ΒΗΜΑ 1

Ερώτημα α

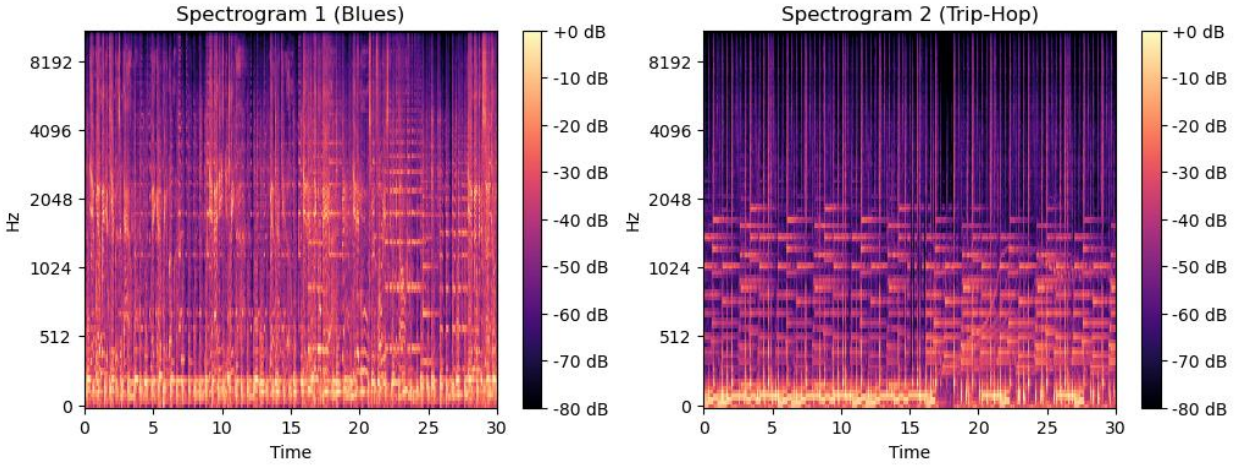
Με χρήση της numpy διαβάζουμε από το αρχείο fma_genre_spectrograms/train_labels.txt δύο γραμμές με διαφορετικές επισημειώσεις (labels) και τις αποθηκεύουμε σε μία λίστα. Συγκεκριμένα, πρόκειται για τα αρχεία:
['1042.fused.full.npy', '69588.fused.full.npy'], που ανήκουν στο είδος Blues και Trip-Hop αντίστοιχα.

Ερώτημα β

Για τα δύο .npy αρχεία που έχουμε επιλέξει, παίρνουμε τα φασματογραφήματά τους, τα οποία αποτελούνται από τις 128 πρώτες γραμμές του αρχείου.

Ερώτημα γ

Με χρήση της librosa.display.specshow απεικονίζουμε τα φασματογραφήματα των δύο αρχείων (Εικόνα 1)



Εικόνα 1: Φασματογραφήματα για Blues και Trip-Hop

Το φασματογραφήματα απεικονίζουν το φάσμα του συχνοτήτων του ήχου στα δοσμένα σήματα.
Για το πρώτο, παρατηρούμε συνεχή κατανομής ενέργειας σε όλες τις συχνότητες, που είναι ενδεικτικό της ύπαρξης σταθερού ρυθμού. Περισσότερη ενέργεια υπάρχει στις χαμηλές, που είναι σύνηθες για τη μουσική Blues, λόγω της ισχυρής παρουσίας του μπάσου.
Για το δεύτερο, παρατηρούμε ότι υπάρχουν συγκεκριμένα διακριτά σημεία υψηλής ενέργειας, ακολουθούμενο από σημεία χαμηλής. Αυτή η εναλλαγή δείχνει μεταβολές στο ρυθμό, κάτι που είναι συνεπές με τη φύση της ηλεκτρονικής μουσικής.

ΒΗΜΑ 2

Ερώτημα α

Τα φασματογραφήματα που δείξαμε παραπάνω έχουν χρονική διάρκεια 30 sec. Πήραμε τις default τιμές της συνάρτησης specshow της librosa:

- sr = 22050 Hz (ρυθμός δειγματοληψίας)
- hop length = 512 (αριθμός δειγμάτων / παράθυρο)

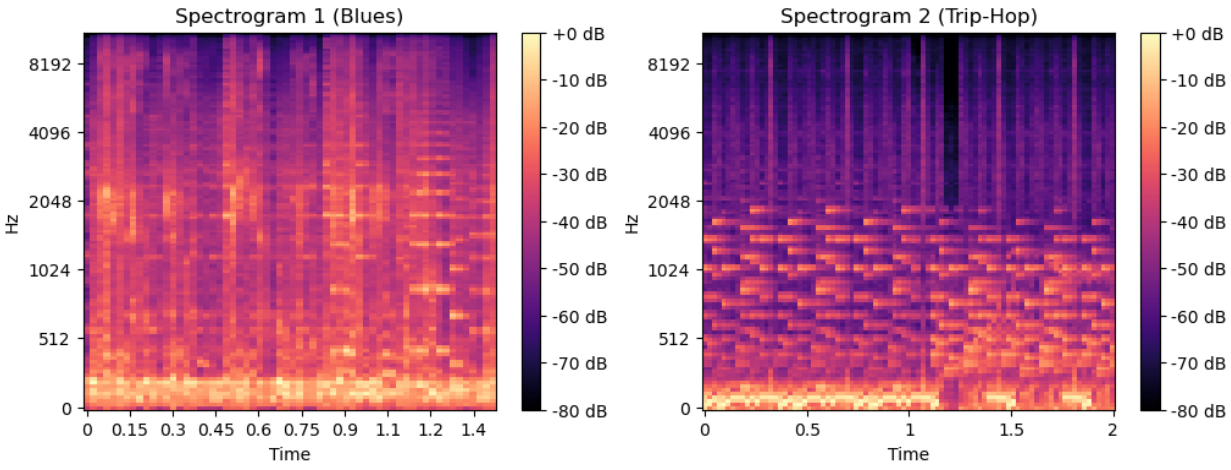
Για το πλήθος των timesteps ισχύει:

Blues spectrogram timesteps = 1293 Trip-Hop spectrogram timesteps = 1293

Η εκπαίδευση ενός LSTM δικτύου με 1293 timesteps ανά sample είναι δύσκολη λόγω της δυσκολίας που έχει το LSTM να μαθαίνει εξαρτήσεις στα δεδομένα σε μεγάλες ακολουθίες. Αυτό οφείλεται στο πρόβλημα των exploding/vanishing gradients, καθώς και στην υψηλή διαστατικότητα του προβλήματος, καθώς κάθε timestep είναι ένα στιγμιότυπο του φάσματος συχνοτήτων. Επιπλέον, κάθε βήμα του φασματογραφήματος έχει ισχυρή συσχέτιση με το προηγούμενο και το επόμενό του, κάτι που εισάγει περιττή πληροφορία στο LSTM, και θα συμβάλλει σε μεγαλύτερο χρόνο εκπαίδευσης.

Ερώτημα β

Χρησιμοποιώντας τα ίδια δύο .npy αρχεία που αντιστοιχούν τώρα σε φασματογραφήματα συγχρονισμένα πάνω στο ρυθμό, και πάλι με χρήση της specshow, απεικονίζουμε εκ νέου τα δύο φασματογραφήματα (Εικόνα 2).



Εικόνα 2: Φασματογραφήματα για τα ίδια αρχεία, συγχρονισμένα στο ρυθμό

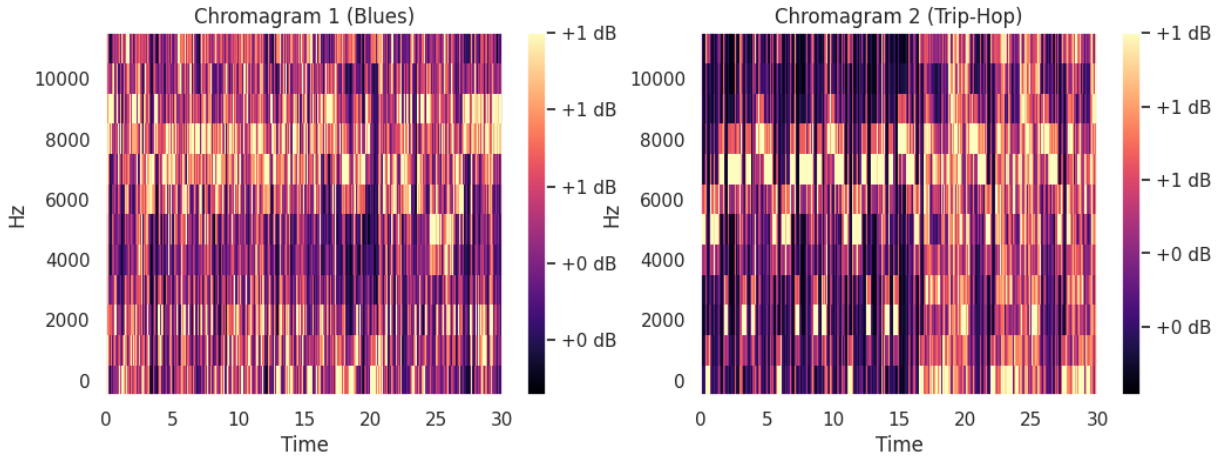
Παρατηρούμε ότι πλέον στον άξονα x, αντί για ίσα χρονικά διαστήματα, υπάρχουν τα χρονικά σημεία όπου ακούγεται το beat της μουσικής. Υπάρχουν αρκετά λιγότερες στήλες, καθώς κάθε στήλη ισοδυναμεί με ένα beat. Επομένως, αν κάποια μοτίβα beat είναι χαρακτηριστικά σε κάποια μουσικά είδη, διευκολύνεται περισσότερο η ταξινόμηση δειγμάτων μουσικής σε κατηγορίες. Επιπλέον τα timesteps είναι πολύ λιγότερα:

```
Blues beat synced spectrogram timesteps = 62
Trip-Hop beat synced spectrogram timesteps = 87
```

Βήμα 3

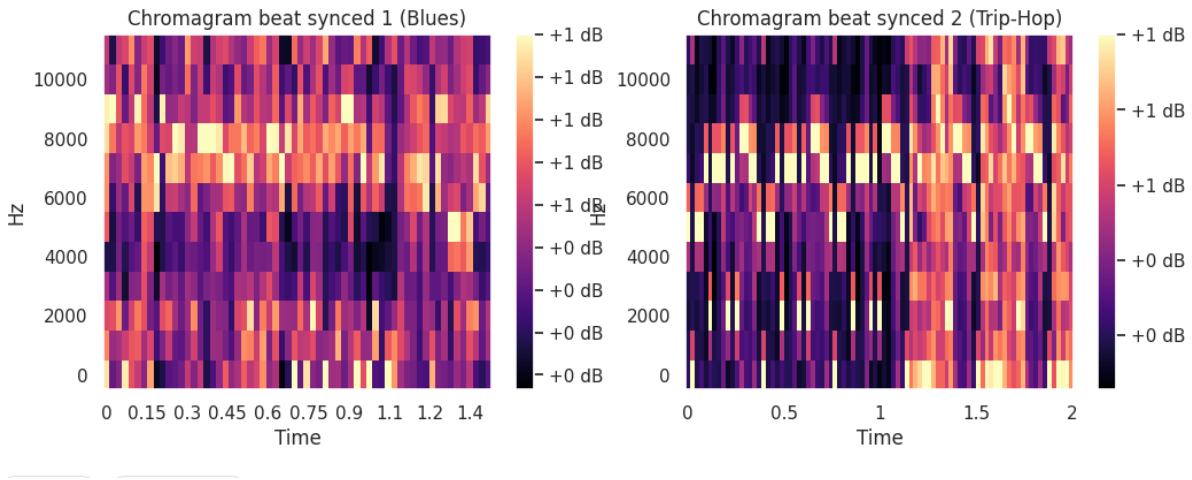
Επαναλάβαμε την ίδια διαδικασία με παραπάνω λαμβάνοντας τα χρωμογραφήματα (που από τα αρχεία μας είναι τα χαρακτηριστικά από το 129 εώς το 140 (12 σύνολο). Ο αριθμός των timesteps παραμένει ο ίδιος. Απεικονίζουμε τα χρωμογραφήματα για τα δύο ίδια αρχεία (Εικόνα 3) όπως και για τα αρχεία που είναι συγχρονισμένα στο ρυθμό (Εικόνα 4).

```
Blues spectrogram timesteps = 1293
Trip-Hop spectrogram timesteps = 1293
```



Εικόνα 3: Χρωμογραφήματα για τα αρχεία των δύο ειδών μουσικής

```
Blues chromagram beat synced timesteps = 62
Trip-Hop chromagram beat synced timesteps = 87
```



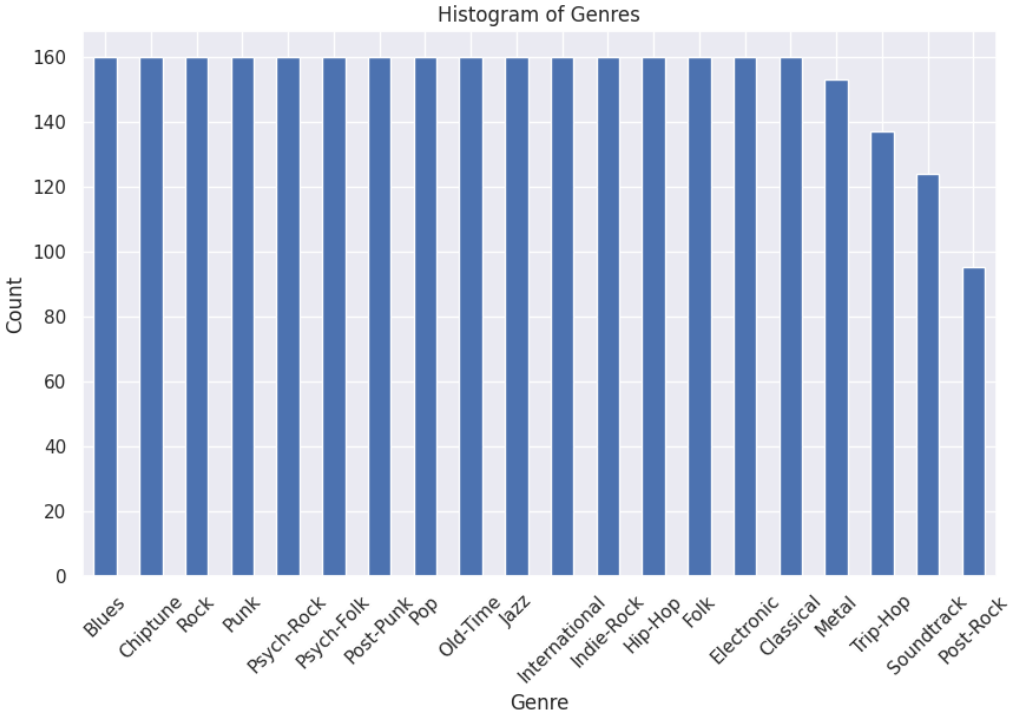
Εικόνα 4: Χρωμογραφήματα για τα συγχρονισμένα στο ρυθμό αρχεία

Στα δύο πρώτα χρωμογραφήματα βλέπουμε τη κατανομή της έντασης του pitch κατά τα 30 δευτερόλεπτα. Παρατηρούμε πως στο κομμάτι Blues έχουμε σταδιακές αλλαγές της έντασης που είναι χαρακτηριστικό του είδους, ενώ στο Trip-Hop έχουμε απότομες αλλαγές (από μαύρο σε κίτρινο) οι οποίες είναι χαρακτηριστικές της Trip-Hop μουσικής.

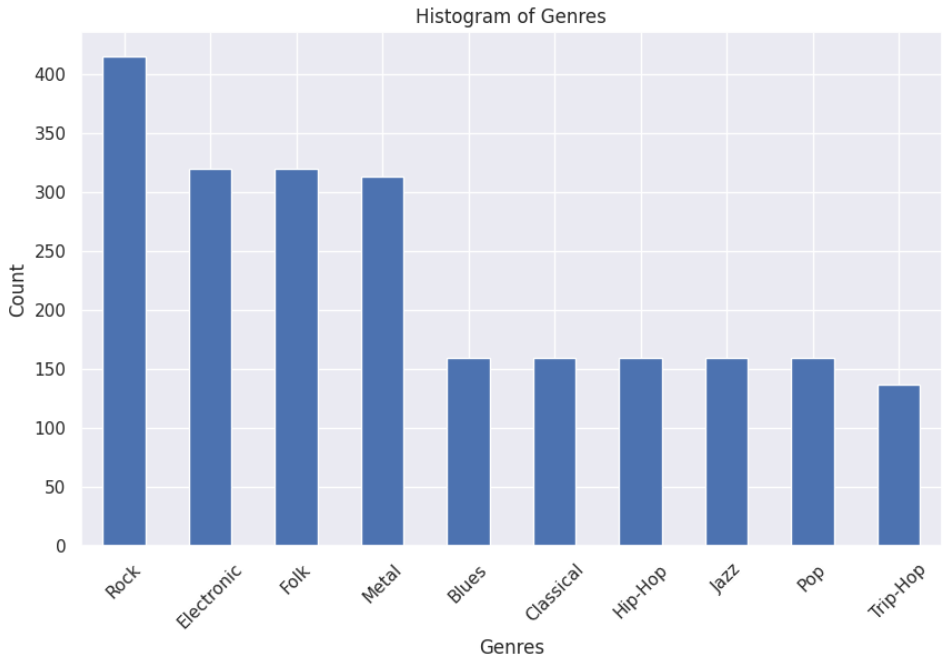
Στα beat-synced χρωμογραφήματα ομοίως με τα φασματογραφήματα έχουν ελαττωθεί οι στήλες και υπάρχουν μόνο τα σημεία αλλαγής του ρυθμού.

Βήμα 4

Σχεδιάσαμε τα ιστογράμματα πριν (Εικόνα 5) και μετά (Εικόνα 6) τη συγχώνευση των ειδών (και την αφαίρεση όσων είχαν λίγα δείγματα).



Εικόνα 5: Ιστόγραμμα των ειδών μουσικής του dataset



Εικόνα 6: Ιστόγραμμα των ομαδοποιημένων μουσικών ειδών

Στο κώδικα που παρέχεται εκτελείται η ανάγνωση και προεπεξεγασία των δεδομένων. Συγκεκριμένα:

- Η torch_train_val_split λαμβάνει το dataset και κατασκευάζει τα Train και Validation Dataloaders που αποτελούνται από Batches τα οποία περιέχουν στοιχεία της μορφής: (data, labels, original_lengths) όπου το data είναι ένας 3D τένσορας διαστάσεων Batch Size x max(timesteps) x features, τα labels είναι η κλάση στην οποία ανήκουν τα στοιχεία του batch και το original length είναι το πλήθος των timesteps πριν γίνει το padding.
- Η read_spectrogram τροποποιήθηκε ώστε να επιστρέφει είτε τα φασματογραφήματα σε κλίμακα mel, δηλαδή τις 128 πρώτες γραμμές, είτε τα χρωμογραφήματα, δηλαδή τις υπόλοιπες 12, είτε και τα δύο (140)
- Η κλάση SpectrogramDataset διαβάζει τα δεδομένα από το φάκελο (train ή test) και δημιουργεί το dataset. Πραγματοποιήσαμε μια μικρή διόρθωση στη μέθοδο get_files_labels, καθώς ήταν διαφορετική η δομή των ονομάτων των test αρχείων σε σχέση με αυτών των train.

Βήματα 5,6

Ερώτημα β

Για να βεβαιωθούμε πως το νευρωνικό δίκτυο λειτουργεί θέσαμε το overfit_batch=True έτσι ώστε να εκπαιδευτεί για πολύ μεγάλο αριθμό επαναλήψεων και για λίγα batches. Πράγματι το σφάλμα έτεινε στο 0, πράμα που σημαίνει πως είχαμε overfitting και το μοντέλο δουλεύει ορθά.

Εκπαιδεύσαμε το μοντέλο με τις εξής υπερπαραμέτρους:

epochs	100
optimizer	Adam
learning rate	10^{-5}
weigh decay	10^{-6}
dropout rate	0.4
early stopping patience	3

Για να αξιολογήσουμε τα μοντέλα μας θα χρησιμοποιήσουμε τις εξής μετρικές:

- Accuracy: Η αναλογία σωστών προβλέψεων προς το συνολικό αριθμό των προβλέψεων
    - Υψηλό accuracy σημαίνει πως το μοντέλο ταξινομεί σωστά τα περισσότερα labels, ωστόσο αν μια κλάση έχει πολύ περισσότερα δείγματα από τις υπόλοιπες μπορεί να μας οδηγήσει σε λάθος συμπεράσματα.
- Precision: Η αναλογία των σωστά ταξινομημένων σε μία κλάση προς όλα όσα ταξινομήθηκαν σε αυτή τη κλάση.
    - Υψηλό precision για μία κλάση σημαίνει πως όταν ένα δείγμα ταξινομείται στη κλάση είναι πολύ πιθανό όντως να ανήκει σε αυτή τη κλάση.
- Recall: Η αναλογία των σωστά ταξινομημένων ως προς όλα όσα ανήκουν στη κλάση (είτε ταξινομήθηκαν σωστά είτε λάθος)
    - Υψηλό recall για μία κλάση σημαίνει ότι το μοντέλο είναι ικανό να ταξινομεί σωστά όσα ανήκουν στη συγκεκριμένη κλάση.
- F1 Score: Ο αρμονικός μέσος όρος των precision και recall.
    - Υψηλό F1 Score για μία κλάση σημαίνει πως για αυτή τη κλάση υπάρχει μια καλή ισορροπία μεταξύ των δύο μετρικών.
    - Είναι πολύ σημαντική μετρική στη περίπτωση μας καθώς έχουμε διαφορετικό πλήθος δειγμάτων για κάθε κλάση.
- Micro Averaged Precision: Η αναλογία των σωστά ταξινομημένων για όλες τις κλάσεις προς το άθροισμα όσων ταξινομήθηκαν στις κλάσεις.
    - Όταν είναι υψηλό τότε, όταν το μοντέλο ταξινομεί σε μία κλάση, συνήθως το κάνει σωστά
- Micro Averaged Recall: Η αναλογία των σωστά ταξινομημένων για όλες τις κλάσεις προς τα σωστά ταξινομημένα + τα λάθος ταξινομημένα σε άλλες κλάσεις
    - Όταν είναι υψηλό, τότε το μοντέλο σπάνια δεν εντοπίζει
- Micro Averaged F1: Ο αρμονικός μέσος όρος των Micro Averaged Precision και Micro Averaged Recall
    - Όταν είναι υψηλό, τότε υπάρχει καλή ισορροπία ανάμεσα στις δύο αυτές μετρικές

ΣΗΜΕΙΩΣΗ: Στη περίπτωση μας, τα micro averaged μεγέθη ταυτίζονται με το accuracy διότι κάθε δείγμα ανήκει σε μία μόνο κλάση

- Macro Averaged: Οι μέσοι όροι των metrics για όλες τις κλάσεις. Δίνουν μια γενική εικόνα για τις μετρικές του μοντέλου

Μεγάλη απόκλιση ανάμεσα στο accuracy και στο F1 Score μπορεί να σημαίνει τα εξής:

- Άνιση κατανομή των δειγμάτων στις κλάσεις
- Bias του μοντέλου ως προς κάποιες κλάσεις

Μεγάλη απόκλιση ανάμεσα σε Micro και Macro Averaged F1 Scores μπορεί να σημαίνει τα εξής:

- Άνιση κατανομή των δειγμάτων στις κλάσεις
- Αδυναμία του μοντέλου να ανιχνεύσει συγκεκριμένες κλάσεις

Η επιλογή ανάμεσα σε βελτιστοποίηση του precision ή του recall εξαρτάται από τη φύση του προβλήματος. Σε προβλήματα που το κόστος των λανθασμένα ταξινομημένων σε μια κατηγορία είναι πολύ μεγαλύτερο σε σχέση με το κόστος όσων δεν ταξινομήθηκαν στη κλάση ενώ ανήκουν σε αυτή, απαιτείται υψηλό precision. Τέτοια παραδείγματα είναι τα spam filters και το fraud detection.

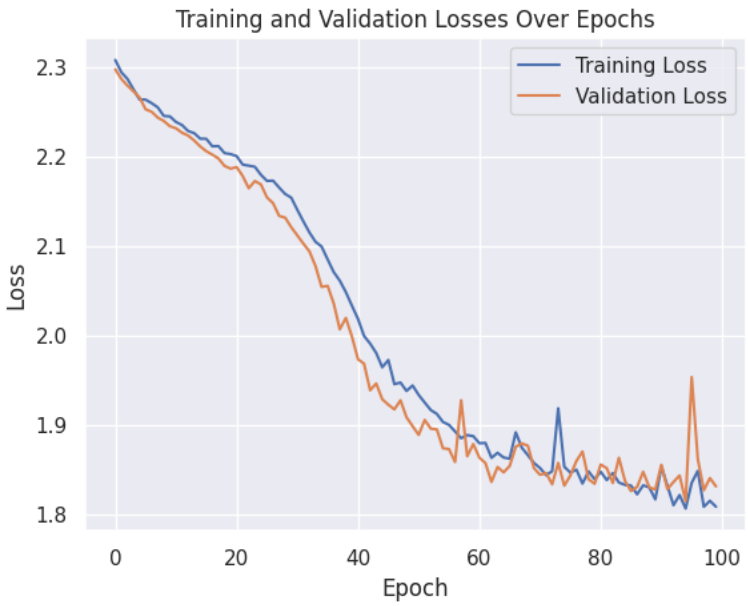
Επομένως το accuracy δεν αρκεί για να επιλέξω ένα μοντέλο επειδή υπάρχει η περίπτωση κάποιες κλάσεις να έχουν λιγότερα δείγματα από άλλες.

Το F1 επίσης δεν είναι αρκετό, καθώς όταν είναι χαμηλό γνωρίζουμε πως ένα εκ των precision και recall (ή και τα δύο) είναι χαμηλό, αλλά όχι ποιο.

Τα Macro Averaged Metrics είναι πολύ χρήσιμα επειδή δίνουν την ίδια βαρύτητα σε κάθε κλάση, ανεξαρτήτως του πλήθους δειγμάτων της.

Ερώτημα γ

Στην αρχή το μοντέλο εκπαιδεύτηκε πάνω στα φασματογραφήματα.



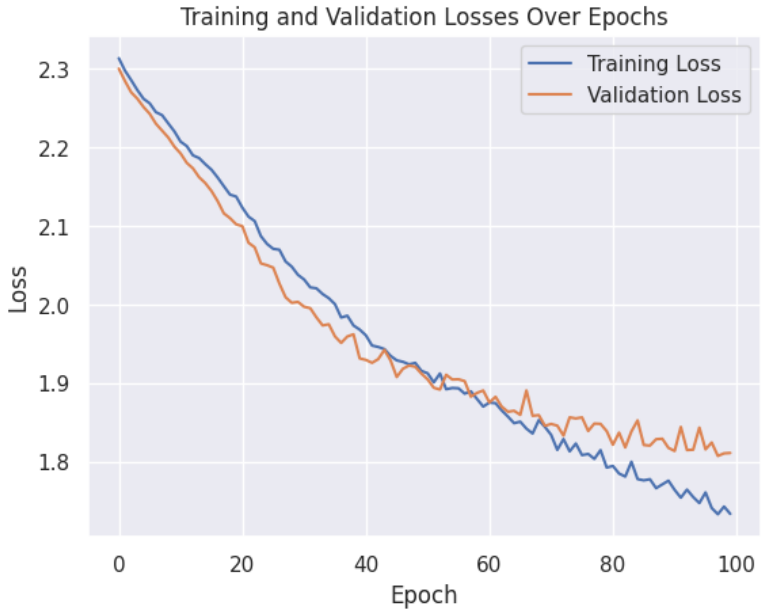
Εικόνα 7: Σφάλματα Εκπαίδευσης και Επαλήθευσης για εκπαίδευση με φασματογραφήματα

	precision	recall	f1-score	support
0	0.00	0.00	0.00	40
1	0.50	0.57	0.53	40
2	0.32	0.70	0.44	80
3	0.31	0.50	0.38	80
4	1.00	0.03	0.05	40
5	0.00	0.00	0.00	40
6	0.38	0.69	0.49	78
7	0.00	0.00	0.00	40
8	0.36	0.29	0.32	103
9	0.00	0.00	0.00	34
accuracy			0.35	575
macro avg	0.29	0.28	0.22	575
weighted avg	0.31	0.35	0.28	575

Πίνακας 1: Οι μετρικές κατηγοριοποίησης του μοντέλου που εκπαιδεύτηκε με φασματογραφήματα

Ερώτημα δ

Αυτή τη φορά δώσαμε ως είσοδο τα beat-synced φασματογραφήματα.



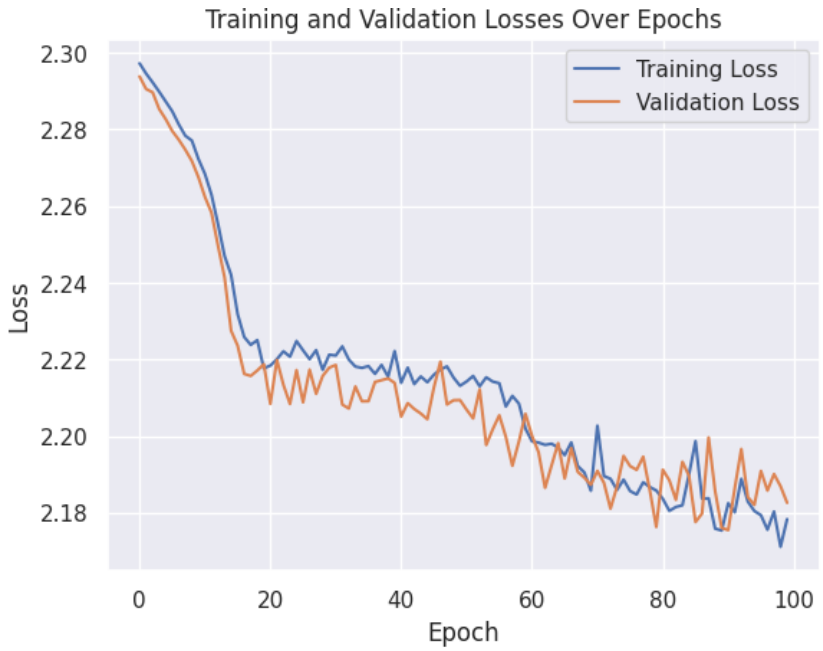
Εικόνα 8: Σφάλματα Εκπαίδευσης και Επαλήθευσης για εκπαίδευση με beat-synced φασματογραφήματα

	precision	recall	f1-score	support
0	0.00	0.00	0.00	40
1	0.39	0.55	0.45	40
2	0.31	0.76	0.44	80
3	0.36	0.56	0.44	80
4	0.29	0.05	0.09	40
5	0.00	0.00	0.00	40
6	0.44	0.59	0.51	78
7	0.00	0.00	0.00	40
8	0.34	0.28	0.31	103
9	0.00	0.00	0.00	34
accuracy			0.36	575
macro avg	0.21	0.28	0.22	575
weighted avg	0.26	0.36	0.28	575

Πίνακας 2: Οι μετρικές κατηγοριοποίησης του μοντέλου που εκπαιδεύτηκε με beat-synced φασματογραφήματα

Ερώτημα ε

Σε αυτό το ερώτημα χρησιμοποιήθηκαν τα χρωμογραφήματα.



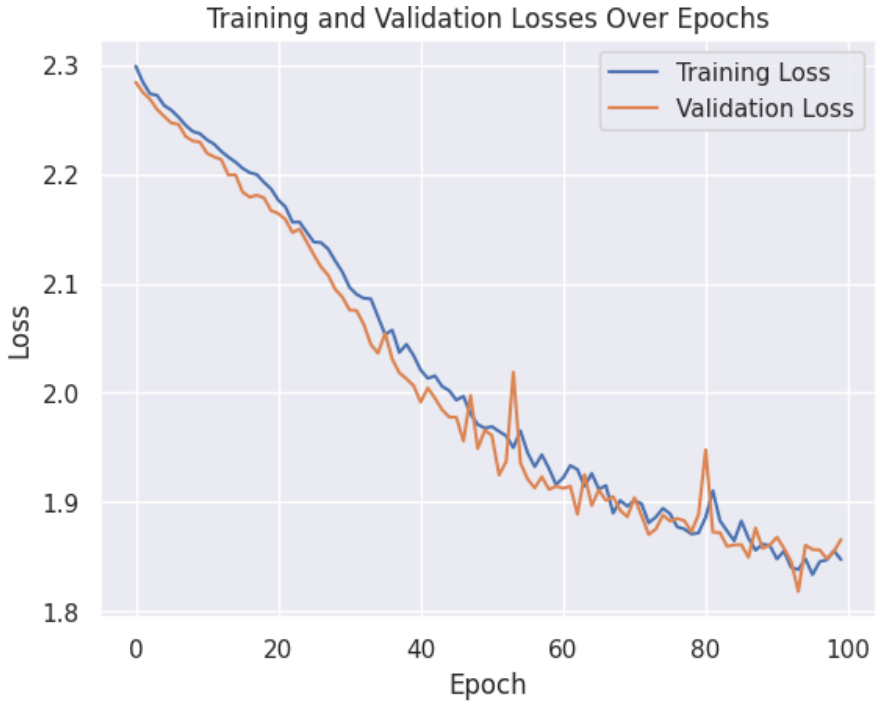
Εικόνα 9: Σφάλματα Εκπαίδευσης και Επαλήθευσης για εκπαίδευση με χρωμογραφήματα

	precision	recall	f1-score	support
0	0.00	0.00	0.00	40
1	0.00	0.00	0.00	40
2	0.00	0.00	0.00	80
3	0.21	0.47	0.29	80
4	0.00	0.00	0.00	40
5	0.00	0.00	0.00	40
6	0.21	0.42	0.28	78
7	0.00	0.00	0.00	40
8	0.15	0.35	0.21	103
9	0.00	0.00	0.00	34
accuracy			0.19	575
macro avg	0.06	0.12	0.08	575
weighted avg	0.08	0.19	0.12	575

Πίνακας 3: Οι μετρικές κατηγοριοποίησης του μοντέλου που εκπαιδεύτηκε με χρωμογραφήματα

Ερώτημα ζ

Τέλος δίνουμε και τα φασματογραφήματα και τα χρωμογραφήματα ως είσοδο.



Εικόνα 9: Σφάλματα Εκπαίδευσης και Επαλήθευσης για εκπαίδευση με φασματογραφήματα και χρωμογραφήματα

	precision	recall	f1-score	support
0	0.00	0.00	0.00	40
1	0.42	0.42	0.42	40
2	0.29	0.72	0.42	80
3	0.37	0.42	0.39	80
4	0.33	0.03	0.05	40
5	0.00	0.00	0.00	40
6	0.41	0.44	0.43	78
7	0.00	0.00	0.00	40
8	0.30	0.46	0.36	103
9	0.00	0.00	0.00	34
accuracy			0.33	575
macro avg	0.21	0.25	0.21	575
weighted avg	0.25	0.33	0.27	575

Πίνακας 10: Οι μετρικές κατηγοριοποίησης του μοντέλου που εκπαιδεύτηκε με φασματογραφήματα και χρωμογραφήματα

Σχολιασμός:

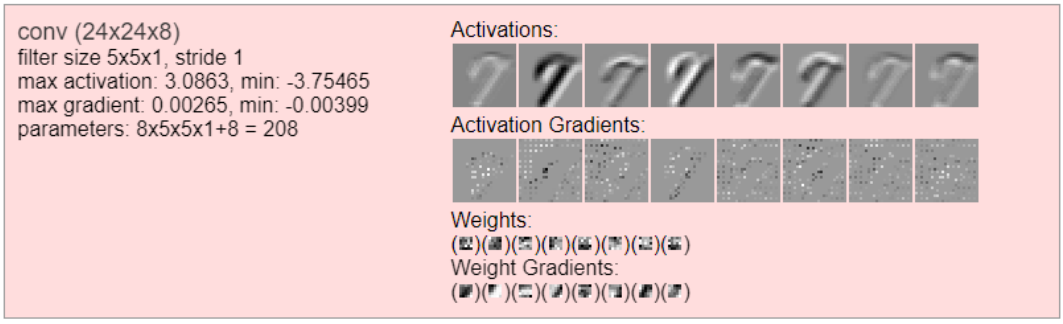
- Παρατηρούμε πως σε όλες τις περιπτώσεις οι κλάσεις 0, 5, 7 και 9 έχουν 0 σωστά ταξινομημένα δείγματα σε αυτές.
- Για τη περίπτωση των beat-synced spectrograms το μοντέλο εκπαιδεύτηκε πολύ πιο γρήγορα λόγω του σημαντικά μειωμένου αριθμού timesteps.
- Υπάρχει σημαντική διαφορά ανάμεσα στο macro averaged F1 Score και στο accuracy πράμα που επιβεβαιώνει πως κάποιες κλάσεις εκπροσωπούνται λιγότερο από τα δείγματα.
- Τα χρωμογραφήματα δεν αποτελούν καλό είδος δεδομένων εκπαίδευσης, καθώς τα test δείγματα ταξινομήθηκαν μόλις σε 3 κλάσεις, το σφάλμα είναι μεγαλύτερο και είχε υψηλές διακυμάνσεις κατά την εκπαίδευση.

Βήμα 7

Το input layer λαμβάνει ως είσοδο μία εικόνα (24 x 24 x 1). Οι αρχικές διαστάσεις των εικόνων είναι 28 x 28 αλλά επιλέγεται ένα τυχαίο παράθυρο 24 x 24 έτσι ώστε το μοντέλο να γενικεύει (Εικόνα 11). Η πρώτη εικόνα (Activations) αναπαριστά τις αρχικές τιμές που δίδονται ως είσοδο. Τα Activation Gradients δείχνουν ποια τμήματα της εικόνας επηρεάζουν περισσότερο την εκπαίδευση.

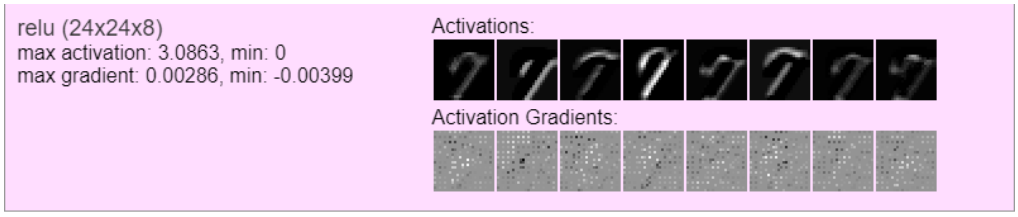
Εικόνα 11: Το input layer

Στο πρώτο επίπεδο συνέλιξης (Εικόνα 12) έχουμε έξοδο 24 x 24 x 8 διότι το επίπεδο εφαρμόζει φίλτρα στις εικόνες εισόδου και παράγει 8 διαφορετικές χαρτογραφήσεις χαρακτηριστικών (feature maps). Κάθε feature map φαίνεται πως αναδεικνύει ένα διαφορετικό χαρακτηριστικό της εικόνας (Activations). Τα Activation Gradients δείχνουν ποια τμήματα των feature maps επηρεάζουν περισσότερο την εκπαίδευση. Τα weights αντιπροσωπεύουν τα φίλτρα που έχει μάθει το convolutional layer. Τα weight gradients δείχνουν τα gradients για το εκάστοτε φίλτρο.



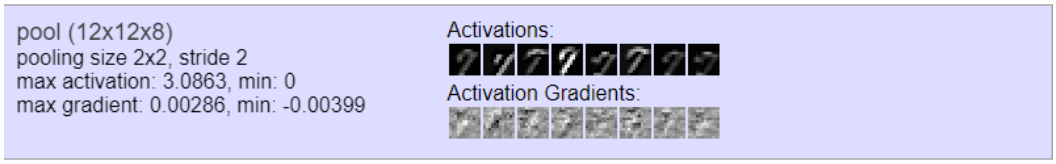
Εικόνα 12: Το Convolutional Layer

Το επόμενο layer περιλαμβάνει την εφαρμογή της συνάρτησης ενεργοποίησης ReLu η οποία διατηρεί τις θετικές τιμές ως έχουν και μηδενίζει τις υπόλοιπες (Εικόνα 13). Για αυτό παρατηρούμε πως οι εικόνες έχουν γίνει πιο μαύρες σε σχέση με το προηγούμενο επίπεδο. Να σημειωθεί πως η ReLu εισάγει μη γραμμικότητα στο μοντέλο η οποία είναι απαραίτητη για την κατανόηση περίπλοκων μοτίβων στα δεδομένα.



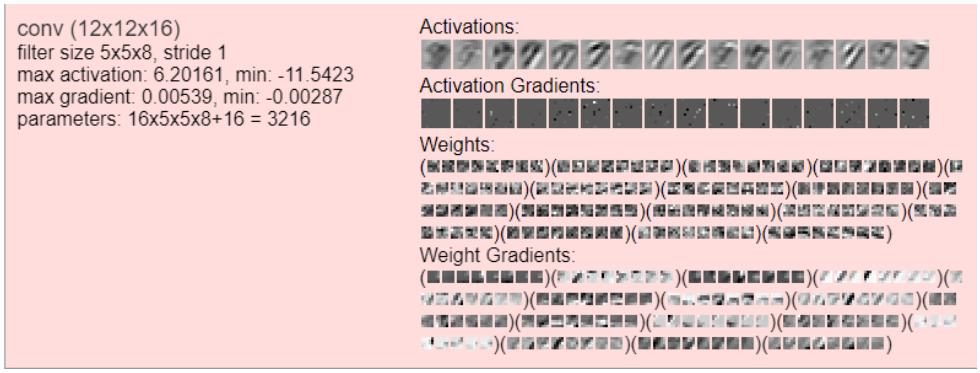
Εικόνα 13: Το ReLu layer

Το pooling layer έχει output size 12 x 12 x 8 διότι με χρήση ενός 2 x 2 παραθύρου μειώνει τις διαστάσεις των εικόνων στο μισό (συνολική μείωση 75%) (Εικόνα 14). Παρατηρούμε πως το ψηφίο παραμένει ευδιάκριτο.

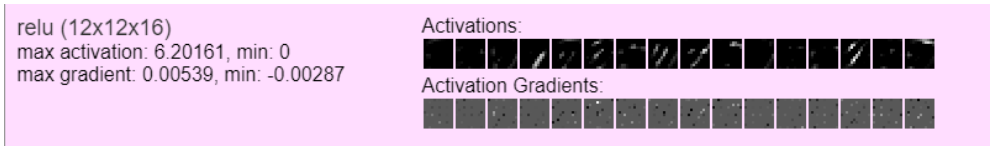


Εικόνα 14: Το pooling layer

Η διαδικασία επαναλαμβάνεται άλλη μία φορά με διαδοχικά convolution, ReLu και pooling layers (Εικόνες 15,16,17) χρησιμοποιώντας 16 φίλτρα αντί 8 που είχαμε στη προηγούμενη ομάδα layers.



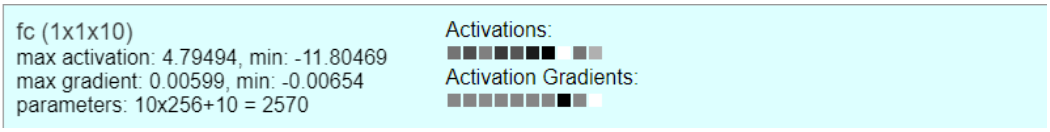
Εικόνα 15: Το 2ο Convolution Layer



Εικόνα 16: Το 2ο ReLu layer

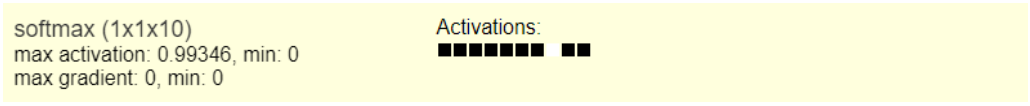
Εικόνα 17: Το 2ο pooling layer

Το fully connected layer (Εικόνα 18) έχει διαστάσεις 1 x 1 x 10 διότι συνδέει 2570 παραμέτρους/νευρώνες από το προηγούμενο επίπεδο σε 10 νευρώνες που αντιστοιχούν στις 10 κατηγορίες (τα 10 ψηφία).



Εικόνα 18: Το fully connected layer

Το softmax layer (Εικόνα 19) μετατρέπει το output του προηγούμενου layer σε πιθανότητες. Επομένως έχουμε το επικρατέστερο ψηφίο.



Εικόνα 19: Το softmax layer

Ερώτημα γ

	Λειτουργία	**Ρόλος**
2D convolution	Ένα φίλτρο/kernel διασχίζει κατά μήκος και πλάτος την εικόνα, υπολογίζοντας το εσωτερικό γινόμενο του φίλτρου και της εισόδου σε κάθε σημείο.	Η εξαγωγή χαρακτηριστικών από την εικόνα εισόδου (π.χ. γωνίες, ακμές).
Batch normalization	Κανονικοποιεί την έξοδο του 2D Convolution Layer ανά batch.	Διατήρηση των μ, σ^2 κάθε layer εντός μικρού εύρους επιτρέποντας γρηγορότερη εκπαίδευση.
ReLU activation	Εφαρμόζει μη γραμμικό μετασχηματισμό στα δεδομένα εισόδου θέτοντας τις αρνητικές τιμές 0 και διατηρώντας τις θετικές.	Εισαγωγή μη γραμμικότητας στο δίκτυο που του επιτρέπει να κατανοήσει περίπλοκα μοτίβα.

		Βοηθάει στο περιορισμό των vanishing gradients.
Max pooling	Χρησιμοποιόντας ένα μικρό παράθυρο, κάνει προσπέλαση της εικόνας επιλέγοντας το pixel του παραθύρου με τη μεγαλύτερη τιμή.	Η μείωση των διαστάσεων της εικόνας με μικρή απώλεια πληροφορίας.

Ερώτημα δ

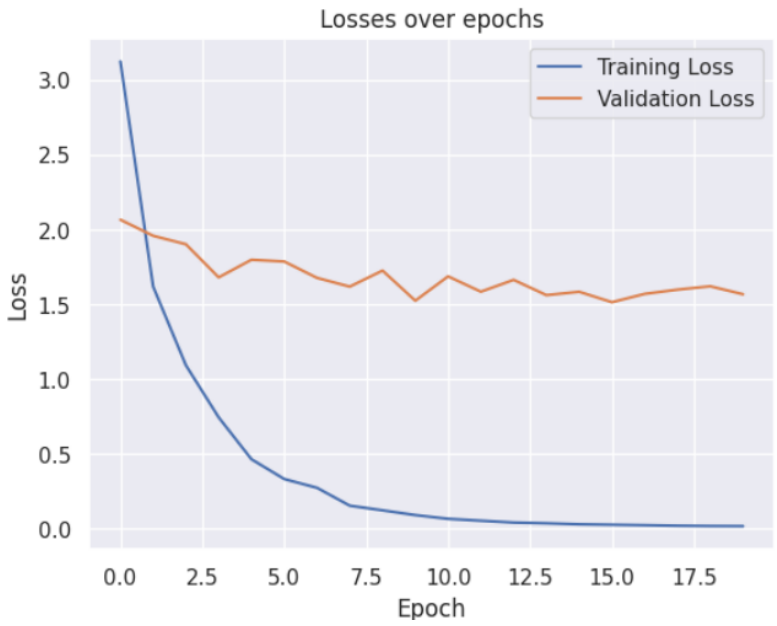
Πραγματοποιούμε τη διαδικασία batch overfitting και μέσα στις 100 πρώτες επαναλήψεις το σφάλμα πήρε πολύ μικρή τιμή ($1.8 * 10^{-5}$). Να σημειωθεί ότι έφτασε σε πολύ μικρό loss μετά από μικρότερο αριθμό επαναλήψεων σε σχέση με το LSTM.

Ερώτημα ε

Υλοποιούμε ένα δίκτυο CNN με 4 layers όπου το κάθε layer εκτελεί τις λειτουργίες 2D convolution, batch normalization, ReLu και max pooling.

Παρατηρούμε πως ενώ το Training Loss φτάνει πολύ χαμηλά, δεν υπάρχει ανάλογη βελτίωση στο Validation Loss, δηλαδή το μοντέλο δεν γενικεύει σωστά. Ενδεχομένως το Validation Set να μην περιέχει επαρκείς πληροφορίες για να αξιολογηθεί σωστά. (Εικόνα 20)

Παρόλα αυτά, το μοντέλο παρουσιάζει καλύτερες μετρικές από το αντίστοιχο LSTM για φασματογραφήματα, με τα macro avg και weighted avg να αυξάνονται σημαντικά. (Πίνακας 11)



Εικόνα 20: Σφάλματα Εκπαίδευσης και Επαλήθευσης για το CNN Network

	precision	recall	f1-score	support
0	0.31	0.20	0.24	40
1	0.51	0.70	0.59	40
2	0.58	0.60	0.59	80
3	0.48	0.56	0.52	80
4	0.54	0.55	0.54	40
5	0.33	0.17	0.23	40
6	0.65	0.60	0.63	78
7	0.05	0.03	0.03	40
8	0.38	0.44	0.41	103
9	0.28	0.38	0.32	34
accuracy			0.46	575
macro avg	0.41	0.42	0.41	575
weighted avg	0.44	0.46	0.45	575

Πίνακας 21: Οι μετρικές κατηγοριοποίησης του CNN

Βήμα 8

Ερώτημα α

Από το Βήμα 5 επιλέχθηκε το μοντέλο που έχει ως είσοδο τα beat-synced αρχεία καθώς σημείωσε τη μεγαλύτερη ακρίβεια.

Από τη στιγμή που πλέον έχουμε να κάνουμε με συνεχή ζητούμενα μεγέθη, δηλαδή με πρόβλημα παλινδρόμησης, τα μοντέλα τροποποιήθηκαν ώστε να αξιοποιούν την κλάση Regressor και το Μέσο Τετραγωνικό Σφάλμα (MSE) για τη συνάρτηση σφάλματος.

Να σημειωθεί πως καθώς δεν δίνεται αρχείο με labels για το test multitask_dataset, επεξεργαστήκαμε τη συνάρτηση torch_train_val_split ώστε να παρέχει τη δυνατότητα δημιουργίας και test_dataloader απομονώνοντας ένα ποσοστό του train_dataset.

Spearman Correlation

Καθώς δεν μπορούμε να χρησιμοποιήσουμε τις μετρικές που χρησιμοποιήθηκαν στον ταξινομητή, στη θέση τους θα αξιοποιηθεί το Spearman Correlation. Το Spearman Correlation αξιολογεί τη μονοτονική σχέση δύο μεταβλητών, αν δηλαδή τείνουν να αυξάνονται ή να ελαττώνονται μαζί αλλά όχι απαραίτητα με τον ίδιο ρυθμό. Υπολογίζεται από την εξής φόρμουλα (1). Λαμβάνει τιμές στο [-1,1], με -1/1 να δηλώνουν τέλεια αρνητική/θετική μονοτονική συσχέτιση και το 0 να δηλώνει την απουσία της.

$$\rho = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} \quad (1)$$

, όπου:

- $cov(R(X), R(Y))$*:* Η συνδιασπορά των βαθμών (ranks) των δύο τυχαίων μεταβλητών
- $\sigma_{R(X)}, \sigma_{R(Y)}$*:* Οι τυπικές αποκλίσεις των βαθμών των δύο τυχαίων μεταβλητών

Ερωτήματα β, γ, δ, ε

Εκπαιδεύοντας τα δύο μοντέλα πάνω στα τρία μεγέθη λάβαμε τις τιμές του Spearman Correlation:

backbone / task	valence	energy	danceability
LSTM	0.1237	0.2576	0.1724
CNN	-0.1537	-0.2764	0.1249

Παρατηρούμε ότι το καλύτερο αποτέλεσμα επιτυγχάνεται για τη μέτρηση του energy με χρήση του CNN. Υπάρχει μικρή βελτίωση ως προς το valence και το energy αλλά όχι ως προς το danceability.

Βήμα 9

Ερώτημα α

Στο paper How transferable are features in deep neural networks? παρουσιάζεται η δυνατότητα μεταφοράς ενός αριθμού αρχικών layers ενός εκπαιδευμένου δικτύου σε ένα καινούριο που δεν έχει εκπαιδευτεί. Η βασική ιδέα είναι πως τα αρχικά layers μαθαίνουν χαρακτηρίστηκα τα οποία είναι κοινά σε διαφορετικά προβλήματα. Παρόλα αυτά ακόμα και τα προβλήματα να είναι πολύ διαφορετικά μεταξύ τους, είναι προτιμότερο να μεταφερθούν τα layers από την αρχικοποίηση τυχαίων βαρών.

Ερωτήματα β, γ, δ, ε

Θα επιλεχτεί το CNN μοντέλο για φασματογραφήματα για την αρχική εκπαίδευση καθώς είχε σημειώσει τα μεγαλύτερα accuracy, micro avg και weighted avg.

Εκπαιδεύσαμε το δίκτυο πάνω στα φασματογραφήματα για 5 εποχές αφού πρώτα του φορτώσαμε τα εκπαιδευμένα βάρη για τον άξονα valence.

Το spearman correlation υπολογίζεται 0.0222. Δεν γίνεται αντιληπτή κάποια βελτίωση με χρήση αυτής της τεχνικής.

Βήμα 10

Ερώτημα α

Στο paper One Model To Learn Them All μελετάται η δυνατότητα συνδυασμού δεδομένων από διαφορετικούς τομείς (domains). Παρατηρείται ότι η απόδοση βελτιώνεται όταν τα δεδομένα είναι λίγα, υποδηλώνοντας τη συσχέτιση πληροφοριών που περιέχουν οι τομείς. Όταν τα δεδομένα είναι πολλά δεν παρατηρείται βελτίωση, παρόλα αυτά.

Ερωτήματα β,γ,δ

Υλοποιήθηκε καινούριες κλάσεις MultitaskRegressor και MultitaskLoss ώστε να μπορούμε να αξιοποιήσουμε και τους τρεις άξονες. Με 10 επαναλήψεις λάβαμε:

valence	0.3584
energy	0.5122
danceability	0.4133

Παρατηρούμε πως η τεχνική Multitask Learning βελτίωσε αισθητά το Spearman Correlation και για τους 3 άξονες. Επομένως, υπήρχε όντως κρυμμένη συσχετισμένη πληροφορία ανάμεσα στους άξονες. Αυτό μπορούσε κανείς να το υποπτευθεί και διαισθητικά καθώς τραγούδια με υψηλή ενέργεια συνήθως προκαλούν έντονα συναισθήματα που οδηγούν στο χορό.