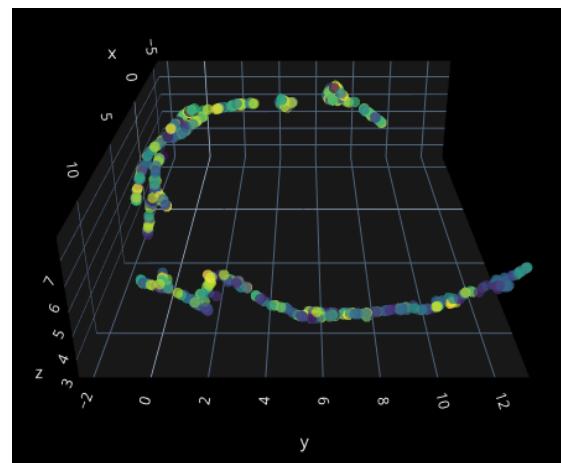
(a) LSTM (F1-Score: 0.34)



(b) CNN (F1-Score: 0.38)



(c) AST (F1-Score: 0.42)

(d) Worst LSTM (on Chromagrams)  
(F1-Score: 0.23)Figure 25: **3D UMAP** visualizations across the best models & the worst model.

Σε περίπτωση που τα στιγμιότυπα που έχουν επισυναφθεί δεν είναι αρκετά πειστικά στον αναγνώστη, υπάρχουν και οι 3d εκδόσεις αυτών αποθηκευμένες στον εκτελεσμένο κώδικα, πιο συγκεκριμένα στο step\_11.ipynb. Θα προχωρήσω στον σχολιασμό των παραπάνω οπτικοποίησεων και την συμπερασματολογία που προκύπτει από αυτές.

Στην UMAP οπτικοποίηση των latents representations του **LSTM**, την οποία αναλύσαμε και νωρίτερα, παρατηρούμε δύο σχετικά επίπεδες επιφάνειες, μια μεγαλύτερη και μια μικρή, οι οποίες συγκλίνουν σε ένα κοινό ολικό ελάχιστο. Να προσθέσω ότι αυτό το μοντέλο είναι η καλύτερη εκδοχή του LSTM μοντέλου και παρουσιάζει micro-average precision, micro-average recall & micro-average F1-score = 0.23 .

Στην συνέχεια, δεξιά, έχουμε την UMAP οπτικοποίηση των latents representations του **CNN** μοντέλου. Εδώ παρατηρούμε μια μεγαλύτερη διασπορά των σημείων, συγχριτικά με κάμψη άλλο μοντέλο, κάτι που πιθανόν μας πληροφορεί ότι η αποδοτικότητα του CNN περιορίζεται σε τοπικούς συσχετισμούς, δηλαδή τοπικές χωρικές ιδιότητες των δεδομένων αλλά χάνει πιο σύνθετες, χρονικές/ρυθμικές πληροφορίες. Για αυτό δηλαδή, παρατηρώ έναν σχετικά καλό διαχωρισμό των κλάσεων χωρίς overlaps αλλά η συνολική εικόνα είναι άναρχη/άτακτη επειδή το CNN αδύνατεί να κατανοήσει πιο περίπλοκες συσχετίσεις, αφού αντιμετωπίζει τα δεδομένα περισσότερο ως εικόνες, παρά ως διαδοχική πληροφορία. Η καλύτερη εκδοχή του CNN μοντέλου παρουσιάζει micro-average precision, micro-average recall & micro-average F1-score = 0.38 .

Ακολουθεί η UMAP οπτικοποίηση των latents representations του **AST** μοντέλου. Εδώ παρατηρούμε μια δομή τύπου μανιφόλδ, όπου τα δεδομένα είναι οργανωμένα σε διαδρομές ή συστάδες που ακολουθούν έναν ελικοειδή σχηματισμό. Η συγκεκριμένη δομή του scatter plot, μας υποδεικνύει ότι τα δεδομένα εμφανίζουν σημαντικές μη γραμμικές σχέσεις και ότι οι υποκείμενες σχέσεις διατηρούνται εν πολλοίς, αφότου μειώθηκαν οι διαστάσεις. Η καλύτερη εκδοχή του AST μοντέλου παρουσιάζει micro-average precision, micro-average recall & micro-average F1-score = 0.42 .

Τέλος, βλέπουμε την UMAP οπτικοποίηση των latents representations του **χειριστού LSTM** μοντέλου, το οποίο έχει εκπαιδευθεί στο Chromagrams Dataset για το Genre Classification Task. Παρατηρούμε μια τοξοειδή γεωμετρία, όπου τα δεδομένα να σχηματίζουν μια συνεχή καμπύλη στον χώρο. Η συγκεκριμένη δομή πιθανώς προκύπτει από την αρχιτεκτονική του μοντέλου, που είναι αποδοτικό στη διαδοχική ανάλυση αλλά λιγότερο βέλτιστο στη μοντελοποίηση πολυδιάστατων χαρακτηριστικών, όπως αυτά που αποτυπώνονται στα χρωμογραφήματα. Η κατανομή δεν είναι ομοιόμορφη. Βλέποντας το γράφημα αυτό όπου οι κλάσεις δεν φαίνεται να διαχωρίζονται, μπορούμε να μαντέψουμε ότι η απόδοση του θα είναι αρκετά χειρότερη από τα υπόλοιπα μοντέλα. Η χειρότερη εκδοχή του LSTM μοντέλου παρουσιάζει micro-average precision, micro-average recall & micro-average F1-score = 0.23 Η κακή απόδοση αυτού το μοντέλου, οφείλεται στο ότι τα chromagrams περιέχουν πληροφορία μόνο για τη τονικότητα, δηλαδή τις νότες, και όχι για την χροιά, τις συχνότητες & τις χρονικές εξαρτήσεις που είναι χρίσιμες για την ταξινόμηση μουσικών ειδών. Αυτό έχει ως αποτέλεσμα να χάνεται η πλούσια φασματική και ρυθμική πληροφορία που περιέχουν τα spectrograms, καθιστώντας τα chromagrams ακατάλληλα για genre classification.

Συμπερασματικά, το AST on Mel-Spectrograms υπερέχει ως το πιο αποδοτικό μοντέλο, χάρη στην ικανότητά του να αποτυπώνει τις μη γραμμικές και πολυδιάστατες σχέσεις των φασματογραφημάτων. Το CNN δείχνει αξιοπρεπή απόδοση, εστιάζοντας χυρίως σε τοπικά χαρακτηριστικά. Το LSTM είναι λιγότερο αποδοτικό, λόγω του περιορισμού του σε μακροπρόθεσμες εξαρτήσεις. Ενώ η χειρότερη εκδοχή του LSTM μοντέλου, που εκπαιδεύθηκε πάνω σε chromagrams, υποδεικνύει ότι τα τελευταία δεν αποτελούν κατάλληλη αναπαράσταση για το συγκεκριμένο task. Τελικά, το AST αναδεικνύεται ως η βέλτιστη επιλογή για Genre Classification και το UMAP ως ο βέλτιστος τρόπος μείωσης διαστασιμότητας των δεδομένων μας.