ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Σύνθεση ήχων περιβάλλοντος πόλης με χρήση αλγορίθμων βαθιάς μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γεώργιος Κωνσταντίνος Κ. Μελέτης

Επιβλέπων : Γεώργιος Στάμου
Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2020

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Σύνθεση ήχων περιβάλλοντος πόλης με χρήση αλγορίθμων βαθιάς μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γεώργιος Κωνσταντίνος Κ. Μελέτης

Επιβλέπων : Γεώργιος Στάμου
Αν. Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18η Νοεμβρίου 2020.

.....	.....	.....
Γεώργιος Στάμου	Στέφανος Κόλλιας	Ανδρέας-Γεώργιος Σταφυλοπάτης
Αν. Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π

Αθήνα, Νοέμβριος 2020

.....
Γεώργιος Κωνσταντίνος Κ. Μελέτης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η εξέλιξη των νευρωνικών δικτύων και της μηχανικής μάθησης γίνεται ολοένα και πιο ραγδαία, με τις δυνατότητες τους να ξεπερνούν κάθε νέο όριο που συναντούν. Ένα πρόβλημα που τεστάρει τις δυνατότητες των νευρωνικών δικτύων σε πολύ υψηλό επίπεδο, είναι η διαδικασία παραγωγής νέων δεδομένων. Η κατηγορία των δικτύων που ασχολούνται με την διαδικασία παραγωγής νέων δεδομένων έχει λιγότερο από μια δεκαετία που άρχισε να αναπτύσσεται, διότι οι απαιτήσεις σε υπολογιστικούς πόρους είναι πολύ υψηλές. Τα πιο γνωστά δίκτυα παραγωγής δεδομένων είναι τα Generative Adversarial Networks (GANs) και οι Variational Autoencoders (VAEs).

Η παραγωγή ήχου με τη χρήση νευρωνικών δικτύων έχει επικεντρωθεί στην παραγωγή ήχων με υψηλό βαθμό οργάνωσης, όπως η μουσική. Στον τομέα αυτό έχουν εξελιχθεί πολύ ικανά δίκτυα, των οποίων οι δυνατότητες προσεγγίζουν γοργά τις δυνατότητες ενός ικανού συνθέτη μουσικής. Παρόλα αυτά, δεν έχει δοθεί τόση σημασία στην παραγωγή ήχων με υψηλό βαθμό τυχαιότητας, όπως οι ήχοι του φυσικού περιβάλλοντος, καθώς η τυχαιότητα που καλούμαστε να αντιμετωπίσουμε αυξάνει αυτόματα τις απαιτήσεις σε υπολογιστική ισχύ. Έτσι δεν φαίνεται να έχει δοθεί μια εκτίμηση για το ποια κατηγορία δικτύων ή ποια μορφή αναπαράστασης τέτοιων ήχων είναι η ιδανική ώστε να υλοποιηθεί μια παραγωγική διαδικασία.

Η παρούσα εργασία ασχολείται με την σύνθεση ήχων με υψηλό βαθμό τυχαιότητας, όπως οι ήχοι που συναντά κανείς σε ένα αστικό περιβάλλον και με το κατά πόσο ένα νευρωνικό δίκτυο μπορεί από μια οπτική αναπαράσταση των ήχων αυτών να παράγει παρόμοιους αλλά εντελώς νέους ήχους. Οι αναπαραστάσεις που ελέγχονται και αξιολογούνται στα πλαίσια της εργασίας, είναι καθαρά οπτικές αναπαραστάσεις του ήχου και αφορούν το φασματογράφημα (Spectrogram), το φασματογράφημα mel (Mel-spectrogram) και τους συντελεστές συχνότητας Cepstral του Mel (Mel-Frequency Cepstral Coefficients - MFCCs). Το δίκτυο που επιλέχθηκε ως βάση για έρευνα, είναι ο VAE και συγκεκριμένα μια παραλλαγή του, η οποία στηρίζεται στο συνδυασμό των συνελικτικών δικτύων (CNN) με τον VAE και ονομάζεται Convolutional Variational Autoencoder (CVAE). Τα αποτελέσματα έδειξαν πως οι οπτικές αναπαραστάσεις του ήχου μπορεί να έχουν λιγότερο κόστος σε μνήμη, αλλά έχουν ως αποτέλεσμα την απώλεια σημαντικής πληροφορίας. Η αναπαράσταση που έδωσε τα καλύτερα αποτελέσματα φάνηκε να ήταν το φασματογράφημα-Mel, με την αναπαράσταση MFFC να ακολουθεί και τέλος το απλό φασματογράφημα. Η αρχιτεκτονική των βαθιών νευρωνικών δικτύων που χρησιμοποιήθηκε φάνηκε να παίζει μικρό ρόλο σε σχέση με την ποιότητα και την ποσότητα των δεδομένων εκπαίδευσης, καθώς φάνηκε να προσεγγίζουν το μέγιστο των δυνατοτήτων τους με βάση τα δεδομένα που τους παρασχέθηκαν.

Τα αποτελέσματα δείχνουν πως η καταλληλότητα των VAE για την παραγωγή νέων ήχων στηρίζεται κατά κύριο λόγο στην ποιότητα και την ποσότητα των διαθέσιμων δεδομένων. Η παρούσα υλοποίηση δείχνει πρώιμα αλλά σημαντικά αποτελέσματα πάνω στην παραγωγή ήχου υψηλής τυχαιότητας από δίκτυα VAE και βάζει τις βάσεις για πιο εξελιγμένα παραγωγικά μοντέλα που χωρίς αμφιβολία θα δημιουργηθούν στο προσεχές μέλλον.

Λέξεις Κλειδιά: Variational Autoencoder, Παραγωγικά μοντέλα, Παραγωγή ήχου, Συνελικτικά Νευρωνικά Δίκτυα, Συνελικτικός Variational Autoencoder, Αναπαράσταση ήχου, Sound Design.

Abstract

The evolution of neural networks and machine learning is becoming more and more rapid and they consistently surpass any new limit they encounter. One problem that tests the capabilities of neural networks at a very high level is the process of generating new data. The category of networks involved in the process of generating new data has been developing for less than a decade, because the computing resources demands are very high. The most well-known data generation networks are Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs).

Sound production using neural networks has focused on producing sounds with a high degree of organization, such as music. Very capable networks have developed in this field, the capabilities of which are rapidly approaching the capabilities of a competent music composer. However, not as much importance has been given to the production of sounds with a high degree of randomness, like the sounds of the natural environment, as the randomness we are called to deal with automatically increases the computing power requirements. Thus no assessment seems to have been made as to which category of networks or what form of representation of such sounds is ideal in order to carry out a data producing process.

This project deals with the synthesis of sounds with a high degree of randomness, such as the sounds encountered in an urban environment and whether a neural network can produce similar but completely new sounds from a visual representation of these sounds. The representations that are checked and evaluated in the context of the work are purely visual representations of sound and concern the spectrogram, the Mel-spectrogram and the Mel-Frequency Cepstral Coefficients (MFCCs). The network chosen as the basis for research is VAE and specifically a variant of it, which is based on the combination of convolutional neural networks (CNNs) with VAE and is called Convolutional Variational Autoencoder (CVAE). The results showed that visual representations of sound may cost less memory, but result in the loss of important information. The representation that gave the best results appeared to be the Mel-spectrogram, followed by the MFFC representation and the normal spectroscopy. The deep neural network architectures used, seemed to play a small role in relation to the quality and quantity of training data, as they seemed to approach their maximum potential based on the data provided to them.

The results show that the suitability of VAE for the production of new sounds is based mainly on the quality and quantity of available data. The present implementation shows early but significant results on the production of high-randomness sound from VAE networks and lays the foundations for the production of more sophisticated models that will certainly be created in the near future.

Key words: Variational Autoencoder, Generative Models, Sound Generation, Convolutional Neural Networks, Convolutional Variational Autoencoder, Sound Representation, Sound Design.

Ευχαριστίες

Η εκπόνηση της παρούσας εργασίας πραγματοποιήθηκε σε μια παράξενη εποχή, όχι μονό για εμένα, αλλά για ολόκληρη την ανθρωπότητα. Υπεύθυνος για αυτό είναι φυσικά ο ιός SARS-CoV-2 (covid-19), ο οποίος άλλαξε τις συνθήκες ζωής και τις συνήθειες όλων. Κάτω, λοιπόν, από αυτές τις πρωτόγνωρες συνθήκες η ανάγκη για στήριξη και ώθηση από τον περίγυρο έγινε μεγαλύτερη από ποτέ.

Αρχικά, θέλω να ευχαριστήσω τον επιβλέποντα καθηγητή αυτής της διπλωματικής εργασίας, κύριο Γιώργο Στάμου, καθώς πέρα από την επίβλεψη αυτής της εργασίας, αποτέλεσε σημαντικό σύμβουλο, βοηθό και δάσκαλο καθ' όλη την διάρκεια των σπουδών μου και μου έδωσε την έμπνευση για την κατεύθυνση που επέλεξα να ακολουθήσω, αυτή της Επιστήμης Δεδομένων.

Ακόμη, θα ήθελα να ευχαριστήσω τον Έντμοντ Ντερβάκο (Edmund Dervakos), ο οποίος ήταν διαθέσιμος ανά πάσα ώρα να προσφέρει βοήθεια, έμπνευση και νέες ιδέες, απαραίτητες για να ολοκληρωθεί τελικά αυτή η εργασία. Πέρα από τις τεχνικές του γνώσεις, αποτέλεσε σημαντικό υποστηρικτή, όταν τα ψυχολογικά αποθέματα έμοιαζαν να τελειώνουν.

Επίσης, ένα μεγάλο ευχαριστώ αξίζει στους καλούς μου φίλους, οι οποίοι ήταν δίπλα μου όλα αυτά τα χρόνια κάνοντας με έναν καλύτερο άνθρωπο. Σε αυτούς συγκαταλέγω και τους αγαπημένους μου τετράποδους φίλους, οι οποίοι βρίσκονταν συνεχώς δίπλα μου τις αμέτρητες ώρες μελέτης των φοιτητικών αυτών χρόνων.

Τέλος, θέλω να ευχαριστήσω την οικογένεια μου, που αποτέλεσε και αποτελεί το σημαντικότερο μου στήριγμα σε κάθε στάδιο της ζωής μου. Αφιερώνω αυτή την προσπάθεια σε αυτούς και ιδιαίτερα στον παππού μου, Γιώργο, ο οποίος μπορεί να μην βρίσκεται πια κοντά μου αλλά μου έδωσε τα εφόδια και τη δύναμη να κυνηγήσω τα όνειρα μου.

Γεώργιος Κωνσταντίνος Κ. Μελέτης

Περιεχόμενα

Πίνακας σχημάτων

1

1. Εισαγωγή

1.1 Κίνητρο

Η επιθυμία του ανθρώπου να χειραγωγεί τον ήχο μας οδηγεί χιλιετίες πίσω. Μάλιστα η απαρχή της σύνθεσης μουσικής δεν μπορεί να εντοπιστεί ακριβώς μιας και ξεκινάει πριν από την καταγραφή οποιασδήποτε ιστορικής πηγής. Από τότε μέχρι σήμερα και παράλληλα με την εξέλιξη του, το ανθρώπινο είδος έχει εφεύρει πολλούς νέους τρόπους για να παράγει ήχους, οι οποίοι κάνουν χρήση των διαθέσιμων μέσων και τεχνολογιών της εκάστοτε εποχής.

Ο όρος σχεδίαση ήχου (sound design) χρησιμοποιήθηκε επίσημα για πρώτη φορά το 1979 στην ταινία Apocalypse Now του Francis Ford Coppola, για να χαρακτηρίσει την δουλειά του Walter Murch. Η σχεδίαση ήχου όμως υπήρχε πολύ πριν από αυτό και μέχρι και σήμερα αποτελεί αναπόσπαστο κομμάτι οποιασδήποτε κινηματογραφικής ή θεατρικής παραγωγής. Χάρη στην εξέλιξη της τεχνολογίας η σχεδίαση ήχου ξεκίνησε να παίζει καθοριστικό ρόλο στην ανάπτυξη των ηλεκτρονικών παιχνιδιών και της εικονικής πραγματικότητας, που αποτελεί ενδεχομένως το επόμενο βήμα στη βιομηχανία της ψυχαγωγίας και του θεάματος. Παρόλα αυτά έως τώρα ο σχεδιασμός του ήχου γίνεται με έναν αρκετά ντετερμινιστικό τρόπο. Αυτό σημαίνει ότι σε ένα βιντεοπαιχνίδι που οδηγούμε ένα όχημα θα ακούμε ξανά και ξανά τον ίδιο ήχο από τον κινητήρα και αφού παίξουμε για μισή ώρα θα έχουμε ακούσει κάθε δυνατό ήχο που μπορεί να παράγει. Πως θα ήταν όμως αν το παιχνίδι ήξερε τι ήχους μπορεί να παράγει ο συγκεκριμένος κινητήρας και μπορούσε να δημιουργεί καινούργιους ήχους καθώς παίζουμε; Η εργασία αυτή αποτελεί ένα βήμα προς αυτή την κατεύθυνση, αφού προσπαθούμε να κάνουμε έναν υπολογιστή να μάθει να παράγει καινούργιους φυσικούς ήχους . Οι δυνατότητες από εκεί και πέρα είναι αμέτρητες καθώς το κόστος και ο χρόνος σε σχέση με την παραδοσιακή σχεδίαση ήχου πέφτουν κατακόρυφα, ενώ οι δυνατότητες παραγωγής νέων ήχων ολοένα και αυξάνονται, αφού κάθε ήχος που έχει ηχογραφηθεί η παραχθεί σε στούντιο για χρήση σε μια ταινία, για παράδειγμα, δεν θα πετιέται αλλά θα γίνεται πηγή (είσοδος του δικτύου) για την παραγωγή νέων παρόμοιων ήχων.

Στην επόμενη ενότητα θα ορίσουμε το αντικείμενο της εργασίας, τοποθετώντας το σε ένα συγκεκριμένο πλαίσιο και θα διατυπώσουμε επακριβώς τα ερωτήματα που καλείται να απαντήσει.

1.2 Αντικείμενο διπλωματικής

Σκοπός της εργασίας είναι η δημιουργία και η εκπαίδευση κατάλληλων νευρωνικών δικτύων, βαθιάς μηχανικής μάθησης, προκειμένου αυτά να μάθουν να παράγουν νέους ήχους που δεν έχουν δει ποτέ. Πιο συγκεκριμένα, τα νευρωνικά αυτά δίκτυα θα περνούν από την διαδικασία της εκπαίδευσης, κατά την οποία θα βλέπουν στην είσοδο τους έναν μεγάλο όγκο από διαφορετικούς ηχούς που συναντά κανείς στο περιβάλλον μιας πόλης (βλ. Κεφάλαιο 4). Σκοπός του δικτύου δεν θα είναι να αναπαράγει στην έξοδο του αυτό που είδε στην είσοδο, αλλά να μάθει τα χαρακτηριστικά που προσδιορίζουν έναν ήχο και άρα να μπορεί όταν του ζητηθεί να παράγει έναν νέο ήχο που θα έχει αυτά τα χαρακτηριστικά. Η διαδικασία αυτή μπορεί να είναι ξεκάθαρη σε κάποιον που διαθέτει γνώσεις γύρω από τα νευρωνικά δίκτυα και την μηχανική μάθηση αλλά προκειμένου να απλοποιήσουμε τον κεντρικό σκοπό της εργασίας για έναν οποιονδήποτε αναγνώστη παρουσιάζουμε το παράδειγμα που ακολουθεί στη συνέχεια.

Ας υποθέσουμε πως το νευρωνικό μας δίκτυο είναι ένα παιδί ηλικίας τριών ετών. Ας υποθέσουμε επίσης πως το παιδί αυτό μέχρι εκείνη την ηλικία δεν έχει ακούσει ποτέ στη ζωή του κανέναν απολύτως ήχο. Αυτό σημαίνει πως αν κάποιος του ζητήσει “να κάνει το σκυλάκι”, το παιδί δεν θα έχει ιδέα πως να το κάνει, αφού δεν έχει ακούσει ποτέ του πως γαβγίζει ένας σκύλος. Το ίδιο φυσικά ισχύει και για οποιονδήποτε ήχο του ζητηθεί να παράγει. Ξημερώνει μια μέρα και ξαφνικά το παιδί ξεκινάει να ακούει. Όταν θα ακούσει για πρώτη φορά έναν σκύλο και παράλληλα ένας μεγαλύτερος είναι εκεί να του υποδείξει πως αυτό που άκουσε είναι πράγματι ένας σκύλος, τότε θα ξεκινήσει η διαδικασία της εκπαίδευσης. Αν ζητηθεί από το παιδί εκ νέου να παράγει έναν ήχο σαν σκύλος, αυτή τη φορά το παιδί θα έχει την τάση να μιμηθεί αυτό ακριβώς που άκουσε μιας και πρόκειται για τον μοναδικό σκύλο που έχει ακούσει. Αυτό σημαίνει πως δεν γνωρίζει τα χαρακτηριστικά του γαβγίσματος ακόμα, άρα δεν γνωρίζει πως αυτό μπορεί να είναι πιο βαρύ ή πιο ψηλό, πιο γρήγορο ή πιο αργό, με διάρκεια ή πιο κοφτό. Αυτά θα ξεκινήσει να τα μαθαίνει όταν θα έχει ακούσει δεκάδες ή εκατοντάδες σκυλιά και τότε είναι που αν του ζητηθεί σε ανύποπτη στιγμή να παράγει το γάβγισμα ενός σκύλου, θα παράγει έναν νέο ήχο ως αποτέλεσμα των γνώσεων που έχει γύρω από τα χαρακτηριστικά του γαβγίσματος. Με πολύ παρόμοιο τρόπο αντιμετωπίζουμε τα δίκτυα μας στην παρούσα εργασία.

Δυστυχώς σε έναν υπολογιστή δεν έχουμε ακόμη τις εκπληκτικές δυνατότητες του ανθρωπίνου μυαλού, σε ότι αφορά την αναλυτική ικανότητα και την διαθέσιμη μνήμη. Επομένως το αντικείμενο της εργασίας οφείλει να διαφοροποιηθεί από το παραπάνω παράδειγμα σε μερικά σημεία. Ένα πολύ σημαντικό σημείο είναι η επιλογή που θα κάνουμε στον τρόπο με τον οποίο θα αναπαραστήσουμε τον ήχο, έτσι ώστε να μπορέσουμε να αντλήσουμε όσο περισσότερη πληροφορία γίνεται χωρίς παράλληλα να γίνεται σπατάλη μνήμης.

1.3 Χρησιμότητα εργασίας

Στο Κεφάλαιο 3 αναλύουμε κάποιες δημοσιεύσεις από τις οποίες αντλήσαμε πολλές και σημαντικές πληροφορίες που βοήθησαν στην διεκπεραίωση της παρούσας εργασίας. Όπως θα παρατηρήσουμε και αργότερα, παρά το γεγονός ότι έχουν γραφτεί αμέτρητες δημοσιεύσεις γύρω από τη σύνθεση ήχου, μόνο ελάχιστες από αυτές έχουν ασχοληθεί με την παραγωγή φυσικών ήχων του περιβάλλοντος. Το συντριπτικό ποσοστό ασχολείται με τη σύνθεση μουσικής και φωνής που αποτελούν φυσικά δυο ακόμη πολύ ενδιαφέροντα, αλλά διαφορετικά προβλήματα. Έτσι η εργασία αυτή έρχεται να συμπληρώσει ένα κομμάτι που δεν έχει μελετηθεί αρκετά ακόμη, αλλά σίγουρα θα απασχολήσει την κοινότητα της μηχανικής μάθησης στο κοντινό μας μέλλον, μιας και όπως ήδη

αναφέραμε η σχεδίαση φυσικού ήχου παίζει καθοριστικό ρόλο σε μερικές από τις πιο κερδοφόρες βιομηχανίες του κόσμου.

Μια πιθανή εφαρμογή της παρούσας εργασίας, ενδεχομένως ακόμη και αυτούσιας, έκανε την εμφάνιση της όσο η εκτέλεση της εργασίας βρισκόταν ακόμη σε εξέλιξη. Η εργασία αυτή εκτελέστηκε κατά ένα μεγάλο μέρος της εν μέσω της πανδημίας του SARS-CoV-2 (covid-19). Κατά το διάστημα αυτό, όπως είναι γνωστό όλες οι αθλητικές δραστηριότητες είχαν διακοπεί. Κατά την επανεκκίνηση τους πολλές αθλητικές ομοσπονδίες αποφάσισαν στα γήπεδα να ακούγονται από τα μεγάφωνα των γηπέδων ήχοι από το κοινό, μιας και πραγματικό κοινό δεν ήταν δυνατό να παρευρεθεί. Έτσι αναγκάστηκαν να έρθουν σε επαφή με δημιουργούς αθλητικών ηλεκτρονικών παιχνιδιών, προκείμενου να προμηθευτούν ήχους από μεγάλα κοινά και οπαδούς, που φυσικά είναι περιορισμένα και επαναλαμβανόμενα. Παρότι λοιπόν μπορεί να αντιλαμβανόμαστε τον φυσικό ήχο ως κάτι δεδομένο, η απουσία αυτού γίνεται πάντα αισθητή αμέσως. Με αυτόν τον τρόπο γίνεται ξεκάθαρο, πως η ύπαρξη μιας έξυπνης μηχανής που μπορεί να παράγει αληθοφανείς φυσικούς ήχους, που περιλαμβάνουν και το στοιχείο της τυχαιότητας, σε αντίθεση με τις μέχρι τώρα μεθόδους σύνθεσης ήχου, έχει αρκετούς λόγους ύπαρξης.

Παράλληλα στο πλαίσιο της εργασίας ασχολούμαστε με τη δημιουργία μοντέλων, όπως ο συνελικτικός (convolutional) variational autoencoder, που δεν έχουν μελετηθεί ιδιαίτερα αλλά παρουσιάζουν εξαιρετικό ενδιαφέρον και προοπτικές εξέλιξης.

1.4 Οργάνωση κειμένου

Στη συνέχεια παραθέτουμε μια σύντομη περίληψη των κεφαλαίων που ακολουθούν:

Στο Κεφάλαιο 2 γίνεται αναφορά σε όλες τις προαπαιτούμενες γνώσεις που χρειάζονται για την πλήρη κατανόηση της εργασίας, σε ότι αφορά τη μηχανική μάθηση, τα στοιχεία των νευρωνικών δικτύων καθώς και βασικές έννοιες και τεχνικές της ψηφιακής επεξεργασίας ήχου.

Στο Κεφάλαιο 3 παραθέτουμε κάποιες από τις σημαντικότερες δημοσιεύσεις από την βιβλιογραφία και επικεντρωνόμαστε σε διάφορα ξεχωριστά βοηθητικά σημεία που μας παρείχε η κάθε μια. Στόχος αυτού του κεφαλαίου δεν είναι η περίληψη των δημοσιεύσεων αλλά ο εντοπισμός και η αναφορά σε συγκεκριμένα σημεία αυτών, τα οποία φάνηκαν χρήσιμα στη διεκπεραίωση της εργασίας.

Στο Κεφάλαιο 4 προχωράμε στις μεθόδους που χρησιμοποιήθηκαν στην παρούσα εργασία. Αρχικά αναλύουμε το σύνολο δεδομένων που χρησιμοποιήσαμε. Στη συνέχεια παρουσιάζεται η προεπεξεργασία των δεδομένων και οι τρόποι αναπαράστασης αυτών. Τέλος παραθέτουμε τις αρχιτεκτονικές των δικτύων που δοκιμάστηκαν για τις ανάγκες τις παρούσας εργασίας.

Στο Κεφάλαιο 5 παρατίθενται τα αποτελέσματα των πειραμάτων μας. Ακόμη γίνεται σύγκριση των αποτελεσμάτων με σκοπό να εντοπιστεί το μοντέλο με τα καλύτερα αποτελέσματα μεταξύ αυτών που κατασκευάσαμε.

Στο Κεφάλαιο 6 αναλύουμε τα συμπεράσματα που προέκυψαν από την εργασία και τα αποτελέσματα αυτής καθώς και μερικές από τις πιθανές μελλοντικές εξελίξεις όσον αφορά τη σχεδίαση ήχου κάνοντας χρήση μηχανικής μάθησης και συγκεκριμένα variational autoencoders.

2

2. Χρήσιμες έννοιες

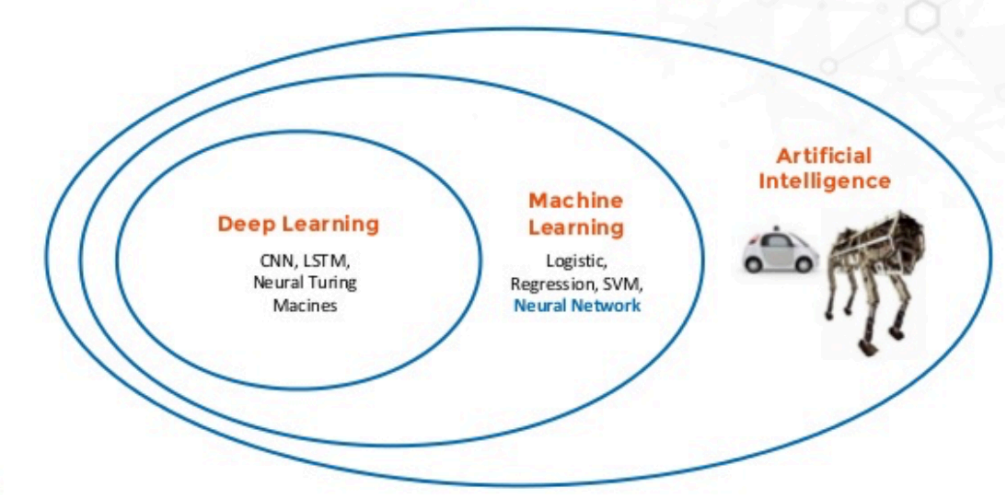
2.1 Μηχανική μάθηση

Κάθε ζωντανός οργανισμός στον πλανήτη μας επιτελεί κάποιας μορφής μάθηση. Όσο πιο εξελιγμένη είναι η νοημοσύνη ενός οργανισμού τόσο πιο σύνθετα είναι και τα πράγματα τα οποία μπορεί να κατανοήσει και να μάθει. Ο άνθρωπος βρίσκεται στη κορυφή αυτής της πυραμίδας των νοημόνων οργανισμών μιας και η ικανότητα μάθησης που κατέχει είναι πολύ πιο ισχυρή από οποιονδήποτε άλλο ζωντανό οργανισμό.

Φτάνοντας στην εποχή της άνθησης των υπολογιστικών συστημάτων αναδύθηκε μια μορφή μάθησης που δεν χαρακτηρίζει πλέον ζωντανούς οργανισμούς, αλλά υπολογιστικά συστήματα, η μηχανική μάθηση. Η μηχανική μάθηση πλέον αφορά ένα αυτόνομο και ευρύ επιστημονικό πεδίο, αρχικά όμως προέκυψε σαν παρακλάδι της τεχνητής νοημοσύνης και αυτός είναι και ο κύριος λόγος που πολλές φορές οι δύο αυτές έννοιες συγχέονται. Η τεχνητή νοημοσύνη αφορά ένα πεδίο της επιστήμης των υπολογιστών που αναπτύσσεται από το 1956 και θα μπορούσαμε να αφιερώσουμε πολλές σελίδες παρουσιάζοντας τα αμέτρητα ενδιαφέροντα επιτεύγματα της, παρόλα αυτά θα αρκεστούμε στο να τονίσουμε την κύρια διαφορά μεταξύ αυτής και της μηχανικής μάθησης. Ο σκοπός της τεχνητής νοημοσύνης επικεντρώνεται στην δημιουργία έξυπνων αλγορίθμων, οι οποίοι προσομοιώνουν την ανθρώπινη συμπεριφορά και μπορούν να εκτελεστούν από υπολογιστές προκειμένου να φέρουν εις πέρας την επιθυμητή εργασία. Για τον σκοπό αυτό οι προγραμματιστές αναλαμβάνουν να μάθουν σε έναν υπολογιστή όλες τις πιθανές πληροφορίες και δεδομένα που μπορεί να χρειάζεται προκειμένου να πραγματοποιήσει την συγκεκριμένη εργασία. Σε αυτό ακριβώς το σημείο συναντάται και η ειδοποιός διαφορά με την μηχανική μάθηση. Η μηχανική μάθηση αφορά την διαδικασία κατά την οποία ο προγραμματιστής δίνει σε έναν υπολογιστή όλα τα απαραίτητα εργαλεία προκειμένου να μάθει μόνος του να φέρνει εις πέρας μια εργασία. Για να γίνει πιο κατανοητή η διαφορά παραθέτουμε το έξης παράδειγμα:

Φανταζόμαστε πως γυρνάμε σε μια εποχή οπού οι άνθρωποι δεν χρησιμοποιούσαν κανένα εργαλείο και σκοπός μας είναι μέχρι να φύγουμε να τους κάνουμε τη ζωή ευκολότερη, με το να τους παρέχουμε διάφορα βασικά εργαλεία. Για τον σκοπό αυτό έχουμε δυο πιθανές λύσεις. Αρχικά θα μπορούσαμε να απομονωθούμε και να ξεκινήσουμε να φτιάχνουμε δεκάδες εργαλεία για πολλές μέρες και την τελευταία να πλησιάζαμε τους ανθρώπους, να τους δείχναμε πως να τα

χρησιμοποιούν και να φεύγαμε. Ο σκοπός σίγουρα θα επιτευχθεί και η ζωή τους θα γίνει πιο εύκολη. Από την άλλη όμως υπάρχει μια δεύτερη εναλλακτική, η οποία είναι να συγκεντρώσουμε όλους τους ανθρώπους και μπροστά στα ματιά τους να αρχίσουμε σιγά σιγά να κατασκευάζουμε ένα σφυρί χρησιμοποιώντας πέτρα και ξύλο. Σε αυτό το σενάριο δεν θα πάρει πολύ χρόνο μέχρι και οι ίδιοι να αρχίσουν να πειραματίζονται και να κάνουν απόπειρες να κατασκευάσουν το δικό τους σφυρί. Αυτή η διαφορετική προσέγγιση προσφέρει στους ανθρώπους κάτι σημαντικότερο από τα ίδια τα εργαλείο και αυτό είναι η γνώση του να τα κατασκευάζεις. Αυτή είναι και η προσέγγιση που υλοποιεί η μηχανική μάθηση.



Σχήμα 2.1: Η σχέση μεταξύ των πεδίων της τεχνητής νοημοσύνης (AI), της μηχανικής μάθησης (ML) και της βαθιάς μηχανικής μάθησης (DL).

Αφού παρουσιάσαμε τη μηχανική μάθηση διαισθητικά, μπορούμε τώρα να παραθέσουμε και τον πιο τυπικό ορισμό της. Ο ορισμός της μηχανικής μάθησης που έχει επικρατήσει, ως ο πιο πλήρης, είναι αυτός που έδωσε ο Mitchell (1997) και την ορίζει ως εξής:

> *«Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία E ως προς κάποια κλάση εργασιών T και μετρό απόδοσης P, αν η απόδοση του σε εργασίες από το T, όπως μετριέται από το P, βελτιώνεται μέσω της εμπειρίας E.»*

Στη συνέχεια, για λόγους κατανόησης και πληρότητας του ορισμού, παραθέτουμε τις έννοιες των συμβόλων T(εργασία), E(εμπειρία) και P(μέτρο επίδοσης) (Ian Goodfellow, 2016).

Η Εργασία (Task), *T*

Η έννοια της εργασίας δεν πρέπει να συγχέεται με την έννοια της μάθησης. Η μάθηση είναι το μέσο μας για την επίτευξη της ικανότητας εκτέλεσης της εργασίας. Για παράδειγμα, αν θέλουμε ένα ρομπότ να μπορεί να περπατάει, τότε το περπάτημα είναι η εργασία. Για την επίτευξη αυτού του στόχου θα μπορούσαμε να προγραμματίσουμε το ρομπότ προκειμένου να μάθει να περπατάει, ή θα μπορούσαμε να προσπαθήσουμε να γράψουμε απευθείας ένα πρόγραμμα που καθορίζει πως να εκτελέσει την εργασία του περπατήματος.

Οι εργασίες μηχανικής μάθησης περιγράφονται συνήθως ως προς τον τρόπο με τον οποίο το σύστημα μηχανικής μάθησης χειρίζεται ένα δείγμα. Ένα δείγμα είναι ένα σύνολο χαρακτηριστικών

που έχουν μετρηθεί ποσοτικά από κάποιο αντικείμενο ή γεγονός που θέλουμε να επεξεργαστεί το σύστημα μηχανικής μάθησης. Αντιπροσωπεύουμε συνήθως ένα δείγμα ως ένα διάνυσμα $x \in R^n$ όπου κάθε είσοδος x_i του διανύσματος είναι ένα ακόμη χαρακτηριστικό. Για παράδειγμα, τα χαρακτηριστικά μιας εικόνας είναι συνήθως οι τιμές όλων των εικονοστοιχείων στην εικόνα. Ανάλογα με τον τελικό σκοπό υπάρχουν πολλοί διαφορετικοί τύποι εργασιών. Ενδεικτικά στη συνέχεια παρουσιάζουμε τους δύο πιο γνωστούς, την ταξινόμηση (classification) και την παλινδρόμηση (regression), καθώς και την σύνθεση που μας απασχολεί στην εργασία αυτή:

- **Ταξινόμηση (Classification):** Στα προβλήματα ταξινόμησης, ο σκοπός είναι η κατασκευή μιας συνάρτησης $f : R^n \rightarrow \{1,...,k\}$, που χαρτογραφεί κάθε δείγμα της εισόδου σε μια από k κατηγορίες. Όταν $y = f(x)$ το μοντέλο αντιστοιχεί μια είσοδο που περιγράφεται από τον διάνυσμα εισόδου x σε μια κατηγορία που έχει ετικέτα (label) y. Σε μια παραλλαγή της εργασίας της ταξινόμησης αντί για μια κατηγορία στην έξοδο, λαμβάνουμε ένα διάνυσμα που για κάθε κατηγορία μας υποδεικνύει την πιθανότητα η συγκεκριμένη είσοδος να ανήκει σε αυτήν.

- **Παλινδρόμηση (Regression):** Σε αυτή την περίπτωση εργασίας, η συνάρτηση που καλείται να κατασκευάσει ο αλγόριθμος μηχανικής μάθησης είναι της μορφής $f : R^n \rightarrow R$, δηλαδή η έξοδος είναι μια πραγματική τιμή ενώ η είσοδος μπορεί να έχει οποιοδήποτε μήκος. Η παλινδρόμηση μοιάζει αρκετά με την ταξινόμηση αφού διαφέρουν μόνο στην έξοδο. Ένα πρόβλημα παλινδρόμησης είναι, για παράδειγμα, η πρόβλεψη της τιμής ενός σπιτιού αν ως είσοδο έχουμε διάφορα στοιχεία, όπως το μέγεθος, την ηλικία του, την περιοχή στην οποία βρίσκεται και άλλα διάφορα στοιχεία.

- **Σύνθεση και δειγματοληψία (Synthesis and sampling):** Σε αυτόν τον τύπο εργασίας, ζητείται από τον αλγόριθμο μηχανικής μάθησης να δημιουργήσει νέα δείγματα που είναι παρόμοια με αυτά των δεδομένων εκπαίδευσης. Η σύνθεση και η δειγματοληψία μέσω μηχανικής μάθησης μπορεί να είναι χρήσιμες για εφαρμογές πολυμέσων, όπου μπορεί να είναι δαπανηρό ή βαρετό να δημιουργήσουμε μεγάλους όγκους περιεχομένου με το χέρι. Σε ορισμένες περιπτώσεις, θέλουμε η διαδικασία δειγματοληψίας ή σύνθεσης να παράγει κάποιο συγκεκριμένο είδος εξόδου δεδομένης μιας εισόδου. Για παράδειγμα, σε μια εργασία σύνθεσης φωνής, παρέχουμε μια γραπτή πρόταση και ζητάμε από το πρόγραμμα να παράγει μια ακουστική κυματομορφή που περιέχει μια προφορική εκδοχή αυτής της πρότασης. Αυτό είναι ένα είδος δομημένης εργασίας εξόδου, αλλά με δεδομένο ότι δεν υπάρχει μια απόλυτα σωστή έξοδος για κάθε είσοδο και επιθυμούμε ρητά να υπάρχει ποικιλομορφία στην έξοδο, προκειμένου αυτή να φαίνεται πιο φυσική και ρεαλιστική.

Το Μέτρο Απόδοσης (Performance Measure), *P*

Για να αξιολογήσουμε την απόδοση ενός αλγορίθμου μηχανικής μάθησης, πρέπει να σχεδιάσουμε ένα ποσοτικό μέτρο της απόδοσής του. Συνήθως αυτό το μέτρο απόδοσης P είναι ειδικό για την εργασία T που εκτελείται από το σύστημα.

Για εργασίες όπως η ταξινόμηση, συχνά υπολογίζουμε την ακρίβεια (accuracy) του μοντέλου. Η ακρίβεια είναι το ποσοστό των δειγμάτων εισόδου για τα οποία το μοντέλο παράγει τη σωστή έξοδο. Η εκτίμηση της απόδοσης σε αυτή τη περίπτωση μπορεί να γίνει και κάνοντας χρήση του ποσοστού σφάλματος (error rate), το ποσοστό δηλαδή των παραδειγμάτων για τα οποία το μοντέλο παράγει εσφαλμένη έξοδο. Συνήθως ενδιαφερόμαστε για το πόσο καλά αποδίδει ο

αλγόριθμος μηχανικής μάθησης σε δεδομένα που δεν έχει ξαναδεί, αφού αυτό καθορίζει πόσο καλά θα λειτουργήσει στον πραγματικό κόσμο. Επομένως, αξιολογούμε αυτά τα μέτρα απόδοσης χρησιμοποιώντας ένα σύνολο δεδομένων ελέγχου (test set) που είναι ξεχωριστό από τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση του συστήματος μηχανικής εκμάθησης.

Βάση του απλού παραδείγματος της ακρίβειας στην ταξινόμηση μπορεί να φαίνεται πως η επιλογή του μέτρου απόδοσης είναι απλή και αντικειμενική, συχνά όμως είναι πολύ πιο δύσκολο να επιλεχθεί ένα μέτρο απόδοσης που αντιστοιχεί καλά στην επιθυμητή συμπεριφορά του συστήματος. Για παράδειγμα κατά την εκτέλεση μιας εργασίας παλινδρόμησης καλούμαστε να αποφασίσουμε το εξής: θα πρέπει να τιμωρούμε περισσότερο το σύστημα εάν κάνει συχνά μεσαίου μεγέθους λάθη ή αν σπάνια κάνει πολύ μεγάλα λάθη; Σε αυτή την περίπτωση δεν υπάρχει σωστή και λάθος επιλογή , αφού αυτή εξαρτάται καθαρά από την εφαρμογή. Σε άλλες περιπτώσεις, γνωρίζουμε ποια ποσότητα θα θέλαμε να μετρήσουμε ιδανικά, αλλά η μέτρησή της δεν είναι πρακτική. Κάτι τέτοιο θα δούμε πως συμβαίνει και στην παρούσα εργασία (βλ. Κεφάλαιο 4), όπου βρισκόμαστε αντιμέτωποι με ένα υπολογιστικά δυσεπίλυτο ολοκλήρωμα. Ο τρόπος που προσπερνάμε αυτό το πρόβλημα, τόσο στην εργασία αλλά και γενικότερα, είναι να σχεδιάσουμε ένα εναλλακτικό κριτήριο που εξακολουθεί να αντιστοιχεί στους σχεδιαστικούς στόχους ή να βρούμε ικανοποιητική προσέγγιση στο ήδη υπάρχον κριτήριο.

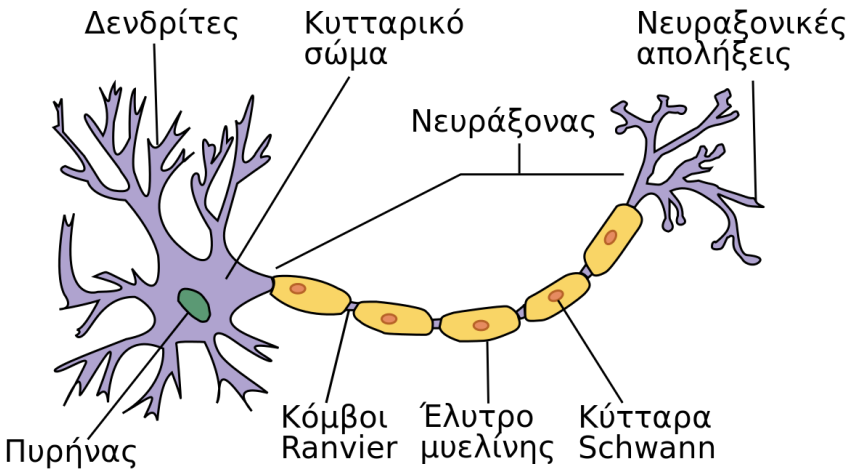
Η Εμπειρία (Experience), *E*

Οι αλγόριθμοι μηχανικής μάθησης μπορούν να ταξινομηθούν σε δύο κατηγορίες ως αλγόριθμοι επιβλεπόμενης και μη επιβλεπόμενης μάθησης. Ο διαχωρισμός αυτός γίνεται με βάση το είδος της εμπειρίας που τους επιτρέπεται να έχουν κατά τη διάρκεια της διαδικασίας εκπαίδευσης.

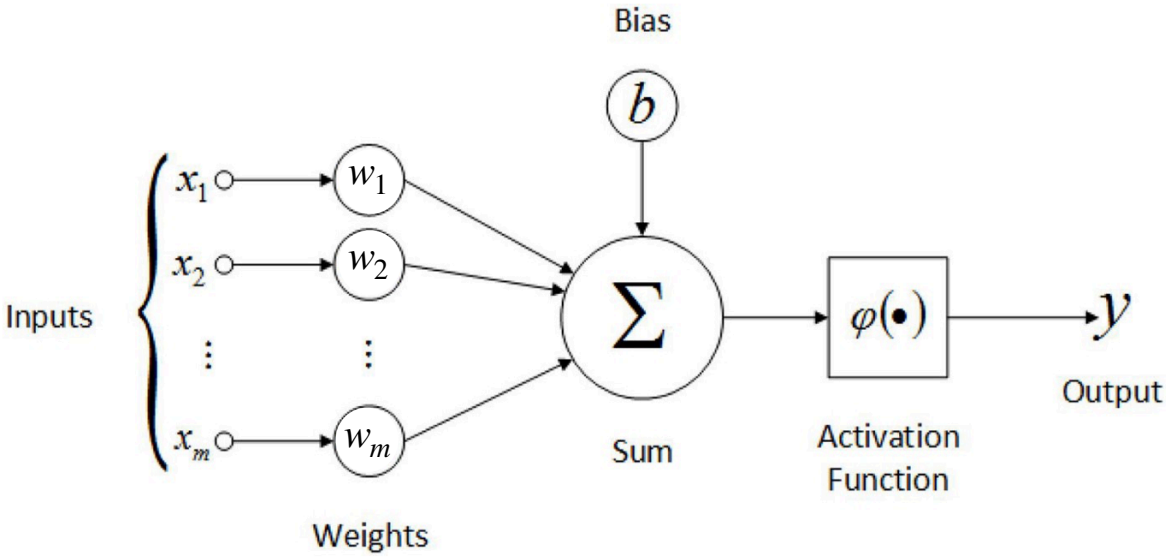
Οι αλγόριθμοι επιβλεπόμενης μάθησης αντιμετωπίζουν ένα σύνολο δεδομένων που περιέχει άγνωστα αρχικά χαρακτηριστικά, αλλά κάθε δείγμα σχετίζεται επίσης με μια ετικέτα ή έναν στόχο, η οποία έχει δοθεί πιθανότατα από κάποιον άνθρωπο. Η εμπειρία που αποκτάται με αυτό τον τρόπο δημιουργεί άμεσες συσχετίσεις μεταξύ των χαρακτηριστικών ενός δείγματος εισόδου και της αντίστοιχης ετικέτας που το συνοδεύει.

Οι αλγόριθμοι μη επιβλεπόμενης μάθησης, από την άλλη πλευρά, αντιμετωπίζουν σύνολα δεδομένων με ποικίλα χαρακτηριστικά, χωρίς να διαθέτουν κάποια ετικέτα, αλλά αποκτούν εμπειρία με το να εξάγουν και να μαθαίνουν χρήσιμες ιδιότητες της δομής αυτού του συνόλου δεδομένων. Τα παραγωγικά μοντέλα που θα μελετήσουμε στη συνέχεια της εργασίας προσπαθούν να αποκτήσουν εμπειρία μέσω αυτής της τακτικής.

Η εφαρμογή της μηχανικής μάθησης στην πράξη γίνεται με τη χρήση πολλών διαφορετικών μοντέλων, τα οποία έχουν αναπτυχθεί με σκοπό να βελτιστοποιήσουν τον τρόπο που μια μηχανή μαθαίνει ως προς μια συγκεκριμένη εργασία. Δηλαδή, βάση του ορισμού, μπορούμε να πούμε πως έχουν αναπτυχθεί διαφορετικά μοντέλα που το καθένα αποσκοπεί στην αύξηση του μέτρου απόδοσης P για μια συγκριμένη εργασία (ή ένα σύνολο εργασιών) T. Μερικά από τα πιο γνωστά μοντέλα είναι τα δέντρα απόφασης (decision trees), οι μηχανές διανυσμάτων υποστήριξης (support vector machines), η ανάλυση παλινδρόμησης (regression analysis), τα Μπεϋζιανά δίκτυα, οι γενετικοί αλγόριθμοι και φυσικά τα νευρωνικά δίκτυα. Όλα τα προαναφερθείσα δίκτυα παρουσιάζουν εξαιρετικό ενδιαφέρον τόσο ως προς τη θεωρία και τη δόμηση, όσο και προς τα αποτελέσματα που παράγουν στα προβλήματα που καλούνται να λύσουν. Για τις ανάγκες της παρούσας εργασίας κρίνεται σκόπιμο να γίνει μια εμβάθυνση στα τεχνητά νευρωνικά δίκτυα (ANNs), στην αμέσως επόμενη ενότητα, καθώς αυτά είναι που θα μας απασχολήσουν και στα επόμενα κεφάλαια.



(α)



(β)

Σχήμα 2.2: (α) Η δομή ενός πραγματικού νευρώνα. (β) Η δομή ενός νευρώνα Perceptron.

Τέλος, πρώτου κλείσουμε αυτή την ενότητα θα απαντήσουμε στην εύλογη απορία που μπορεί να γεννηθεί σε κάποιον αναγνώστη και αφορά το γιατί να έχουμε ανάγκη από τόσα διαφορετικά μοντέλα για την υλοποίηση της μηχανικής μάθησης. Η απάντηση είναι πολύ απλή και έχει να κάνει με το γεγονός πως διαφορετικά προβλήματα κρίνουν διαφορετικής αντιμετώπισης. Δεν υπάρχει μοντέλο που να μπορεί να χαρακτηριστεί καθολικά το καλύτερο, αλλά μόνο το καλύτερο για κάποια συγκεκριμένη εργασία. Το γνωστό θεώρημα του μη δωρεάν γεύματος (no free lunch theorem) απαντάει ακριβώς σε αυτή την ερώτηση και αξίζει να τονιστεί πως η αναζήτηση του βέλτιστου μοντέλου, για κάποια συγκεκριμένη εργασία, αποτελεί μια από τις μεγαλύτερες προκλήσεις για τους μηχανικούς μηχανικής μάθησης.

2.2 Νευρωνικά δίκτυα

Ο όρος νευρωνικά δίκτυα, παραδοσιακά, αναφέρεται στα δίκτυα νευρώνων στον εγκέφαλο των θηλαστικών. Οι νευρώνες (Σχ. 2.2) αποτελούν τις θεμελιώδεις μονάδες πραγματοποίησης υπολογισμών του εγκεφάλου. Οι συνδέσεις, γνωστές και ως συνάψεις, που αναπτύσσουν μεταξύ τους σχηματίζουν δίκτυα, τα οποία είναι υπεύθυνα για την επεξεργασία πιο σύνθετων προβλημάτων και δεδομένων. Ο τρόπος σχηματισμού των δικτύων αυτών αποτελεί ένα σύνθετο και περίπλοκο πρόβλημα το οποίο έχει άμεση σχέση με τα ερεθίσματα που δέχονται οι νευρώνες. Μάλιστα ο κάθε νευρώνας μπορεί να αποκρίνεται με διαφορετικό τρόπο και σε διαφορετικές εισόδους, παράγοντας επίσης διαφορετικές εξόδους.

Ακριβώς αυτή η περιγραφή της δομής ενός πραγματικού εγκέφαλου, ήταν και η έμπνευση για την κατασκευή των τεχνητών νευρωνικών δικτύων. Φυσικά με τους υπολογιστικούς πόρους που διαθέτουμε είναι ακόμη αδύνατο να προσομοιώσουμε τους 86 δισεκατομμύρια νευρώνες που διαθέτει κατά μέσο όρο το ανθρώπινο μυαλό. Έτσι δημιουργήσαμε το μοντέλο του τεχνητού νευρώνα (Σχ. 2.2β), το οποίο παρά το γεγονός ότι είναι εξαιρετικά απλοποιημένο, είναι ικανό να προσομοιώσει τη βασική λειτουργία ενός πραγματικού νευρώνα σε ότι αφορά ένα ερέθισμα (είσοδος) και μια απόκριση (έξοδος). Παρά το γεγονός ότι τα τεχνητά νευρωνικά δίκτυα είναι αρκετά απλοποιημένα, σε σχέση με τον ανθρώπινο εγκέφαλο, ήδη καταφέρνουν να είναι το ίδιο αποδοτικά και πολλές φορές αποδοτικότερα από τον άνθρωπο στην εκτέλεση κάποιας εργασίας. Μερικά τέτοια παραδείγματα είναι η αναγνώριση εικόνας, όπου τα αποτελέσματα των νευρωνικών δικτύων είναι συγκρίσιμα με αυτά των ανθρώπων καθώς και η αυτόματη μετάφραση κειμένου, οπού τα νευρωνικά δίκτυα υπερτερούν.

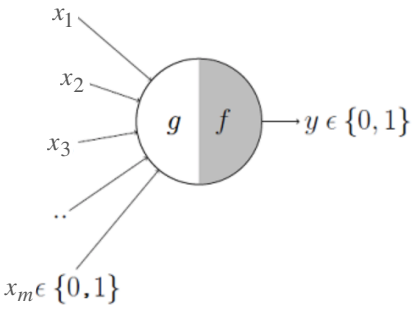
Στη συνέχεια της ενότητας παρουσιάζουμε την δομή και την λειτουργία του τεχνητού νευρώνα, του νευρώνα perceptron, το δίκτυο πρόσθιας τροφοδότησης (feed forward network), τα συνελικτικά νευρωνικά δίκτυα και τους autoencoders. Όλες αυτές οι έννοιες είναι απαραίτητες προκειμένου να προχωρήσουμε στην επόμενη ενότητα και τα παραγωγικά μοντέλα.

2.2.1 Τεχνητός νευρώνας

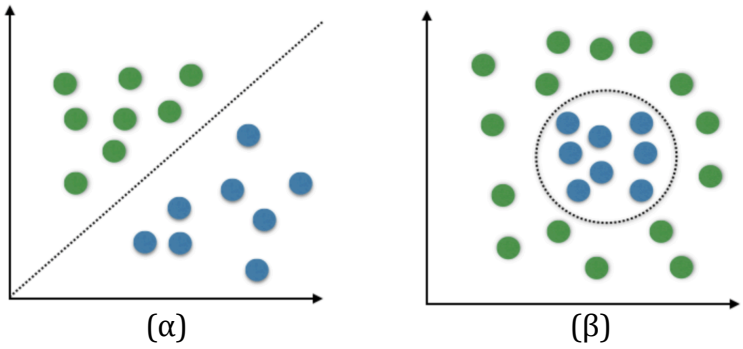
Όπως αναφέραμε, ο τεχνητός νευρώνας αποτελεί μια προσέγγιση ενός πραγματικού νευρώνα, απλοποιημένου σε σημείο που να είναι υλοποιήσιμος από τα υπολογιστικά συστήματα που διαθέτουμε. Το 1943 οι Warren McCulloch και Walter Pitts πρότειναν το πρώτο μοντέλο τεχνητού νευρώνα (Threshold Logic Unit) βασισμένοι στους βιολογικούς νευρώνες (Σχ. 2.3). Στο τεχνητό νευρώνα, η πληροφορία ρέει πάντα προς μια κατεύθυνση, από αριστερά προς τα δέξια. Αυτό σημαίνει πως σε καμία φάση της λειτουργίας του νευρώνα η πληροφορία δεν γυρνάει προς τα

πίσω μέσω βρόχων ανάδρασης. Έχοντας ξεκαθαρίσει αυτόν τον περιορισμό προχωράμε στην περιγραφή της λειτουργίας του νευρώνα (Warren McCulloch, 1943)(Walter Pitts, 1943).

Αρχικά δέχεται εισόδους $x_1, x_2, \ldots, x_m$ όπου $x_m = 0$ ή $x_m = 1$. Στη συνέχεια αθροίζει τις εισόδους (συνάρτηση g) και ελέγχει αν το άθροισμα ξεπερνά ένα κατώφλι (συνάρτηση f). Αν το ξεπερνάει ο νευρώνας ενεργοποιείται και βγάζει στην έξοδο την τιμή 1 αλλιώς μένει ανενεργός, δηλαδή παράγει την τιμή 0. Ο νευρώνας δουλεύει μόνο με γραμμικά διαχώρισμα δεδομένα, δηλαδή με δεδομένα που μπορούν να διαχωριστούν τέλεια με τη χρήση ενός υπερεπιπεδού που χωρίζει το χώρο σε δυο μέρη. Στο χώρο των δύο διαστάσεων αυτό σημαίνει πως υπάρχει μια ευθεία γραμμή που μπορεί να διαχωρίσει τα δεδομένα έτσι ώστε όλα τα δεδομένα που ανήκουν στην ίδια ομάδα να είναι από την ίδια μεριά της ευθείας (Σχ. 2.4).



Σχήμα 2.3: Νευρώνας McCulloch-Pitts.



Σχήμα 2.4: (α) Γραμμικά διαχώρισμα δεδομένα. (β) Μη γραμμικά διαχώρισμα δεδομένα.

2.2.2 Ο νευρώνας Perceptron

Ο νευρώνας perceptron είναι μια εξέλιξη του νευρώνα των McCulloch και Pitts και προτάθηκε από τον Frank Rosenblatt το 1957(Σχ. 2.2β). Η διαφορά με τον απλό τεχνητό νευρώνα είναι ότι ο νευρώνας perceptron μπορεί να δεχτεί είσοδο οποιαδήποτε τιμή μεταξύ 0 και 1. Ακολουθεί η περιγραφή της λειτουργίας του perceptron (Frank Rosenblatt, 1957).

Υποθέτουμε $x_1, x_2, \ldots, x_m$ εισόδους στον νευρώνα. Αρχικά οι είσοδοι αυτοί πολλαπλασιάζονται με βάρη $w_1, w_2, \ldots, w_m$. Ακόμη έχουμε έναν επιπλέον όρο, ο οποίος ονομάζεται πόλωση ή κατώφλι ή μεροληψία (bias, threshold) και παίζει το ρόλο ενός εξωτερικού παράγοντα. Το συνολικό

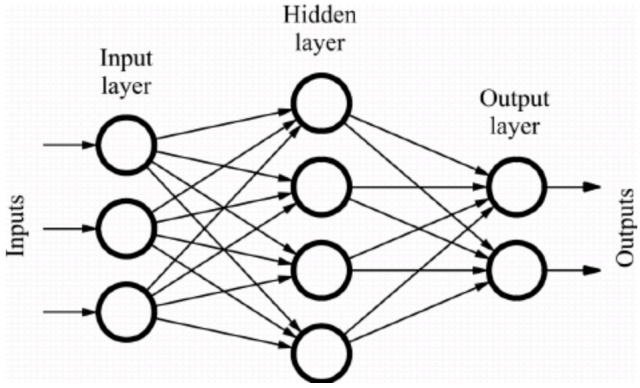
άθροισμα των σταθμισμένων εισόδων με την πόλωση μας δίνει την τελική είσοδο στο νευρώνα, η οποία ονομάζεται τοπικό πεδίο. Στη συνέχεια μια συνάρτηση, γνωστή ως συνάρτηση ενεργοποίησης (activation function) ή συνάρτηση μεταφοράς (squashing function) εφαρμόζεται στο τοπικό πεδίο και έτσι παράγεται η τελική έξοδος του νευρώνα. Όπως αναφέραμε αυτή η συνάρτηση δεν παράγει μόνο διάδικες τιμές, αλλά και κάθε πιθανή τιμή μεταξύ αυτών.

Ο νευρώνας perceptron, όπως και ο απλός τεχνητός νευρώνας μπορεί να δουλέψει μόνο σε γραμμικά διαχειρίσιμες κλάσεις. Παρόλα αυτά είναι αρκετά πιο ευέλικτος από τον απλό νευρώνα χάρη της προσθήκης των βαρών για την στάθμιση της εισόδου. Ο νευρώνας perceptron είναι η βασική δομική μονάδα των νευρωνικών δικτύων που παρουσιάζονται στην επόμενη ενότητα.

2.2.3 Δίκτυα Πρόσθιας Τροφοδότησης

Το νευρωνικό δίκτυο πρόσθιας τροφοδότησης (feedforward neural network) (Σχ. 2.5) αποτελεί τον πρώτο χρονικά και πιο απλό τύπο νευρωνικού δικτύου. Πρόκειται για ένα δίκτυο που όπως μαρτυρά η ονομασία του τροφοδοτεί την πληροφορία μόνο προς μια κατεύθυνση, από την είσοδο προς την έξοδο.

Ο χαρακτηρισμός των δομών αυτών ως δίκτυα δικαιολογείται από το γεγονός ότι η αναπαριστώνται συνήθως μέσω της σύνθεσης διαφόρων συναρτήσεων. Ένα δίκτυο πρόσθιας τροφοδότησης θα μπορούσε να παρομοιαστεί με έναν κατευθυνόμενο ακυκλικό γράφο, που περιγράφει τον τρόπο με τον οποίο διαφορετικές συναρτήσεις προσεγγίζονται μαζί. Για παράδειγμα, μπορούμε να έχουμε τρεις συναρτήσεις $f^{(1)}$, $f^{(2)}$ και $f^{(3)}$ συνδεδεμένες αλυσιδωτά, ώστε να παραχθεί η συνάρτηση $f^{(3)}$ ($f^{(2)}$ ($f^{(1)}$ (x)). Κοιτώντας τώρα από την οπτική του δικτύου το $f^{(1)}$ είναι το πρώτο στρώμα (layer) του δικτύου, το $f^{(2)}$ το δεύτερο στρώμα κτλ.. Όσο αυτή η αλυσίδα μεγαλώνει με την προσθήκη κι άλλων συναρτήσεων, τόσο μεγαλώνει και το βάθος του δικτύου. Τα δεδομένα που δίνονται στην είσοδο του δικτύου αυτού και ονομάζονται δεδομένα εκπαίδευσης, είναι υπεύθυνα μόνο για τη συμπεριφορά του στρώματος εξόδου (output layer). Η συμπεριφορά των ενδιάμεσων στρωμάτων δεν καθορίζεται με κανέναν τρόπο από τα δεδομένα εκπαίδευσης, αλλά από τον αλγόριθμο μάθησης, για αυτό αυτά τα ενδιάμεσα στρώματα είναι γνωστά και ως κρυφά στρώματα (hidden layers). Ο χαρακτηρισμός νευρωνικά, όπως έχουμε ήδη αναφέρει, αφορά το γεγονός ότι σε πολλά σημεία προσπαθεί να αντλήσει πληροφορίες από την νευροεπιστήμη (Ian Goodfellow, 2016).



Σχήμα 2.5: Δίκτυο πρόσθιας τροφοδότησης με ένα κρυφό στρώμα.

Καίρια σημασία έχει η επιλογή των συναρτήσεων ενεργοποίησης των κρυφών στρωμάτων του δικτύου, καθώς αυτή θα καθορίσει σε μεγάλο βαθμό και τα αποτελέσματα που θα παράγει. Άλλοι παράγοντες που επηρεάζουν τη λειτουργία ενός τέτοιου δικτύου είναι ο αριθμός των κρυφών στρωμάτων, ο αριθμός των νευρώνων που απαρτίζουν κάθε ένα από τα κρυφά στρώματα, όπως επίσης και οι συνδέσεις που θα υλοποιηθούν από ένα στρώμα στο επόμενο.

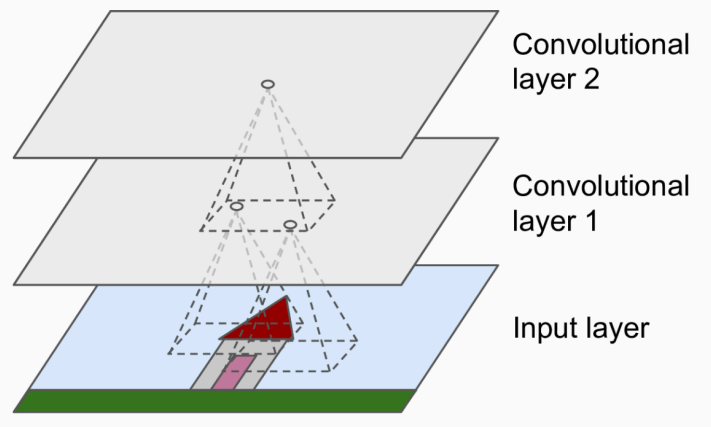
Πριν κλείσουμε αυτή την ενότητα θα αναφερθούμε στην έννοια των αλγορίθμων οπισθοδιάδοσης (backpropagation). Οι αλγόριθμοι αυτοί αποτελούν τον τρόπο με τον οποίο υλοποιείται η διαδικασία της μάθησης και το δίκτυο μαθαίνει μέσα από τα λάθη του. Η ονομασία τους δείχνει ξεκάθαρα πως αφορούν κάποια διαδρομή προς τα πίσω μέσα στο δίκτυο. Συγκεκριμένα πρόκειται για μια αναδρομική διαδικασία, που στοχεύει στην αλλαγή των βαρών του δικτύου προκείμενου να βελτιωθεί η απόδοση του, κάτι που συμβαίνει προς την αντίθετη κατεύθυνση από αυτή που διαδίδεται η πληροφορία. Η αλλαγή των βαρών γίνεται μέσω του υπολογισμού των κλίσεων κάνοντας χρήση του κανόνα της αλυσίδας (gradient-based optimization).

2.2.4 Συνελικτικά νευρωνικά δίκτυα

Τα συνελικτικά νευρωνικά δίκτυα (convolutional neural networks - CNN) αποτελούν μια κατηγορία νευρωνικών δικτύων που προέκυψαν από την απόπειρα αντιγραφής της λειτουργίας του οφθαλμού. Πιο συγκεκριμένα η ανάπτυξη των πρώτων συνελικτικών νευρωνικών δικτύων βασίστηκε στην εργασία των Hubel και Wiesel, οι οποίοι το 1968 δημοσίευσαν την δουλειά τους πάνω στη μελέτη του οπτικού φλοιού θηλαστικών, όπως οι γάτες και οι μαϊμούδες.

Από τη φύση τους τα συνελικτικά νευρωνικά δίκτυα φάνηκαν τα καταλληλότερα για να χρησιμοποιηθούν στην Όραση Υπολογιστών. Οι δυνατότητες τους όμως δεν σταματούν εκεί αφού, όπως θα δούμε στην συνέχεια της εργασίας είναι κατάλληλα και για την επεξεργασία και αναγνώριση μουσικής και γενικά ήχου. Ο χαρακτηρισμός συνελικτικά προκύπτει από τον τύπο των κρυφών επιπέδων από τα οποία αποτελούνται. Τα κρυφά στρώματα που συναντά κανείς σε ένα συνελικτικό νευρωνικό δίκτυο είναι τα ακόλουθα:

- Συνελικτικά στρώματα (convolutional layers),
- Πλήρως συνδεδεμένα στρώματα (fully connected layers),
- Στρώματα κανονικοποίησης (normalization layers)
- Στρώματα συγκέντρωσης (pooling layers)



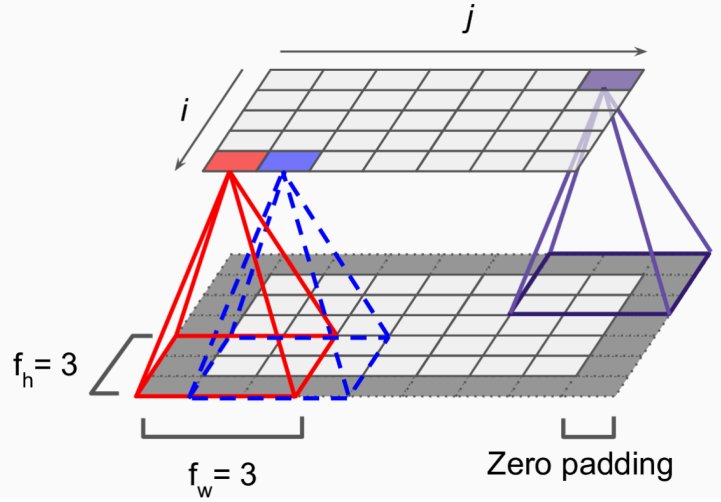
Σχήμα 2.6: Συνελικτικά στρώματα με ορθογώνια δεκτικά πεδία.

Στο τέλος ενός συνελικτικού στρώματος είναι σύνηθες να εφαρμόζεται κάποια μη γραμμική συνάρτηση ενεργοποίησης (activation function). Μερικές τέτοιες συναρτήσεις είναι η σιγμοειδής, η υπερβολική εφαπτομένη, οι συναρτήσεις ReLU και SeLU. Προκείμενου να γίνει κατανοητή η χρησιμότητα της συνάρτησης ενεργοποίησης στα συνελικτικά δίκτυα, πρέπει πρώτα να κατανοήσουμε την αρχή λειτουργίας τους. Όπως θα δούμε και στη συνέχεια, τα συνελικτικά δίκτυα έχουν σχεδιαστεί με σκοπό τη λήψη ορισμένων χαρακτηριστικών από εικόνες, που ξεκινούν με χαρακτηριστικά χαμηλού επιπέδου στο αρχικό στρώμα και πολύ υψηλότερου επιπέδου όσο αυξάνονται τα στρώματα. Με το χαρακτηρισμό υψηλού επιπέδου εννοούνται τα χαρακτηριστικά τα οποία περιέχουν γενικεύσεις ως προς μια συγκεκριμένη κατηγορία δεδομένων. Αυτή η γενίκευση δεν θα ήταν δυνατή με τη χρήση γραμμικών συναρτήσεων και άρα η αφαιρετική ικανότητα και η ισχύς των δικτύων θα περιοριζόταν σε χαρακτηριστικά πολύ χαμηλού επιπέδου. Αυτός είναι και ο λόγος, για τον οποίο η χρήση μη γραμμικών συναρτήσεων ενεργοποίησης είναι καθοριστική για την απόδοση των συνελικτικών νευρωνικών δικτύων.

Η έξοδος ενός συνελικτικού δίκτυο μπορεί να έχει διαφορετική μορφή, μέγεθος και σημασία, ανάλογα με το ποιος είναι ο σκοπός του δικτύου. Για παράδειγμα στην ταξινόμηση, η πολυδιάστατη είσοδος αντιστοιχίζεται σε μια κατανομή πιθανότητας (με πιθανές τιμές όλες τις πιθανές κλάσεις ταξινόμησης), η οποία υλοποιείται μέσω της συνάρτησης SoftMax που ακολουθεί ύστερα από ένα ή περισσότερα πλήρως συνδεδεμένα στρώματα. Στη συνέχεια παρουσιάζονται περιληπτικά τα πιο βασικά συστατικά στοιχεία ενός συνελικτικού νευρωνικού δικτύου.

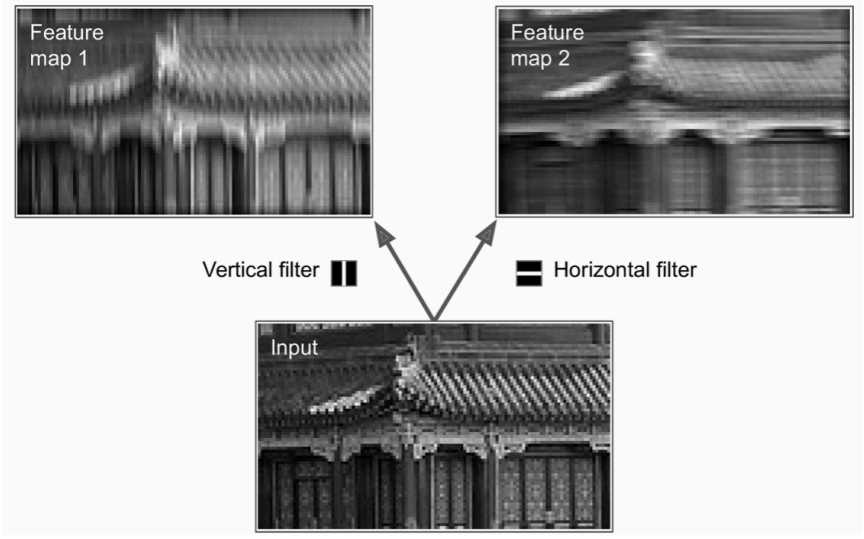
- **Συνελικτικό στρώμα (Convolutional layer):** Το πιο σημαντικό δομικό στοιχείο ενός συνελικτικού νευρωνικού δικτύου είναι το συνελικτικό στρώμα. Οι νευρώνες στο πρώτο συνελικτικό στρώμα δεν συνδέονται με κάθε εικονοστοιχείο στην εικόνα εισόδου (όπως συμβαίνει στα πλήρως συνδεδεμένα δίκτυα), αλλά μόνο σε εικονοστοιχεία που βρίσκονται μέσα στα δεκτικά πεδία (receptive fields) (Σχ. 2.6). Με τη σειρά του, κάθε νευρώνας στο δεύτερο συνελικτικό στρώμα συνδέεται μόνο με νευρώνες που βρίσκονται μέσα σε ένα μικρό ορθογώνιο στο πρώτο στρώμα. Αυτή η αρχιτεκτονική επιτρέπει στο δίκτυο να επικεντρωθεί σε χαρακτηριστικά χαμηλού επιπέδου στο πρώτο κρυφό επίπεδο και, στη συνέχεια, να εξάγει χαρακτηριστικά υψηλότερου επιπέδου στο επόμενο κρυφό επίπεδο και ούτω καθεξής. Αυτή η ιεραρχική δομή είναι κοινή σε εικόνες πραγματικού κόσμου και αποτελεί έναν από τους λόγους για τους οποίους τα συνελικτικά νευρωνικά δίκτυα λειτουργούν τόσο καλά στην αναγνώριση εικόνας. Ένας νευρώνας που βρίσκεται στη σειρά i και στήλη j ενός δεδομένου στρώματος, συνδέεται με τις εξόδους των νευρώνων του προηγούμενου στρώματος που βρίσκονται στις σειρές i έως $i + f_h - 1$ και στήλες από j έως $j + f_w - 1$, όπου f_h και f_w είναι το ύψος και το πλάτος του δεκτικού πεδίου. Προκειμένου ένα επίπεδο να έχει το ίδιο ύψος και πλάτος με το προηγούμενο επίπεδο, είναι σύνηθες να προσθέτετε μηδενικά γύρω από τις εισόδους, μια πρακτική γνωστή και ως zero padding (Σχ. 2.7).

- **Φίλτρα/Πυρήνες (Filters/Convolution Kernels):** Τα βάρη ενός νευρώνα μπορούν να αναπαρασταθούν σαν ένας πίνακας στο μέγεθος του δεκτικού πεδίου. Αυτά τα βάρη αναφέρονται συχνά και ως οι παράμετροι προς εκμάθηση του δικτύου (learnable parameters). Ο πίνακας των βαρών ονομάζεται συνήθως πυρήνας (kernel) αλλά συναντάται και με την ονομασία φίλτρο (filter). Κατά την εκπαίδευση, ο πυρήνας μετατοπίζεται κατά μήκος του ύψους και του πλάτους της εικόνας και σε κάθε νέα θέση, κάνοντας χρήση της πράξης της συνέλιξης, υπολογίζεται το γινόμενο μεταξύ του πυρήνα και των τιμών που βλέπει το δεκτικό πεδίο στην εικόνα. Μέσα από την διαδικασία αυτή προκύπτει μια νέα αναπαράσταση της εικόνας, γνωστή και ως χάρτης χαρακτηριστικών (feature map). Όπως φαίνεται και στην Σχήμα 2.8 οι χάρτες χαρακτηριστικών εκφράζουν την επίδραση του πυρήνα σε διαφορετικά



Σχήμα 2.7: Δεκτικά πεδία και zero padding.

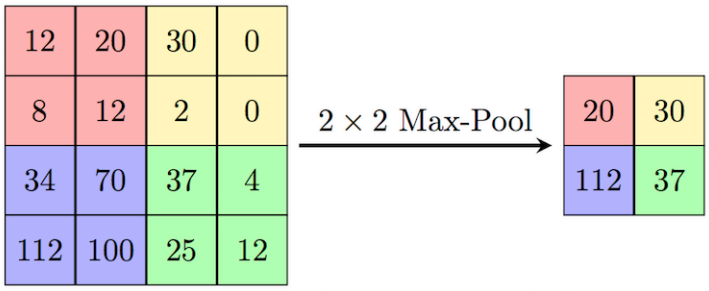
μέρη της εικόνας. Η απόσταση σε κάθε διαδοχική μετακίνηση του πυρήνα ονομάζεται βήμα (stride) και γενικά δεν υπάρχει μια επιλογή του που να ικανοποιεί όλα τα πιθανά προβλήματα. Συνήθως το βήμα είναι μικρότερο από το μέγεθος του πυρήνα και άρα αυτό σημαίνει πως έχουμε επικαλύψεις. Τέλος αναφέρουμε πως στα σύγχρονα προβλήματα, οι εικόνες που δίνονται στην είσοδο του δικτύου είναι έγχρωμες και άρα αποτελούνται από 3 κανάλια: κόκκινο (R), πράσινο (G), μπλε (B). Βάση αυτού και οι πυρήνες που χρησιμοποιούνται θα πρέπει να έχουν και αυτοί βάθος 3.



Σχήμα 2.8: Χρήση δύο διαφορετικών φίλτρων για παραγωγή δυο διαφορετικών χαρτών χαρακτηριστικών (feature maps).

- **Στρώμα Συγκέντρωσης (Pooling Layer):** Ο στόχος των στρωμάτων συγκέντρωσης (pooling layers) είναι να πραγματοποιήσουν υποδειγματοληψία (δηλαδή να συρρικνώσουν) την εικόνα εισόδου προκειμένου να μειωθεί το υπολογιστικό φορτίο, η χρήση της μνήμης και ο αριθμός των παραμέτρων. Ακριβώς όπως στα συνελικτικά στρώματα, κάθε νευρώνας σε ένα στρώμα συγκέντρωσης συνδέεται με τις εξόδους ενός περιορισμένου αριθμού νευρώνων στο προηγούμενο στρώμα, που βρίσκεται μέσα σε ένα μικρό ορθογώνιο δεκτικό πεδίο. Το μέγεθός του, το βήμα και ο τύπος του στρώματος συγκέντρωσης αποτελούν τις υπερπαραμέτρους του

και πρέπει να οριστούν πριν την εφαρμογή του. Κάποιοι τύποι στρωμάτων συγκέντρωσης που



Σχήμα 2.9: Στρώμα μέγιστης συγκέντρωσης (max pooling layer), μεγέθους 2×2 με βήμα 2.

συναντώνται συχνά είναι ο μέσος όρος, σταθμισμένος ή μη (average pooling), η ευκλείδεια νόρμα και η συγκέντρωση μεγίστου (max pooling), η οποίο αποτελεί και την συνηθέστερη επιλογή. Σημειώνουμε πως ένας νευρώνας συγκέντρωσης δεν έχει βάρη (Σχ. 2.9).

2.2.5 Autoencoder

Στην ενότητα αυτή παρουσιάζουμε το τελευταίο θεμελιώδες δίκτυο, στο πλαίσιο δημιουργίας της απαραίτητης θεωρητικής βάσης για την παρούσα εργασία, τους autoencoders. Στη συνέχεια τόσο του κεφαλαίου όσο και της εργασίας, θα αναφερόμαστε σε αυτούς με την αγγλική ονομασία autoencoders, διότι η αντίστοιχη ελληνική ονομασία, αυτοκωδικοποιητές, δεν συνηθίζεται. Οι autoencoders είναι ένα νευρωνικό δίκτυο που έχει ως βασικό σκοπό να αντιγράψει την είσοδο που του δίνεται, στην έξοδο του. Στο εσωτερικό, υπάρχει ένα κρυφό επίπεδο h, το οποίο είναι υπεύθυνο για την απεικόνιση της εισόδου σε έναν χώρο μικρότερης διάστασης και περιγράφει μια συμπιεσμένη και κωδικοποιημένη αναπαράσταση της εισόδου. Ένα τέτοιο δίκτυο, έστω με είσοδο x και έξοδο r, μπορεί να θεωρηθεί ότι αποτελείται από δύο μέρη:

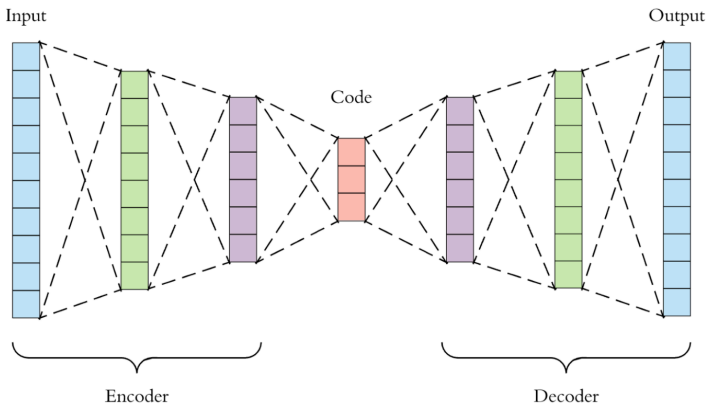
1. Μια συνάρτηση κωδικοποιητή $h = f(x)$.
2. Έναν αποκωδικοποιητή, ο οποίος παράγει την ανακατασκευασμένη είσοδο, $r = g(h)$.

Σκοπός του δικτύου είναι να καταφέρει να εξάγει σημαντικά χαρακτηριστικά της εισόδου και να τα αποθηκεύει στην ενδιάμεση κωδικοποίηση (latent space/code). Άρα δεν πρέπει να θεωρείται ως ένας απλός αλγόριθμος συμπίεσης δεδομένων. Το δίκτυο του κωδικοποιητή μπορεί να είναι ένα πλήρως συνδεδεμένο δίκτυο, ένα συνελικτικό δίκτυο ή ένα οποιοδήποτε άλλο νευρωνικό δίκτυο που είναι συμβατό με τα δεδομένα εισόδου. Ο αποκωδικοποιητής είναι σχεδόν πάντα ένα ίδιο δίκτυο με αυτό της εισόδου, αλλά με τα στρώματα ανεστραμμένα (Σχ. 2.10). Εάν ένας autoencoder καταφέρει να μάθει να ορίζει τέλεια $g(f(x)) = x, \forall x$, τότε δεν πετυχαίνει το σκοπό του, αφού μάλλον έχει καταφέρει να συμπιέζει την είσοδο χωρίς όμως να εξάγει χρήσιμα χαρακτηριστικά από τα δεδομένα εισόδου.

Διαισθητικά, καταλαβαίνουμε πως η συνολική αρχιτεκτονική ενός autoencoder δημιουργεί ένα σημείο συμφόρησης (bottleneck), το οποίο διασφαλίζει ότι, μόνο το κύριο δομημένο μέρος των πληροφοριών μπορεί να περάσει και να ανακατασκευαστεί. Κοιτάζοντας το γενικότερο πλαίσιο, η οικογένεια E των εξεταζόμενων κωδικοποιητών ορίζεται από την αρχιτεκτονική δικτύου του κωδικοποιητή, η οικογένεια D των εξεταζόμενων αποκωδικοποιητών ορίζεται από την

αρχιτεκτονική του δικτύου του αποκωδικοποιητή και η αναζήτηση του ζεύγους κωδικοποιητή και αποκωδικοποιητή που ελαχιστοποιεί το σφάλμα ανακατασκευής γίνεται με με αλγορίθμους σύγκλισης ελάττωσης της παραγώγου πάνω από τις παραμέτρους αυτών των δικτύων.

Συνήθως, τα δίκτυα των autoencoders, περιορίζονται στο να ανακατασκευάζουν ένα συγκεκριμένο σύνολο δεδομένων, το οποίο απαρτίζεται από συγκεκριμένα διακριτά χαρακτηριστικά. Ανάλογα με τη χρήση τους υπάρχουν πολλά και διαφορετικά είδη autoencoders, όπως είναι οι autoencoders αποθορυβοποίησης και κανονικοποίησης. Όπως θα δούμε και στη συνέχεια του κεφαλαίου, μια σύγχρονη και πολύ σημαντική παραλλαγή των autoencoders είναι οι variational autoencoders που χαρακτηρίζονται ως παραγωγικά μοντέλα, αφού μπορούν πέρα από απλή ανακατασκευή της εισόδου, να παράγουν και εντελώς νέα δεδομένα, από την κατανομή των δεδομένων εισόδου. Τα παραγωγικά μοντέλα και οι variational autoencoders παρουσιάζονται στην επόμενη ενότητα.



Σχήμα 2.10: Autoencoder.

2.3 Παραγωγικά μοντέλα (Generative models)

Στη παρούσα ενότητα γίνεται η απαραίτητη εισαγωγή στα μοντέλα που είναι ικανά να παράγουν νέα δεδομένα, γνωστά και ως παραγωγικά μοντέλα (generative models). Γενικά ένα παραγωγικό (ή γενετικό) μοντέλο μπορεί να οριστεί ως εξής :

> *«Ένα παραγωγικό μοντέλο περιγράφει τον τρόπο δημιουργίας ενός συνόλου δεδομένων, από την οπτική ενός πιθανοτικού μοντέλου. Με δειγματοληψία από αυτό το μοντέλο, καθίσταται δυνατή η παραγωγή νέων δεδομένων.»*

Ένα τέτοιο μοντέλο οφείλει να έχει πιθανοτικό και όχι ντετερμινιστικό χαρακτήρα. Τα μοντέλα που έχουμε παρουσιάσει μέχρι τώρα αποτελούν ντετερμινιστικά μοντέλα, αφού για την ίδια είσοδο παράγουν πάντοτε την ίδια έξοδο. Στη γενική περίπτωση ο τρόπος με τον οποίο φτιάχνουμε μοντέλα, τα οποία είναι σε θέση να παράγουν νέα δεδομένα περιλαμβάνει την προσέγγιση κάποια άγνωστης κατανομής πιθανότητας ενός γνωστού συνόλου δεδομένων. Εφόσον η προσέγγιση αυτής της κατανομής πιθανότητας είναι καλή, μπορούμε ύστερα με απλή δειγματοληψία να πάρουμε νέα δείγματα που θα μπορούσαν να ανήκουν στο αρχικό σύνολο δεδομένων, όντας αληθοφανή.

Η παρουσίαση των παραγωγικών μοντέλων εμπεριέχει, όπως ήδη αναφέρθηκε, πολλούς ορισμούς και έννοιες από την θεωρία πιθανοτήτων και τη στατιστική. Επειδή η παράθεση και η

εμβάθυνση σε αυτές τις έννοιες ξεφεύγει από το βασικό σκοπό της εργασίας, αυτές παρουσιάζονται περιληπτικά, προκειμένου να είναι ξεκάθαρη και πλήρης η παρουσίαση του variational autoencoder και η εξαγωγή της συνάρτησης κόστους αυτού. Στην αμέσως επόμενη ενότητα παρουσιάζουμε τις απαραίτητες έννοιες από την θεωρία πιθανοτήτων.

2.3.1 Στοιχεία πιθανοτήτων

Τυχαία μεταβλητή

Μια τυχαία μεταβλητή (τ.μ.), είναι μια μεταβλητή ικανή να περιγράψει το αποτέλεσμα ενός τυχαίου πειράματος. Συνήθως δηλώνουμε την ίδια την τυχαία μεταβλητή με ένα πεζό γράμμα με απλή γραμματοσειρά και τις τιμές που μπορεί να πάρει με πεζά γράμματα και πλάγια γραμματοσειρά. Για παράδειγμα, x_1 και x_2 είναι και οι δύο πιθανές τιμές που μπορεί να πάρει η τυχαία μεταβλητή x. Για διανυσματικές μεταβλητές, γράφουμε την τυχαία μεταβλητή ως $\mathbf{x}$ και μία από τις τιμές της ως $\boldsymbol{x}$. Από μόνη της, μια τυχαία μεταβλητή είναι απλώς μια περιγραφή των καταστάσεων που είναι δυνατές. Πρέπει να συνδυαστεί με μια κατανομή πιθανότητας που καθορίζει πόσο πιθανή είναι κάθε μία από αυτές τις καταστάσεις.

Μια τυχαία μεταβλητή μπορεί να πάρει ένα σύνολο δυνατών τιμών, σε κάθε μία από τις οποίες αντιστοιχεί μια πιθανότητα, αν είναι διακριτή, ή μια πυκνότητα πιθανότητας αν είναι συνεχής. Από μόνη της, μια τυχαία μεταβλητή είναι απλώς μια περιγραφή των καταστάσεων που είναι δυνατές. Για να έχει φυσικό νόημα πρέπει να συνδυαστεί με μια κατανομή πιθανότητας που καθορίζει πόσο πιθανή είναι κάθε μία από αυτές τις καταστάσεις.

Κατανομή πιθανότητας

Μια κατανομή πιθανότητας είναι μια περιγραφή του πόσο πιθανό είναι μια τυχαία μεταβλητή ή ένα σύνολο τυχαίων μεταβλητών να λάβει καθεμία από τις πιθανές καταστάσεις της. Ο τρόπος με τον οποίο περιγράφονται τις κατανομές πιθανότητας εξαρτάται από το εάν οι μεταβλητές είναι διακριτές ή συνεχείς.

Όταν ασχολούμαστε με συνεχείς τυχαίες μεταβλητές, περιγράφουμε τις κατανομές πιθανότητας χρησιμοποιώντας μια συνάρτηση, γνωστή ως συνάρτηση πυκνότητας πιθανότητας (σ.π.π.). Για να είναι μια συνάρτηση σ.π.π. πρέπει να ικανοποιεί τις ακόλουθες ιδιότητες:

- Το πεδίο ορισμού της να περιλαμβάνει όλες τις πιθανές τιμές της τ.μ. x.
- $\forall x \in \mathrm{x}, p(x) > 0$
- $\int p(x)\, dx = 1$

Μια σ.π.π. $p(x)$ δεν δίνει απευθείας την πιθανότητα μιας συγκεκριμένης κατάστασης, αλλά την πιθανότητα να βρεθούμε μέσα σε περιοχή με όγκο δx δίνεται από το $p(x)\,\delta x$

Μπορούμε να ολοκληρώσουμε την σ.π.π. για να βρούμε την πραγματική μάζα πιθανότητας ενός συνόλου σημείων. Συγκεκριμένα, η πιθανότητα ότι το x βρίσκεται σε κάποιο σύνολο $\mathbb{S}$ δίνεται από το ολοκλήρωμα του $p(x)$ πάνω σε αυτό το σύνολο. Για μια μεταβλητή, η πιθανότητα ότι το x βρίσκεται στο διάστημα [a, b] δίνεται από τον υπολογισμό $\int_{[a,b]} p(x)\, dx$.

Όσον άφορα τις διακριτές μεταβλητές, η κατανομή πιθανότητας πάνω σε αυτές μπορεί να περιγραφεί χρησιμοποιώντας μια συνάρτηση μάζας πιθανότητας (σ.μ.π). Συνήθως δηλώνουμε τις συναρτήσεις μάζας πιθανότητας με κεφαλαίο P. Για να είναι μια συνάρτηση P, σ.μ.π μιας τυχαίας μεταβλητής x, πρέπει να ικανοποιεί τις ακόλουθες ιδιότητες:

- Το πεδίο ορισμού της να περιλαμβάνει όλες τις πιθανές τιμές της τ.μ. x.
- $\forall x \in \mathrm{x},\ 0 \leq P(x) \leq 1$. Ένα αδύνατο ενδεχόμενο έχει πιθανότητα 0. Αντίθετα, ένα βέβαιο ενδεχόμενο, που είναι δηλαδή εγγυημένο ότι θα συμβεί έχει πιθανότητα 1 και κανένα ενδεχόμενο δεν έχει περισσότερες πιθανότητες να συμβεί.
- $\sum_{x \in \mathrm{x}} P(x) = 1$

Δεσμευμένη πιθανότητα

Σε πολλές περιπτώσεις, μας ενδιαφέρει η πιθανότητα κάποιου ενδεχόμενου, δεδομένου ότι κάποιο άλλο ενδεχόμενο έχει συμβεί. Αυτό ονομάζεται δεσμευμένη πιθανότητα. Η πιθανότητα να συμβεί το ενδεχόμενο B δεδομένου ότι έχει συμβεί το ενδεχόμενο A συμβολίζεται ως $P(B \mid A)$. Ένας πολύ σημαντικός τύπος για τον υπολογισμό της δεσμευμένης πιθανότητας είναι ο εξής:

$$P(B \mid A) = \frac{P(A, B)}{P(B)}$$

Δεν μπορούμε να υπολογίσουμε την δεσμευμένη πιθανότητα που εξαρτάται από ένα συμβάν που δεν συμβαίνει ποτέ, δηλαδή όταν $P(B) = 0$. Αντίστοιχα ορίζονται και οι δεσμευμένες κατανομές πιθανότητας.

Περιθώρια κατανομή πιθανότητας

Αν x και y τυχαίες μεταβλητές των οποίων η από κοινού κατανομή είναι κάλα ορισμένη, τότε η κατανομή πιθανότητας της x, ονομάζεται περιθώρια κατανομή πιθανότητας (marginal probability distribution) της x και υπολογίζεται από τον τύπο:

$$p(\mathrm{x}) = \int \mathrm{p(x, y)}\, \mathrm{dy}$$

Για διακριτές τυχαίες μεταβλητές x, y ισχύει:

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y)$$

Θεώρημα Bayes

Το θεώρημα Bayes είναι ένας τρόπος για να ενημερώνουμε την πεποίθηση μας καθώς νέα δεδομένα κάνουν την εμφάνιση τους. Η πιθανότητα μιας υπόθεσης, B, δεχόμενοι κάποια νέα δεδομένα A, δηλώνεται, $P(B \mid A)$ και δίνεται από τον τύπο:

$$P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A)}$$

Αναμενόμενη τιμή

Η αναμενόμενη τιμή μιας συνάρτησης $f(x)$ σε σχέση με την κατανομή πιθανότητας $P(\mathrm{x})$ είναι η μέση τιμή που παίρνει η f όταν το x προέρχεται από τη P. Για συνεχείς μεταβλητές η αναμενόμενη τιμή υπολογίζεται ως:

$$\mathbf{E}_{\mathrm{x}\sim p}[f(\mathrm{x})] = \int p(x) f(x)\, dx$$

Για διακριτές μεταβλητές ο αντίστοιχος τύπος είναι :

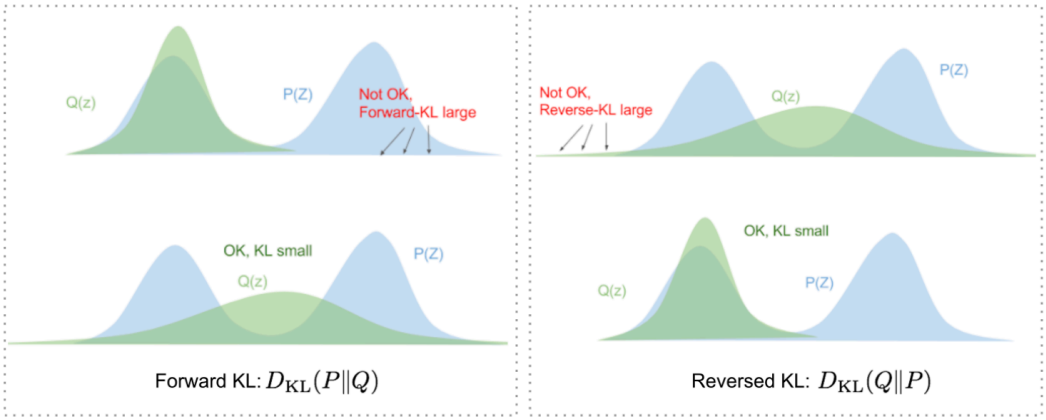
$$\mathbf{E}_{\mathrm{x}\sim P}[f(\mathrm{x})] = \sum_{x} P(x) f(x)$$

Απόκλιση Kullback–Leibler (KL divergence)

Αν έχουμε δύο ξεχωριστές κατανομές πιθανότητας $P(\mathrm{x})$ και $Q(\mathrm{x})$ πάνω στην ίδια τυχαία μεταβλητή x, μπορούμε να μετρήσουμε πόσο διαφέρουν αυτές οι δύο κατανομές κάνοντας χρήση της απόκλισης Kullback-Leibler:

$$D_{KL}(P\|Q) = \mathbf{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbf{E}_{\mathrm{x}\sim P}\left[\log P(x) - \log Q(x)\right]$$

Είναι σημαντικό να τονίσουμε πως η απόκλιση KL είναι πάντα θετική. Ακόμη αν και διαισθητικά ερμηνεύεται ως μια απόσταση μεταξύ δύο κατανομών, πρακτικά δεν ισχύει κάτι τέτοιο αφού δεν τηρεί την ιδιότητα της συμμετρίας, δηλαδή $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$.



Σχήμα 2.11: Ασυμμετρία απόκλισης KL.

Προσεγγιστικός Συμπερασμός (approximate inference)

Πάμε να διατυπώσουμε το πρόβλημα του προσεγγιστικού συμπερασμού, όπως αυτό μπορεί να οριστεί κατάλληλα στο πλαίσιο της μηχανικής μάθησης και συγκεκριμένα του variational autoencoder που παρουσιάζεται στην επόμενη ενότητα.

Έστω ένα σύνολο παρατηρήσεων $\mathbf{x}$ και ένα σύνολο λανθανουσών μεταβλητών $\mathbf{z}$, με από κοινού συνάρτηση κατανομής πιθανότητας $p(\mathbf{x}, \mathbf{z})$. Το πρόβλημα του συμπερασμού σε αυτή την περίπτωση είναι να βρούμε την εκ των υστέρων κατανομή $p(\mathbf{z} \mid \mathbf{x})$. Από το θεώρημα του Bayes ο υπολογισμός αυτός φαίνεται, αρχικά, να είναι αρκετά εύκολος:

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{z})\, p(\mathbf{z})}{p(\mathbf{x})}$$

Όπου το $p(\mathbf{z})$ είναι η εκ των προτέρων κατανομή και $p(\mathbf{x} \mid \mathbf{z})$ η πιθανοφάνεια. Παρόλα αυτά η απλότητα με την οποία εξηγεί το θεώρημα του Bayes τον παραπάνω υπολογισμό, κρύβει την πραγματική του πολυπλοκότητα. Στις περισσότερες των περιπτώσεων το $p(\mathbf{x})$ είναι μη επιλύσιμο, διότι πρέπει να υπολογιστεί ως περιθώρια συνάρτηση κατανομής από τον τύπο:

$$p(\mathbf{x}) = \int_z p(\mathbf{x} \mid \mathbf{z})\, p(\mathbf{z})\, d\mathbf{z}$$

Το ολοκλήρωμα που παρουσιάζεται σε αυτό τον τύπο απαιτεί εκθετικό χρόνο υπολογισμού, πάνω σε όλους τους πιθανούς συνδυασμούς των λανθανουσών μεταβλητών $\mathbf{z}$. Σε αυτό ακριβώς το σημείο έρχεται η έννοια του προσεγγιστικού συμπερασμού, για να επιλύσει αυτό το πρόβλημα.

Οι μέθοδοι εφαρμογής του προσεγγιστικού συμπερασμού μπορούν να χωριστούν σε δύο κατηγορίες, ντετερμινιστικές και στοχαστικές. Οι στοχαστικές μέθοδοι στηρίζονται κατά κύριο λόγο στην αναγωγή του προβλήματος του συμπερασμού, σε ένα πρόβλημα δειγματοληψίας από την εκ των υστέρων κατανομή κάνοντας χρήση κάποιων αλγορίθμων, οι οποίοι είναι γνωστοί ως Markov-Chain Monte Carlo (MCMC) αλγόριθμοι. Στην παρούσα εργασία οι μέθοδοι προσεγγιστικού συμπερασμού που θα μας απασχολήσουν είναι αυτές που χαρακτηρίζονται ως ντετερμινιστικές. Σε αυτή την περίπτωση το πρόβλημα του συμπερασμού ανάγεται σε ένα πρόβλημα βελτιστοποίησης, το οποίο μπορεί να λυθεί ευκολότερα και αποδοτικότερα. Έτσι, εφόσον δεν μπορούμε να υπολογίσουμε ακριβώς την ζητούμενη κατανομή επιλέγουμε να την προσεγγίσουμε, μέσω μιας άλλης κατανομής $q(\mathbf{z} \mid \mathbf{x})$, η οποία μπορεί να υπολογιστεί. Συγκεκριμένα η μέθοδος η οποία θα χρησιμοποιήσουμε για αυτή την προσέγγιση ονομάζεται variational inference και την αναφέρουμε με την αγγλική ονομασία μιας και δεν φαίνεται να υπάρχει κάποια καθιερωμένη ονομασία στα ελληνικά.

Με τη μέθοδο variational inference υπολογίζουμε την εκ των υστέρων κατανομή με μια οικογένεια κατανομών $q_\lambda(\mathbf{z} \mid \mathbf{x})$. Οι παράμετροι λ υποδεικνύουν την οικογένεια των κατανομών. Για παράδειγμα, αν μιλάμε για την οικογένεια των γκαουσιανών κατανομών, οι παράμετροι θα είναι η μέση τιμή και διακύμανση των λανθανουσών μεταβλητών για κάθε στοιχείο του συνόλου δεδομένων, δηλαδή: $\lambda_{x_i} = (\mu_{x_i}, \sigma^2_{x_i})$. Για να υπολογίσουμε πόσο κοντά είναι η προσέγγιση που κάναμε στην πραγματική κατανομή, μπορούμε να κάνουμε χρήση της απόκλισης KL σε συνδυασμό με τον ορισμό της δεσμευμένης κατανομής πιθανότητας για την εκ των υστέρων κατανομή $p(\mathbf{z} \mid \mathbf{x})$:

$$D_{KL}(q_\lambda(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z} \mid \mathbf{x})) = \mathbf{E}_q\big[\log q_\lambda(\mathbf{z} \mid \mathbf{x})\big] - \mathbf{E}_q\big[\log p(\mathbf{x}, \mathbf{z})\big] + \log p(\mathbf{x})$$

Ο τελικός σκοπός είναι να υπολογίσουμε τις παραμέτρους λ οι οποίες ελαχιστοποιούν αυτή την απόκλιση, το οποίο συντελεί και το πρόβλημα βελτιστοποίησης που καλούμαστε να επιλύσουμε.

$$q_\lambda^*(\mathbf{z} \mid \mathbf{x}) = \arg\max_\lambda \lambda D_{KL}(q_\lambda(\mathbf{z} \mid \mathbf{x}) \mid\mid p(\mathbf{z} \mid \mathbf{x}))$$

Παρατηρούμε όμως πως ο όρος $p(\mathbf{x})$ παρουσιάζεται στην εξίσωση και όπως αναφέραμε πρόκειται για έναν όρο που είναι μη υπολογίσιμος. Ξαναγράφουμε την εξίσωση ως εξής:

$$\log p(\mathbf{x}) = D_{KL}(q_\lambda(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z}|\mathbf{x})) + \mathbf{E}_q\big[\log p(\mathbf{x}, \mathbf{z})\big] - \mathbf{E}_q\big[\log q_\lambda(\mathbf{z}|\mathbf{x})\big] \qquad (1)$$

Μέσα από την εξίσωση αυτή αναγνωρίζουμε έναν νέο όρο που ονομάζουμε κατώτατο όριο απόδειξης (Evidence Lower Bound), ονομασία που προκύπτει αφού ως evidence χαρακτηρίζεται ο παρονομαστής του θεωρήματος του Bayes, $p(\mathbf{x})$:

$$ELBO(\lambda) = \mathbf{E}_q\big[\log p(\mathbf{x}, \mathbf{z})\big] - \mathbf{E}_q\big[\log q_\lambda(\mathbf{z}|\mathbf{x})\big] \qquad (2)$$

Άρα τελικά η εξίσωση (1) γράφεται:

$$\log p(\mathbf{x}) = D_{KL}(q_\lambda(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z}|\mathbf{x})) + ELBO(\lambda) \geq ELBO(\lambda) \qquad (3)$$

Εφόσον η απόκλιση KL είναι πάντα θετική, αυτό σημαίνει πως η ελαχιστοποίηση αυτής της απόκλισης είναι ισοδύναμη με την μεγιστοποίηση του ELBO. Έτσι μπορούμε να απαλλαχθούμε από την απόκλιση KL που περιέχει έναν μην υπολογίσιμο όρο.

Μια πλήρης μαθηματική ανάλυση για την εξαγωγή των τύπων της παρούσας ενότητας παρουσιάζεται στο **Παράρτημα Α**. Στην επόμενη ενότητα θα εντάξουμε όσα παρουσιάσαμε στο πλαίσιο του variational autoencoder και θα εξάγουμε τους τελικούς τύπους υπολογισμού της συνάρτησης κόστους που τον χαρακτηρίζει.

2.3.2 Variational Autoencoder

Ο variational autoencoder (VAE) (D. Kingma, M. Welling, 2014) αποτελεί ένα από τα πιο διαδεδομένα μοντέλα στο χώρο των παραγωγικών μοντέλων και παράλληλα αποτελεί το μοντέλο με το οποίο θα ασχοληθούμε στη συνέχεια της εργασίας. Η δημιουργία τους έρχεται να αντιμετωπίσει κάποιους περιορισμούς των παραδοσιακών autoencoders, όσον αφορά την παραγωγή νέων δεδομένων. Με μια πρώτη σκέψη, θα μπορούσε να φανταστεί κανείς πως, αφού οι απλοί autoencoders κωδικοποιούν χαρακτηριστικά των δεδομένων εισόδου στην ενδιάμεση αναπαράσταση και εφόσον αυτή η αναπαράσταση είναι επαρκώς κανονικοποιημένη, θα μπορούσαμε να επιλέξουμε τυχαία ένα σημείο και να το αποκωδικοποιήσουμε, παίρνοντας έτσι ένα νέο δεδομένο στην έξοδο. Ωστόσο, η συνέχεια της ενδιάμεσης κωδικοποίησης στους autoencoders είναι δύσκολο να επιτευχθεί, καθώς εξαρτάται άμεσα από την κατανομή των δεδομένων στον αρχικό χώρο, τη διάσταση της ενδιάμεσης κωδικοποίησης και την αρχιτεκτονική του κωδικοποιητή. Έτσι, είναι αρκετά δύσκολο, έως αδύνατο, να διασφαλιστεί, a priori, ότι ο κωδικοποιητής θα οργανώσει τον ενδιάμεσο χώρο με τρόπο έξυπνο και συμβατό με τη διαδικασία παραγωγής δεδομένων που μόλις περιγράψαμε.

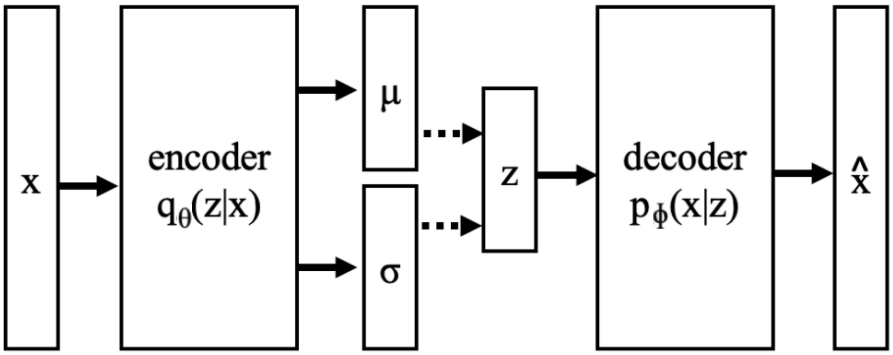
Αυτή η αδυναμία των απλών autoencoders δεν θα πρέπει να μας προκαλεί καμία έκπληξη, αφού δεν κατασκευάστηκαν ποτέ με σκοπό να μπορούν να πετύχουν μια τέτοια οργάνωση που θα

επέτρεπε την παραγωγή νέων δεδομένων. Όπως ήδη αναφέραμε, ο σκοπός τους είναι η ανακατασκευή των δεδομένων εισόδου με τις μικρότερες δυνατές απώλειες. Η συνάρτηση σφάλματος, μάλιστα που καλείται να βελτιστοποιήσει ένας autoencoder τον σπρώχνει προς αυτή τη κατεύθυνση, αφού στοχεύει αποκλειστικά στην ελαχιστοποίηση της διαφοράς που υπάρχει μεταξύ του δεδομένου εξόδου του αποκωδικοποιητή με το δεδομένο εισόδου του κωδικοποιητή.

Για να γίνει εφικτή η χρησιμοποίηση του αποκωδικοποιητή του autoencoder για γενετικούς σκοπούς, θα πρέπει να διασφαλίσουμε ότι η ενδιάμεση κωδικοποίηση είναι επαρκώς κανονική. Ένας τρόπος να επιτευχθεί αυτό είναι να εισαγάγουμε έναν όρο που θα επιβάλει ρητά την κανονικότητα κατά τη διάρκεια της εκπαίδευσης. Καταλήγουμε λοιπόν στον εξής άτυπο ορισμό για τους VAE:

> *«Ένας VAE μπορεί να οριστεί ως ένας autoencoder του οποίου η εκπαίδευση είναι κανονικοποιημένη για να αποφευχθεί η υπερπροσαρμογή (overfitting) και να διασφαλιστεί ότι ο χώρος κωδικοποίησης έχει τέτοιες ιδιότητες, ώστε να είναι εφικτή η παραγωγή νέων δεδομένων.»*

Επομένως για την κατασκευή ενός VAE παίρνουμε σαν βάση έναν απλό autoencoder με τα επιμέρους στοιχεία του, τον κωδικοποιητή και τον αποκωδικοποιητή και θέλουμε να εφαρμόσουμε μια βασική αλλαγή πάνω σε αυτά. Η αλλαγή αυτή είναι τελικώς να μην κωδικοποιούμε μια είσοδο σε ένα μοναδικό σημείο του χώρου της ενδιάμεσης κωδικοποίησης, αλλά να το κωδικοποιούμε σαν μια κατανομή πάνω σε αυτόν τον ενδιάμεσο χώρο. Με βάση αυτό το στόχο το δίκτυο λαμβάνει την μορφή που παρουσιάζεται στη συνέχεια.



Σχήμα 2.12: Αρχιτεκτονική VAE.

Ο κωδικοποιητής έχει είσοδο **x**, έξοδο τις παραμέτρους της ενδιάμεσης αναπαράστασης, **z**, και αποδίδει μια εκ των υστέρων συνάρτηση κατανομής $q(\mathbf{z}|\mathbf{x})$. Η περιγραφή αυτή μπορεί να παραμετροποιηθεί με ένα νευρωνικό δίκτυο μέσω ενός συνόλου βαρών θ. Έτσι συμβολίζουμε τον κωδικοποιητή ως $q_\theta(\mathbf{z}|\mathbf{x})$. Σημειώνουμε ότι ο λανθάνων χώρος είναι στοχαστικός, δηλαδή ο κωδικοποιητής εξάγει παραμέτρους στη $q_\theta(\mathbf{z}|\mathbf{x})$, που συνηθίζεται να περιγράφουν μια γκαουσιανή σ.π.π. και άρα αφορούν τη μέση τιμή και την τυπική απόκλιση ή την διακύμανση. Στη συνέχεια μπορούμε να δειγματοληπτήσουμε από αυτή την κατανομή και να πάρουμε θορυβώδεις αναπαραστάσεις του **z**.

Ο αποκωδικοποιητής έχει είσοδο **z**, έξοδο ένα νέο δεδομένο και αποδίδει μια πιθανοφάνεια $p(\mathbf{x}|\mathbf{z})$. Η περιγραφή αυτή μπορεί να παραμετροποιηθεί με ένα νευρωνικό δίκτυο μέσω ενός συνόλου βαρών φ. Συμβολίζουμε τον αποκωδικοποιητή ως $p_\varphi(\mathbf{x}|\mathbf{z})$. Αφού δειγματοληπτήσουμε

από την ενδιάμεση κατανομή, το δίκτυο του αποκωδικοποιητή μπορεί να αποκωδικοποιήσει αυτή την είσοδο και να παράγει νέα δεδομένα.

Στο μοντέλο του VAE, υπάρχουν μόνο τοπικές λανθάνουσες μεταβλητές (κανένα σημείο δεδομένων δεν μοιράζεται τη λανθάνουσα αναπαράσταση του, $\mathbf{z}$, με τη λανθάνουσα αναπαράσταση άλλου σημείου δεδομένων). Έτσι μπορούμε να αποσυνθέσουμε το ELBO σε ένα άθροισμα όπου κάθε όρος εξαρτάται από ένα μόνο σημείο δεδομένων. Αυτό μας επιτρέπει να χρησιμοποιούμε τον αλγόριθμο σύγκλισης με ελάττωση της παραγώγου (gradient descent) σε ότι αφορά τις παραμέτρους. Είναι σημαντικό να αναφέρουμε πως οι παράμετροι είναι κοινές για όλα τα δείγματα. Προκειμένου να εκφράσουμε την εξίσωση (2) της προηγούμενης ενότητας για ένα μόνο σημείο δεδομένων στον VAE και με βάση το πλαίσιο που περιγράψαμε στις δύο προηγούμενες παραγράφους, αυτή γράφεται:

$$ELBO_i(\theta,\varphi) = \mathbf{E}_q\Big[\log p_\varphi(\mathbf{x}_i,\mathbf{z})\Big] - \mathbf{E}_q\Big[\log q_\theta(\mathbf{z}\,|\,\mathbf{x}_i)\Big] = \mathbf{E}_q\Big[\log p_\varphi(\mathbf{x}_i,\mathbf{z})\Big] - \mathbf{E}_q\Big[\log q_\theta(\mathbf{z}\,|\,\mathbf{x}_i)\Big]$$

$$= \mathbf{E}_q\left[\log \frac{p_\varphi(\mathbf{x}_i\,|\,\mathbf{z})\,p_\varphi(\mathbf{z})}{q_\theta(\mathbf{z}\,|\,\mathbf{x}_i)}\right]$$

$$= \mathbf{E}_q\left[\log \frac{p_\varphi(\mathbf{z})}{q_\theta(\mathbf{z}\,|\,\mathbf{x}_i)} + \log p_\varphi(\mathbf{x}_i\,|\,\mathbf{z})\right]$$

$$= -D_{KL}(q_\theta(\mathbf{z}\,|\,\mathbf{x}_i)\,\|\,p_\varphi(\mathbf{z})) + \mathbf{E}_{\sim q_\theta(\mathbf{z}|\mathbf{x}_i)}\Big[\log p_\varphi(\mathbf{x}_i\,|\,\mathbf{z})\Big] \tag{4}$$

Άρα βάση των εξισώσεων (3) και (4) έχουμε:

$$\log p(\mathbf{x}_i) \geq -D_{KL}(q_\theta(\mathbf{z}\,|\,\mathbf{x}_i)\,\|\,p_\varphi(\mathbf{z})) + \mathbf{E}_{\sim q_\theta(\mathbf{z}|\mathbf{x}_i)}\Big[\log p_\varphi(\mathbf{x}_i\,|\,\mathbf{z})\Big] \tag{5}$$

Συνάρτηση κόστους VAE

Η συνάρτηση κόστους που ελαχιστοποιείται κατά την εκπαίδευση ενός VAE αποτελείται από έναν όρο "ανακατασκευής", που τείνει να κάνει το σχήμα κωδικοποίησης-αποκωδικοποίησης όσο το δυνατόν πιο αποδοτικό, και έναν όρο "κανονικοποίησης", που τείνει να κανονικοποιήσει την οργάνωση του λανθάνοντος χώρου φέρνοντας τις κατανομές που επιστρέφονται από τον κωδικοποιητή όσο το δυνατόν πιο κοντά σε μια τυπική κανονική κατανομή. Αυτός ο όρος κανονικοποίησης εκφράζεται ως η απόκλιση Kulback-Leibler μεταξύ της επιστρεφόμενης κατανομής και μιας τυπικής γκαουσιανής κατανομής. Έτσι προκειμένου να πάρουμε την κλειστή μορφή της συνάρτησης κόστους του VAE επιλέγουμε:

$$p_\varphi(z) \to \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2}\right)$$

$$q_\theta(z\,|\,x_i) \to \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right)$$

Με βάση αυτή την επιλογή και ακολουθώντας τις μαθηματικές πράξεις στο **Παράρτημα Β** καταλήγουμε στην τελική συνάρτηση κόστους του VAE, η οποία έχει την μορφή:

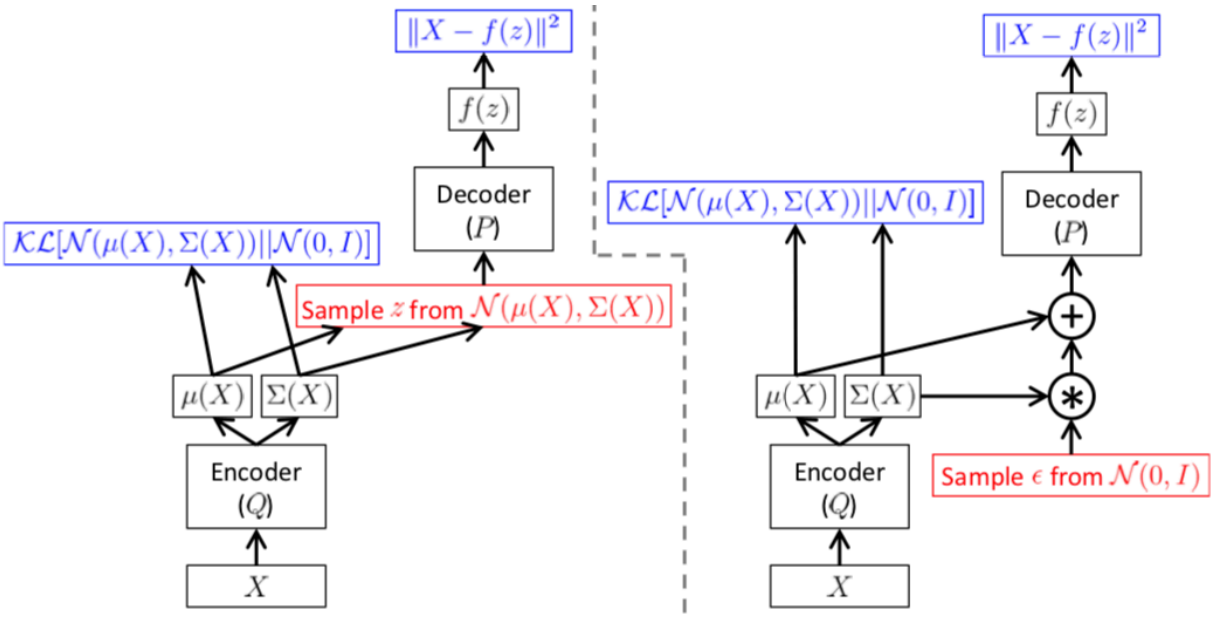
$$\mathscr{L} = \sum_{j=1}^{J} \frac{1}{2}\left[1 + \log(\sigma_i^2) - \sigma_i^2 - \mu_i^2\right] - \frac{1}{L}\sum_{l} \mathbf{E}_{\sim q_\theta(z|x_i)}\left[\log p_\varphi(x_i\,|\,z^{(l,i)})\right]$$

Ο κωδικοποιητής και ο αποκωδικοποιητής είναι ντετερμινιστικές συναρτήσεις. Εφόσον το $p(\mathbf{x}\,|\,\mathbf{z})$ είναι τέτοια συνάρτηση η οποία αντιστοιχίζει το $\mathbf{z}$ στο $\hat{\mathbf{x}}$, μπορούμε να σκεφτούμε την έκφραση ως $p(\mathbf{x}\,|\,\hat{\mathbf{x}})$. Λόγω του ότι υποθέσαμε γκαουσιανή κατανομή ο δεύτερος όρος της εξίσωσης είναι ανάλογος με το τετράγωνο σφάλμα:

$$\log p(\mathbf{x}\,|\,\hat{\mathbf{x}}) \sim \log e^{-|\mathbf{x}-\hat{\mathbf{x}}|^2} \sim (\mathbf{x}-\hat{\mathbf{x}})^2$$

Το τέχνασμα της επαναπαραμετροποίησης (reparametrization trick)

Στην περιγραφή του VAE που δώσαμε μέχρι τώρα υπάρχει ένας βασικός περιορισμός. Το γεγονός ότι υπάρχει ένα σημείο στο οποίο κάνουμε δειγματοληψία από μια κατανομή καθιστά την διαδικασία στοχαστική. Έτσι είναι αδύνατο να ανανεώσουμε τα βάρη του δικτύου μας με τη χρήση αλγορίθμων οπισθοδιάδοσης (back propagation) και άρα η διαδικασία της μάθησης δεν μπορεί να προχωρήσει. Η λύση στο πρόβλημα αυτό είναι γνωστή ως το τέχνασμα την ανά επαναπαραμετροποίησης (ή στα αγγλικά “the reparametrization trick”) και στοχεύει στην στην μεταφορά της διαδικασίας της δειγματοληψίας σε ένα ανεξάρτητο στρώμα εισόδου. Δεδομένης μιας μέσης τιμής μ και μίας τυπικής απόκλισης σ, μπορούμε να πραγματοποιήσουμε



Σχήμα 2.13: Δίκτυο VAE. Αριστερά χωρίς το reparametrization trick και δέξια με χρήση του reparametrization trick.

δειγματοληψία από την κατανομή $\mathcal{N}(\mu, \sigma)$, πραγματοποιώντας ισοδύναμε δειγματοληψία από την κατανομή $\epsilon \sim \mathcal{N}(0, I)$ και υπολογίζοντας στη συνέχεια το $z = \mu + \sigma \odot \epsilon$. Η διαδικασία αυτή φαίνεται σχηματικά στο σχήμα 2.13.

2.4 Ψηφιακή επεξεργασία ήχου

Από τη στιγμή που η παρούσα εργασία ασχολείται με ηχητικά δεδομένα, δεν θα μπορούσε να απουσιάζει από αυτήν μια ενότητα που να περιέχει βασικές έννοιες της ψηφιακής επεξεργασίας σημάτων. Στην ενότητα αυτή, παρουσιάζουμε τις βασικούς ορισμούς και εργαλεία που χρησιμοποιήθηκαν για τον χειρισμό των δεδομένων μας. Μιας και η ψηφιακή επεξεργασία σημάτων αποτελεί ένα πολύ ευρύ πεδίο με αρκετές δυσκολονόητες έννοιες, προσπαθούμε να μην αναλωθούμε σε λεπτομέρειες, αλλά να επικεντρωθούμε στα σημεία που είναι απαραίτητα για την πλήρη κατανόηση της εργασίας μας.

Ο ήχος αποτελεί ένα συνεχές, χρονικά μεταβαλλόμενο σήμα και προτού καταστεί εφικτό να αναλυθεί κάνοντας χρήση ηλεκτρονικού υπολογιστή, το σήμα πρέπει να ψηφιοποιηθεί από μια συσκευή που ονομάζεται μετατροπέας αναλογικού σε ψηφιακό (A/D), ή ψηφιοποιητής. Ο ψηφιοποιητής μετρά επανειλημμένα ή δειγματοληπτεί το στιγμιαίο πλάτος τάσης ενός συνεχώς μεταβαλλόμενου αναλογικού σήματος με συγκεκριμένο ρυθμό δειγματοληψίας (sample rate), που συνήθως πρόκειται για χιλιάδες ή δεκάδες χιλιάδες φορές ανά δευτερόλεπτο. Στην περίπτωση του ηχητικού σήματος, αυτή η χρονικά μεταβαλλόμενη τάση είναι ανάλογη με την ηχητική πίεση σε μια συσκευή όπως ένα μικρόφωνο. Η ψηφιακή αναπαράσταση του σήματος που δημιουργείται από τον ψηφιοποιητή αποτελείται έτσι από μια ακολουθία αριθμητικών τιμών που αντιπροσωπεύουν το πλάτος της αρχικής κυματομορφής σε διακριτά και ισαπέχοντα σημεία στο χρόνο. Τα αρχεία τύπου WAVE (.wav), από τα οποία απαρτίζεται και το σύνολο δεδομένων μας, πρόκειται για μια τέτοια ψηφιακή κωδικοποίηση του ήχου.

Για τις ανάγκες της εργασίας δεν μας αρκεί να έχουμε τον ήχο ως μια ακολουθία δειγμάτων, διότι συνήθως αυτή η μορφή αναπαράστασης του ήχου δεν προσφέρει επαρκή πληροφορία στο δίκτυο ώστε αυτό να είναι σε θέση να μάθει τα εγγενή χαρακτηριστικά μια κατηγορίας ήχων. Έτσι καταφεύγουμε σε άλλες μορφές αναπαράστασης του ήχου. Συγκεκριμένα, οι αναπαραστάσεις του ήχου, που μελετήθηκαν και συγκρίθηκαν στο πλαίσιο της εργασίας είναι οι εξής:

- Φασματογράφημα (Spectrogram)
- Φασματογράφημα Mel (Mel-Spectrogram)
- Συντελεστές συχνότητας Cepstral του Mel (Mel-Frequency Cepstral Coefficients, MFCCs)

Το πιο χρήσιμο εργαλείο για την παραγωγή αυτών των αναπαραστάσεων είναι ο μετασχηματισμός Fourier. Ο μετασχηματισμός Fourier αποτελεί ένα εργαλείο που καταφέρνει να δώσει πληροφορίες για έναν ήχο στο πεδίο της συχνότητας, αντί για το πεδίο του χρόνου. Αυτό καθιστά το φάσμα έναν διαισθητικά ευχάριστο τομέα για να εργαστεί, γιατί μπορούμε να εξετάσουμε οπτικά τα σήματα.

Τα σήματα ήχου που χειριζόμαστε αποτελούν σήματα διακριτού χρόνου και έτσι ο αντίστοιχος μετασχηματισμός Fourier που μας ενδιαφέρει είναι ο διακριτός μετασχηματισμός Fourier (Discrete Fourier Transform-DFT). Ο διακριτός μετασχηματισμός Fourier μετατρέπει μια ακολουθία N μιγαδικών αριθμών $\{\mathbf{x_n}\} := x_0, x_1, \ldots, x_{N-1}$ σε μια άλλη ακολουθία μιγαδικών αριθμών $\{\mathbf{X_k}\} := X_0, X_1, \ldots, X_{N-1}$, που ορίζεται από τον τύπο:

$$X_k = \sum_{n=0}^{N-1} x_n \cdot \mathrm{e}^{-\frac{i2\pi}{N}kn}$$

$$= \sum_{n=0}^{N-1} x_n \cdot \left[\cos\left(\frac{2\pi}{N}kn\right) - i \cdot \sin\left(\frac{2\pi}{N}kn\right)\right]$$

οπού η δεύτερη εξίσωση προκύπτει από την εξίσωση του Euler: $\mathrm{e}^{ix} = \cos x + i \sin x$

Για εισόδους με πραγματικές τιμές, τα θετικά και αρνητικά συστατικά στοιχεία της συχνότητας είναι συζυγείς μιγαδικοί μεταξύ τους, έτσι ώστε να διατηρούμε N μονάδες πληροφοριών. Ωστόσο, δεδομένου ότι τα φάσματα είναι διανύσματα μιγαδικών αριθμών, είναι δύσκολο να τα απεικονίσουμε με αυτή τη μορφή. Όπως θα δούμε και αργότερα όταν θα αναφερθούμε στην επεξεργασία των δεδομένων μας, οι μιγαδικές είσοδοι δεν συνηθίζονται στα νευρωνικά δίκτυα. Μια πρώτη λύση θα ήταν να σχεδιάσουμε το φάσμα πλάτους $|X_k|$ ή φάσμα ισχύος $|X_k|^2$. Λόγω όμως των μεγάλων διαφορών στο εύρος των διαφορετικών συχνοτήτων, δυστυχώς αυτές οι αναπαραστάσεις δεν εμφανίζουν εύκολα σχετικές πληροφορίες. Η τελική επιλογή στην οποία καταφεύγουμε είναι να πάρουμε το λογαριθμικό φάσμα και να το μετατρέψουμε σε decibels, δηλαδή $20log_{10} | X_k |$. Μέσω αυτής της μετατροπής οι ήχοι μπορούν να απεικονιστούν εμφανίζοντας αρκετή πληροφορία.

Ωστόσο, τα σήματα που μελετάμε στην εργασία είναι μη στάσιμα σήματα, δηλαδή το φασματικό και συχνοτικό τους περιεχόμενο αλλάζει μέσα στο χρόνο. Άρα το να ακολουθούσαμε την τακτική που μόλις περιγράψαμε για να μετατρέψουμε έναν ήχο στον τομέα συχνότητας, θα έχει ως αποτέλεσμα να χαθεί μεγάλο μέρος χρήσιμης πληροφορίας. Διαιρώντας το σήμα σε μικρότερα τμήματα, μπορούμε να επικεντρωθούμε στις ιδιότητες σήματος σε μια συγκεκριμένη χρονική στιγμή. Για την διαίρεση του σήματος θα χρειαστεί να κάνουμε χρήση κάποιου παραθύρου. Με τον τρόπο αυτό καταλήγουμε τελικά στον βραχυπρόθεσμο μετασχηματισμό Fourier (Short-Time Fourier Transform - STFT). Ο τρόπος με τον οποίο λαμβάνουμε τον STFT είναι να παραθυροποιήσουμε το σήμα, συνήθως με επικαλυπτόμενα παράθυρα (σε ποσοστό $25\,\%$, $50\,\%$ κτλ) και στη συνέχεια να εφαρμόσουμε διαδοχικούς DFT στα παράθυρα αυτά. Με το παράθυρο και τη λήψη του διακριτού μετασχηματισμού Fourier (DFT) κάθε παραθύρου, λαμβάνουμε τον STFT του σήματος. Συγκεκριμένα, για ένα σήμα εισόδου x_n και παράθυρο w_n, ο μετασχηματισμός ορίζεται ως:

$$STFT\{x_n\}(h,k) = X(h,k) = \sum_{n=0}^{N-1} x_{n+h} w_n \mathrm{e}^{-i2\pi\frac{kn}{N}}$$

όπου το h είναι διακριτό και το k συνεχές. Η μιγαδική ποσότητα $X(h,k)$ παριστάνει την k-στή συνιστώσα Fourier, στο h-στό χρονικό πλαίσιο. Ο τρόπος που υλοποιείται ο STFT σε έναν ηλεκτρονικό υπολογιστή είναι συνήθως με τη χρήση του αλγορίθμου του γρήγορου μετασχηματισμού Fourier (Fast Fourier Transform - FFT).

Ο STFT περιγράφει την εξέλιξη των συνιστωσών συχνότητας με την πάροδο του χρόνου. Όπως και το ίδιο το φάσμα, ένα από τα οφέλη των STFT είναι ότι οι παράμετροι του έχουν μια φυσική και διαισθητική ερμηνεία. Μια ακόμη ομοιότητα με το φάσμα είναι ότι η έξοδος του STFT αποτελείται από μιγαδικές τιμές, αν και η έξοδος του STFT είναι ένας πίνακας και όχι ένα διάνυσμα όπως στο φάσμα. Κατά συνέπεια, δεν μπορούμε να απεικονίσουμε άμεσα την έξοδο του STFT. Συνηθίζεται λοιπόν να απεικονίζουμε το μέτρο του πλάτους του STFT με έναν δισδιάστατο χάρτη θερμότητας με λογαριθμική κλίμακα, γνωστό και ως φασματογράφημα.

Φασματογράφημα Mel

Πολλές μελέτες έχουν δείξει πως οι άνθρωποι δεν αντιλαμβάνονται τις συχνότητες σε γραμμική κλίμακα και πως είναι αρκετά πιο ικανοί στο να αντιλαμβάνονται μικρές διαφορές στις χαμηλότερες συχνότητες από ότι στις υψηλές. Για παράδειγμα, είναι πολύ πιο εύκολο να εντοπιστεί η διαφορά μεταξύ 500 και 1000 Hz, αλλά πολύ πιο δύσκολο να γίνει αντιληπτή κάποια διαφορά μεταξύ 10.000 και 10.500 Hz, παρόλο που η απόσταση μεταξύ αυτών των δύο ζευγαριών είναι η ίδια.

Το 1937, οι Stevens, Volkmann και Newmann πρότειναν μια μονάδα για την μέτρηση του τόνου (pitch) έτσι ώστε ίσες αποστάσεις στον τόνο να ακούγονται εξίσου διαφορετικές στον ακροατή. Αυτή η κλίμακα που ορίζεται βάση αυτής της αρχής, ονομάστηκε κλίμακα mel. Ένα φασματογράφημα mel, δεν είναι τίποτα άλλο από ένα απλό φασματογράφημα που στον άξονα τις συχνότητας έχει την κλίμακα mel. Ο πιο γνωστός τύπος για την μετατροπή μεταξύ των μονάδων από την κλίμακα Hertz στην κλίμακα Mel είναι ο ακόλουθος:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

Συντελεστές συχνότητας Cepstral Mel (MFCCs)

Το cepstrum συχνότητας Mel (Mel-frequency Cepstrum - MFC) είναι μια αναπαράσταση του βραχυπρόθεσμου φάσματος ισχύος (short-term power spectrum) ενός ήχου, που βασίζεται σε έναν γραμμικό μετασχηματισμό συνημίτονου (Discrete Cosine Transform - DCT) ενός λογαριθμικού φάσματος ισχύος σε μια μη γραμμική κλίμακα συχνοτήτων Mel.

Οι συντελεστές συχνότητας Cepstral του Mel (MFCCs) είναι συντελεστές που αποτελούν συλλογικά ένα MFC. Μπορούν να εξαχθούν από έναν τύπο cepstral αναπαράστασης του ηχητικού κλιπ (ένα μη γραμμικό "φάσμα-ενός-φάσματος"). Η διαφορά μεταξύ του απλού cepstrum και του MFC είναι ότι στο MFC οι ζώνες συχνοτήτων απέχουν ίσες αποστάσεις στην κλίμακα Mel, η οποία όπως αναφέραμε, προσεγγίζει την απόκριση του ανθρώπινου ακουστικού συστήματος με μεγαλύτερη ακρίβεια από τις ζώνες συχνοτήτων γραμμικής απόστασης που χρησιμοποιούνται στο απλό cepstrum. Η μέθοδος εξαγωγής των MFFCs είναι η ακόλουθη:

1. Παίρνουμε τον μετασχηματισμό Fourier ενός σήματος.
2. Αντιστοιχίζουμε το φάσμα που λήφθηκε στην κλίμακα Mel, κάνοντας χρήση τριγωνικών επικαλυπτόμενων παραθύρων.
3. Λογαριθμίζουμε για κάθε μια από τις συχνότητες Mel που λήφθηκαν.
4. Παίρνουμε το DCT των λογαριθμισμένων Mel συχνοτήτων, σαν να ήταν σήμα.
5. Οι MFCCs είναι τα πλάτη της εξόδου του DCT.

Τα βήματα που περιγράφηκαν αποτελούν μια υψηλού εμπέδου περιγραφή της εξαγωγής των MFCCs. Η υλοποίηση της μεθόδου αυτής κρύβει αρκετές λεπτομέρειες, οι οποίες όμως δεν αποτελούν αντικείμενο της παρούσας εργασίας. Για της ανάγκες της εργασίας η βιβλιοθήκη librosa παρέχει έτοιμα όλα τα εργαλεία που χρειαζόμαστε για την εξαγωγή των MFCCs.

3

3. Χρήσιμα στοιχεία από την βιβλιογραφία

3.1 Σύγκριση τρόπων αναπαράστασης του ήχου

Ένα σημείο της εργασίας που έπαιξε καθοριστικό ρόλο στην ποιότητα των αποτελεσμάτων είναι η επιλογή της κατάλληλης αναπαράστασης του ήχου, που προορίζεται να δοθεί ως είσοδος στο νευρωνικό μας δίκτυο. Μάλιστα η έρευνα που έγινε πάνω σε αυτό το κομμάτι αποτελεί ένα σημαντικό έμμεσο αποτέλεσμα της παρούσας εργασίας και έναν έμμεσο τρόπο αξιολόγησης της καταλληλότητας των διαφόρων αναπαραστάσεων του ήχου όταν αυτοί προορίζονται ως είσοδοι ενός συνελικτικού νευρωνικού δικτύου με σκοπό την εξαγωγή χαρακτηριστικών υψηλού επιπέδου του ήχου.

Μια έρευνα από την βιβλιογραφία (Huzaifah, 2017) πραγματεύεται ακριβώς αυτό το θέμα της σύγκρισης των διαφορετικών αναπαραστάσεων του ήχου. Συγκεκριμένα οι αναπαραστάσεις που τίθενται υπό σύγκριση είναι ο STFT με γραμμική και Mel κλίμακα, οι MFCCs, ο constant-Q μετασχηματισμός (CQT) και ο συνεχής μετασχηματισμός Wavelet (CWT). Ο τρόπος αξιολόγησης αυτών των μεθόδων γίνεται μέσω της σύγκρισης των επιδόσεων τους πάνω στην ταξινόμηση του UrbanSound8k κάνοντας χρήση διαφορετικών αρχιτεκτονικών κάποιων CNNs.

Τα αποτελέσματα της έρευνας έδειξαν ότι το Mel φασματογράφημα είχε σταθερά καλή απόδοση σε όλα τα σενάρια στα οποία δοκιμάστηκε, δηλαδή με διαφορετικές αρχιτεκτονικές δικτύων και διαφορετικά μεγέθη παραθύρων για τον STFT. Σε κάποια από αυτά τα σενάρια ο απλός STFT καθώς και ο CQT είχαν παρόμοια απόδοση. Ένα γενικότερο συμπέρασμα είναι ότι όλες οι χρονοσυχνοτικές αναπαραστάσεις είχαν καλύτερη επίδοση από τους MFCCs, οι οποίοι θεωρήθηκαν σαν τη μέθοδο αναφοράς στα πειράματα που διεξήχθησαν. Τελικώς η επιλογή μεταξύ του Mel-STFT και του απλού STFT έγκειται στα εγγενή χαρακτηριστικά των ήχων, το μέγεθος και την αρχιτεκτονική των δικτύων και το μέγεθος του παραθύρου στην εξαγωγή των μετασχηματισμών.

Στην παρούσα εργασία, όπως θα δούμε στη συνέχεια, κάνουμε χρήση και των δύο μεθόδων και τα αποτελέσματα παρουσιάζονται στο κεφάλαιο 5 οπού αναλύουμε και τα δικά μας συμπεράσματα σε ότι αφορά τη καταλληλότερη αναπαράσταση του ήχου.

3.2 Συνελικτικός Variational Autoencoder (CVAE)

Ο συνελικτικός VAE (CVAE) πρόκειται για έναν συνδυασμό των δυο ξεχωριστών δικτύων που ήδη παρουσιάσαμε, δηλαδή του CNN και του VAE. Ο CVAE είναι ένα δίκτυο το οποίο είναι αρκετά νέο και με λίγες αναφορές στην βιβλιογραφία σε σύγκριση με άλλα διαδεδομένα νευρωνικά δίκτυα. Η κύρια χρήση του που συναντά κανείς αφορά στην παραγωγή εικόνων, όπως για παράδειγμα προσώπων. Η χρήση του για την παραγωγή ήχου είναι περιορισμένη και αφορά κυρίως δομημένους ήχους όπως σύνθεση μουσικής ή φωνής. Αυτός είναι και ένας από τους λόγους που επιλέξαμε να κάνουμε χρήση του στην παρούσα εργασία και να αξιολογήσουμε την επίδοση του στην παραγωγή νέων περιβαλλοντικών ήχων και να εμπλουτίσουμε με αυτόν τον τρόπο την βιβλιογραφία χαρακτηρίζοντας τα αποτελέσματα του και την καταλληλότητα του πάνω στη συγκεκριμένη εργασία.

Η αρχιτεκτονική του CVAE είναι ίδια με αυτή ενός απλού VAE, με τη διαφορά ότι ο κωδικοποιητής και ο αποκωδικοποιητής του CVAE είναι ένα CNN και ένα transpose CNN αντίστοιχα. Πιο συγκεκριμένα η είσοδος στον κωδικοποιητή είναι δισδιάστατη, με διαστάσεις $M \times N$. Στη συνέχεια ακολουθεί μια σειρά από στρώματα συγκέντρωσης και συνελικτικά στρώματα, που μειώνουν τις διαστάσεις της εισόδου και αυξάνουν τον αριθμό των φίλτρων που εφαρμόζονται σε αυτή αντίστοιχα. Για να φτάσουμε στην ενδιάμεση κρυφή κωδικοποίηση του CVAE, εφαρμόζουμε ένα στρώμα εξομάλυνσης (flatten layer) το οποίο μετατρέπει την είσοδο του σε ένα διάνυσμα με τα ίδια στοιχεία αλλά μιας μόνο διάστασης. Στη συνέχεια προτού πάρουμε την ενδιάμεση κωδικοποίηση μπορούμε να παρεμβάλουμε ένα ή περισσότερα πλήρως συνδεδεμένα στρώματα για να μειώσουμε περαιτέρω το μέγεθος της ενδιάμεσης κωδικοποίησης του CVAE. Το τελικό διάνυσμα που παίρνουμε, έστω μεγέθους K, το χωρίζουμε σε δυο υποδιανύσματα ίδιου μεγέθους, δηλαδή $K/2$, εκ των οποίων το ένα αντιπροσωπεύει τις μέσες τιμές και το άλλο τις διακυμάνσεις (ή τους λογαρίθμους) των χαρακτηριστικών που έχουν εξαχθεί.

Το δίκτυο του αποκωδικοποιητή έχει σαν τελικό σκοπό την παραγωγή μιας δισδιάστατης εξόδου ίδιας διάστασης με την είσοδο του κωδικοποιητή, δηλαδή $M \times N$. Έτσι, αρχικά λαμβάνει σαν είσοδο ένα διάνυσμα μεγέθους $K/2$, το οποίο περιλαμβάνει τις τιμές των χαρακτηριστικών της ενδιάμεσης κωδικοποίησης, όπως αυτές λήφθηκαν από τις κατανομές που διαμορφώθηκαν από τις μέσες τιμές και τις διακυμάνσεις της ενδιάμεσης κωδικοποίησης. Στη συνέχεια ακολουθεί μια σειρά στρωμάτων με σκοπό να καθρεφτίσει τα στρώματα που είδαμε στον κωδικοποιητή. Αυτό σημαίνει πως αρχικά είναι πιθανό να έχουμε μια σειρά από πλήρως συνδεδεμένα στρώματα, στη συνέχεια ένα στρώμα Reshape ώστε να αντιστρέψουμε το στρώμα εξομάλυνσης του κωδικοποιητή και τέλος μια σειρά στρωμάτων που ονομάζονται transposed convolution layers (βλ. ενότητα 2.2.4). Έτσι τελικώς λαμβάνεται μια έξοδος με τις επιθυμητές διαστάσεις.

Φυσικά, όπως κάθε νευρωνικό δίκτυο, η συμπεριφορά του CVAE μπορεί να ρυθμιστεί μέσω των υπερπαραμέτρων των διαφόρων στρωμάτων. Στην ενότητα 4.2 παρουσιάζουμε τις σχεδιαστικές επιλογές που εμείς πραγματοποιήσαμε σε ότι άφορα τον CVAE.

3.3 Ταξινόμηση του UrbanSound8K

Το UrbanSound8K είναι το σύνολο δεδομένων που χρησιμοποιούμε στην εργασία αυτή για την εκπαίδευση των παραγωγικών δικτύων. Στην ενότητα 4.1 υπάρχει η αναλυτική παρουσίαση αυτού του συνόλου δεδομένων καθώς και του τρόπου με τον οποίο το χειριστήκαμε. Στην παρούσα ενότητα παρουσιάζουμε τα αποτελέσματα της ταξινόμησης του συγκεκριμένου συνόλου δεδομένων όπως αυτά παρουσιάζονται στην βιβλιογραφία κατά τη διάρκεια συγγραφής της εργασίας. Σημειώνουμε πως η εργασία της ταξινόμησης έχει διαφορετικές απαιτήσεις από ότι αυτή

της παραγωγής νέων ήχων, μιας και η δεύτερη δεν αρκείται στην εξαγωγή χαρακτηριστικών από τους ήχους, άλλα απαιτεί αυτά τα χαρακτηριστικά να είναι κατάλληλα έτσι ώστε να μπορούν να κατασκευαστούν νέοι ήχοι. Παρόλα αυτά η μελέτη της επίδοσης των διαφόρων μοντέλων ταξινόμησης πάνω στο σύνολο δεδομένων που χρησιμοποιήθηκε είναι ικανή να μας δώσει σημαντικές πληροφορίες για την ευχρηστία αυτού του συνόλου δεδομένων.

Το μεγαλύτερο πλήθος εργασιών πάνω στο Urbansound8K κάνουν χρήση CNN για την ταξινόμηση των ήχων. Οι πιο συνηθισμένες προσεγγίσεις βασίζονται σε φασματογραφήματα και ειδικότερα σε φασματογράφημα κλίμακας Mel.

Ένα δίκτυο με ονομασία PiczakCNN, δημιουργήθηκε το 2015 και ήταν μία από τις πρώτες εφαρμογές CNN στο σύνολο δεδομένων Urbansound8K. Το δίκτυο κάνει χρήση 2 καναλιών log-Mel-φασματογραφημάτων, περιλαμβάνοντας τόσο τις τιμές του απλού φασματογραφήματος όσο και τη διαφορά πρώτης τάξης. Το μοντέλο χρησιμοποιεί 2 συνελικτικά στρώματα, το πρώτο με μέγεθος 57×6 (συχνότητα x χρόνο) και στη συνέχεια 1×3, ακολουθούμενα από δύο πλήρως συνδεδεμένα στρώματα με 5000 νευρώνες το καθένα. Η εργασία συγκρίνει βραχέα (950 ms) έναντι μεγαλύτερων (2,3 δευτερόλεπτα) παραθύρων ανάλυσης και τη μέθοδο majority voting έναντι της μεθόδου probability voting. Η απόδοση στο Urbansound8K κυμάνθηκε από 69 % έως 73 %. Παρατηρήθηκε ότι το probability voting και τα μεγάλα παράθυρα είχαν ελαφρώς καλύτερη απόδοση.

Το SB-CNN (2016) είναι ένα CNN 3 επιπέδων με ομοιόμορφους πυρήνες μεγέθους 5×5 και στρώμα συγκέντρωσης μεγίστου, μεγέθους 4×2. Η εργασία αναλύει επίσης την επίδραση διαφόρων τύπων data augmentation πάνω στο Urbansound8K, συμπεριλαμβανομένων των Time Shift, Pitch Shift, Dynamic Range Compression και Background Noise. Ύστερα από την εφαρμογή όλων αυτών των μεθόδων η επίδοση του μοντέλου αυξήθηκε από 72 % σε 79 %.

Το D-CNN (2017) χρησιμοποιεί μια αρχιτεκτονική μοντέλου που ακολουθεί σε μεγάλο βαθμό αυτό του PiczakCNN, ωστόσο, το δεύτερο στρώμα χρησιμοποιεί τη μέθοδο των dilated convolutions με dilation rate ίσο με 2. Η απόδοση του έφτασε το 81.9 % στο Urbansound8K. Η συνάρτηση ενεργοποίησης LeakyRelu φάνηκε να αποδίδει ελαφρώς καλύτερα από την ReLu που σημείωσε Α.

Μια πρόσφατη εργασία (Z. Zhang et al, 2018) διερεύνησε την επίδραση της μεθόδου mixup για το data augmentation. Αυτή η μέθοδος περιλαμβάνει ουσιαστικά την ανάμειξη των διαφόρων δεδομένων εκπαίδευσης. Το μοντέλο χρησιμοποιεί 4 μπλοκ με 2 συνελικτικά στρώματα το καθένα, με κάθε μπλοκ να ακολουθείται από ένα στρώμα μέγιστης συγκέντρωσης. Το δεύτερο μπλοκ και το τρίτο μπλοκ μαζί σχηματίζουν μια χωρικά χωρισμένη συνέλιξη: το δεύτερο μπλοκ χρησιμοποιεί δύο 3×1 συνελίξεις και το τρίτο μπλοκ χρησιμοποιεί δύο συνελίξεις 1×5. Στα φασματογραφήματα κλίμακας Mel το μοντέλο σημείωσε 74.7 % στο Urbansound8k χωρίς data augmentation, 77.3 % με εφαρμογή μόνο mixup και 82.6 % όταν το time stretch και το pitch shift συνδυάστηκαν με mixup. Μάλιστα όταν έγινε χρήση φασματογραφήματος Gammatone έναντι φασματογραφήματος Mel η απόδοση αυξήθηκε στο 83.7 %, το οποίο φαίνεται ήταν και το state-of-the-art τον Απρίλιο του 2019.

Φαίνεται λοιπόν πως το Urbansound8K είναι ένα σύνολο δεδομένων που παρουσιάζει καλή συμπεριφορά στην εργασία της ταξινόμησης. Οι επιδόσεις που μόλις περιγράψαμε είναι καλύτερες από την αντίστοιχη ενός ανθρώπου στο συγκεκριμένο σύνολο δεδομένων, που βάση της βιβλιογραφίας φάνηκε να κυμαίνεται στο 75 %. Οι καλές επιδόσεις των CNN στο Urbansound8K σε ότι αφορά την ταξινόμηση δεν εξασφαλίζουν την καλή επίδοση του σε ότι αφορά την διαδικασία παραγωγής νέων ήχων. Παρόλα αυτά τα αποτελέσματα αυτά μας υποδεικνύουν πως τα δείγματα

του συνόλου δεδομένων παρουσιάζουν ικανοποιητικά διακριτά χαρακτηριστικά που κατά πάσα πιθανότητα θα φανούν χρήσιμα και στη διαδικασία παραγωγής που θέλουμε να υλοποιήσουμε.

3.4 Τρόποι αξιολόγησης των παραγωγικών μοντέλων

Στην επιβλεπόμενη μάθηση ο τρόπος αξιολόγησης ενός μοντέλου είναι αρκετά προφανής, αφού έχουμε στην διάθεση μας τις απαιτούμενες ετικέτες για τα αντίστοιχα δεδομένα. Αυτό σημαίνει πως μπορούμε να αξιολογήσουμε την επίδοση του μοντέλου πάνω στα δεδομένα εκπαίδευσης με διάφορες μετρικές, όπως για παράδειγμα η ακρίβεια (precision), η ανάκληση (recall) και το F score, οι οποίες αποτελούν μερικές από τις πιο γνωστές μετρικές αξιολόγησης της επίδοσης μοντέλων επιβλεπόμενης μάθησης. Μπορούμε μάλιστα να λάβουμε μια καλή εκτίμηση της επίδοσης του συστήματος και όταν θα κληθεί να αντιμετωπίσει νέα δεδομένα μέσω της χρήσης ενός συνόλου δεδομένων ελέγχου (test set). Οι μετρικές αυτές είναι ικανές να δώσουν το μέτρο της πραγματικής απόδοσης του συστήματος βασισμένες σε αντικειμενικά κριτήρια.

Ο τρόπος αξιολόγησης των παραγωγικών μοντέλων, από την άλλη, είναι ένα περίπλοκο και μη ξεκάθαρο ζήτημα. Ο λόγος για τον οποίο συμβαίνει αυτό είναι πως τα αποτελέσματα ενός παραγωγικού μοντέλου είναι, πολύ συχνά, εντελώς καινούργια δεδομένα. Όταν συμβαίνει αυτό είναι δύσκολο για έναν αλγόριθμο να κρίνει το αποτέλεσμα αντικειμενικά ως προς την σημασιολογία του. Αυτό σημαίνει πως, για παράδειγμα, ένα δίκτυο παραγωγής εικόνων προσώπου δεν αρκεί να παράγει εικόνες υψηλής ανάλυσης για να χαρακτηριστεί αποδοτικό. Πρέπει οι εικόνες που παράγει να αντιπροσωπεύουν αληθοφανή πρόσωπα με πειστικά χαρακτηριστικά. Μάλιστα είναι ξεκάθαρο πως ένα δίκτυο που θα καταφέρει να παράγει αληθοφανή πρόσωπα χαμηλής ποιότητας, χαρακτηρίζεται καλύτερο παραγωγικό δίκτυο από ένα άλλο που θα παράγει εικόνες άριστης ποιότητας που υστερούν όμως στο να παρουσιάσουν ένα πρόσωπο.

Στον VAE καλούμαστε να βελτιστοποιήσουμε μια συνάρτηση κόστους όπως αυτή αναλύθηκε στην ενότητα 2.3.2. Παρόλα αυτά η σύγκλιση και η βελτιστοποίηση της συνάρτησης αυτής δεν εγγυάται ότι τα αποτελέσματα της παραγωγικής διαδικασίας είναι πειστικά και αληθοφανή. Εγγυάται μόνο πως το μοντέλο μας μπορεί να φτάσει στο μέγιστο των δυνατοτήτων του, κάτι που από μόνο του δεν εγγυάται την ποιότητα των αποτελεσμάτων.

Για την αξιολόγηση των αποτελεσμάτων ενός παραγωγικού μοντέλου υπάρχουν δυο διαφορετικές προσεγγίσεις όπως αυτές αναλύονται στην εργασία των T. Salimans, I. Goodfellow et al (2016):

- Αξιολόγηση από άνθρωπο
- Αξιολόγηση μέσω ενός εκπαιδευμένου μοντέλου ταξινόμησης

Στην πρώτη περίπτωση η ανθρώπινη κρίση είναι αυτή που καθορίζει, βάση κάποιων αντικειμενικών κριτηρίων, πόσο καλό χαρακτηρίζεται ένα παραχθέν δεδομένο. Παρόλα αυτά όσο αντικειμενικά και αν είναι τα κριτήρια αξιολόγησης, ο κάθε άνθρωπος κρίνει διαφορετικά κάθε αποτέλεσμα επομένως παραμένει ως ένα βαθμό αβέβαιη η αντικειμενικότητα της αξιολόγησης. Ακόμη παρατηρήθηκε πως αν οι αξιολογητές ενημερωθούν για τυχόν λάθος αξιολογήσεις, αλλάζουν την άποψη τους και είναι σε θέση να εντοπίσουν πιο εύκολα πιθανά λάθη στα νέα δεδομένα.

Στη δεύτερη περίπτωση γίνεται χρήση του inception score. Η λογική πίσω από αυτή την μετρική είναι πως αρχικά κάθε ξεχωριστό παραχθέν δεδομένο πρέπει να έχει χαμηλή εντροπία, ή με άλλα λόγια η συνάρτηση κατανομής πάνω στις κλάσεις αντικειμένων που περιέχονται στο παραχθέν δεδομένο να είναι επικεντρωμένη σε μία κλάση και ταυτόχρονα η περιθώρια κατανομή πάνω σε όλα τα παραχθέντα δεδομένα να έχει υψηλή εντροπία, δηλαδή να υπάρχει ποικιλομορφία

στην έξοδο και η συνάρτηση κατανομής πάνω σε όλα τα δεδομένα να προσεγγίζει την ομοιόμορφη. Το inception score βασίζεται στην ύπαρξη ενός ήδη εκπαιδευμένου μοντέλου ταξινόμησης πάνω σε ένα σύνολο δεδομένων διαφορετικών κλάσεων. Αυτό κατευθείαν φανερώνει έναν βασικό περιορισμό του, ότι δηλαδή δεν μπορεί να χαρακτηρίσει αντικειμενικά δεδομένα που ανήκουν σε κλάσεις που δεν γνωρίζει, δηλαδή σε αυτή την περίπτωση ακόμη και πολύ αξιόλογα παραχθέντα δεδομένα πιθανώς θα λάβουν χαμηλό inception score. Ακόμη είναι πιθανό πως αν το παραγωγικό μοντέλο απλά απομνημονεύσει τα δεδομένα εισόδου και τα αναπαράγει στην έξοδο, θα λάβει υψηλό inception score.

Γενικά το ζήτημα της αξιολόγησης των παραγωγικών μοντέλων είναι ένα ανοιχτό ζήτημα το οποίο δεν έχει μια ξεκάθαρη και τυποποιημένη λύση. Παρόλα αυτά τα περισσότερα παραγωγικά μοντέλα, όπως για παράδειγμα αυτά που παράγουν μουσική, μπορούν να αξιολογηθούν επαρκώς από άνθρωπο για το αν είναι ικανοποιητικά ή όχι τα αποτελέσματα τους. Αυτή τη λογική θα ακολουθήσουμε και στην παρούσα εργασία όπως θα δούμε στο κεφάλαιο 5.

4

4. Η εργασία

4.1 Το σύνολο δεδομένων μας

Στην ενότητα αυτή παρουσιάζουμε το σύνολο δεδομένων που χρησιμοποιήσαμε. Το σύνολο δεδομένων έχει την ονομασία UrbanSound8K και είναι διαθέσιμο για ελεύθερη και μη εμπορική χρήση. Στη συνέχεια της ενότητας παρουσιάζουμε μια πιο εκτενή ανάλυση του συνόλου δεδομένων, τον τρόπο με τον οποίο προεπεξεργαστήκαμε τα δεδομένα καθώς και τους τρόπους αναπαράστασης των ήχων ώστε να μπορούν αργότερα να τροφοδοτηθούν στην είσοδο των δικτύων μας..

4.1.1 Παρουσίαση συνόλου δεδομένων

Το UrbanSound8K είναι ένα σύνολο δεδομένων που, όπως είναι εμφανές και από την ονομασία του, αποτελείται από ήχους αστικού περιβάλλοντος. Συγκεκριμένα περιλαμβάνει 8732 ηχητικά κλιπ (τύπου .wav) με διάρκεια μέχρι και 4 δευτερόλεπτα. Κάθε ένα από αυτά τα ηχητικά κλιπ συνοδεύεται από μια ετικέτα που ανήκει σε μια από 10 κατηγορίες. Οι κατηγορίες που περιλαμβάνονται στο σύνολο δεδομένων είναι οι έξης:

- air_conditioner (κλιματιστικό)
- car_horn (κόρνα αυτοκινήτου)
- children_playing (παιδιά που παίζουν)
- dog_bark (γάβγισμα σκύλου)
- drilling (τρυπάνι)
- engine_idling (κινητήρας σε λειτουργία)
- gun_shot (πυροβολισμός)
- jackhammer (κομπρεσέρ)
- siren (σειρήνα)
- street_music (μουσική δρόμου)

Τα ηχητικά κλιπ δεν έχουν τον ίδιο ρυθμό δειγματοληψίας. Συγκεκριμένα το μεγαλύτερο ποσοστό (61 %) των ηχητικών κλιπ έχει ρυθμό δειγματοληψίας 44.1 kHz, αλλά υπάρχουν και ηχητικά κλιπ με πολύ διαφορετικούς ρυθμούς δειγματοληψίας. Στο Σχήμα 4.1 φαίνονται στην αριστερή στήλη οι ρυθμοί δειγματοληψίας και στην δέξια το ποσοστό των ηχητικών κλιπ που έχει δειγματοληπτηθεί με αυτόν τον ρυθμό.

```
44100     0.614979
48000     0.286532
96000     0.069858
24000     0.009391
16000     0.005153
22050     0.005039
11025     0.004466
192000    0.001947
8000      0.001374
11024     0.000802
32000     0.000458
Name: sample_rate, dtype: float64
```

Σχήμα 4.1: Ρυθμοί δειγματοληψίας δεδομένων.

Τα δεδομένα διαφέρουν επίσης στον αριθμό των καναλιών ήχου. Στο σχήμα που ακολουθεί φαίνονται οι πιθανές τιμές των καναλιών με τα αντίστοιχα ποσοστά.

```
2    0.915369
1    0.084631
Name: num_channels, dtype: float64
```

Σχήμα 4.2: Αριθμός καναλιών δεδομένων.

Ομοιομορφία δεν υπάρχει ούτε στο βάθος bit (bit depth), αφού και σε αυτή την περίπτωση τα αρχεία παίρνουν πολλές διαφορετικές τιμές (Σχ. 4.3).

```
16    0.659414
24    0.315277
32    0.019354
8     0.004924
4     0.001031
Name: bit_depth, dtype: float64
```

Σχήμα 4.3: Βάθος bit δεδομένων.

Όλες αυτές τις διαφορές μεταξύ των αρχείων οφείλουμε να τις εξαλείψουμε, ώστε να παρουσιάζουν μια ομοιομορφία στα χαρακτηριστικά τους προτού τα τροφοδοτήσουμε στο δίκτυο μας. Οι τρόποι που θα μας βοηθήσουν στο να επιτύχουμε αυτή την ομοιομορφία παρουσιάζονται στην αμέσως επόμενη υποενότητα.

Ένα τελευταίο στοιχείο που παρουσιάζουμε σε αυτή την υποενότητα είναι η κατανομή των δεδομένων στις 10 κατηγορίες του συνόλου δεδομένων. Η κατανομή αυτή είναι σχεδόν ομοιόμορφη αφού μόνο οι κατηγορίες car_horn και gun_shot υστερούν σε σχέση με τις άλλες κατηγορίες.

```
air_conditioner     1000
street_music        1000
drilling            1000
dog_bark            1000
children_playing    1000
engine_idling       1000
jackhammer          1000
siren                929
car_horn             429
gun_shot             374
Name: class, dtype: int64
```

Σχήμα 4.4: Κατανομή δεδομένων στις κλάσεις.

4.1.2 Προεπεξεργασία συνόλου δεδομένων

Όπως αναφέρθηκε και στην προηγούμενη υποενότητα, τα δεδομένα μας έχουν μερικές διαφορές στον τρόπο με τον οποίο έχουν συλλεχθεί, που καλούμαστε να εξαλείψουμε.

Το πρώτο βήμα είναι να πετύχουμε ίδιο ρυθμό δειγματοληψίας σε όλα μας τα δεδομένα. Αυτό είναι πολύ εύκολο να επιτευχθεί με την βοήθεια της βιβλιοθήκης librosa της Python. Η librosa είναι μια βιβλιοθήκη ειδικά σχεδιασμένη για την ανάλυση και επεξεργασία ήχου και μουσικής. Για να πετύχουμε λοιπόν ίδιο ρυθμό δειγματοληψίας σε όλα τα δεδομένα αρκεί να φορτώνουμε κάθε αρχείο με τη βοήθεια της εντολής " librosa.load" και να ορίζουμε την συχνότητα δειγματοληψίας σε μια συγκεκριμένη τιμή. Η επιλογή μας ήταν τα 16kHz διότι αποτελεί μια συχνότητα στην οποία δεν χάνεται σημαντική πληροφορία και ταυτόχρονα βοηθάει στη μείωση της διαστατικότητας των δεδομένων, πράγμα πολύ σημαντικό δεδομένων των περιορισμών υπολογιστικών πόρων. Ταυτόχρονα με την ίδια εντολή έχουμε την δυνατότητα να μετατρέψουμε όλους τους ήχους σε μονοφωνικούς.

Στη συνέχεια πρέπει να σιγουρευτούμε πως τα δεδομένα μας έχουν το ίδιο μήκος, το οποίο είναι ακριβώς τα 4 δευτερόλεπτα. Εφόσον τα δεδομένα έχουν δειγματοληπτηθεί στα 16 kHz θέλουμε λοιπόν τα δεδομένα μας να έχουν 64.000 δείγματα. Άρα σε όσα δεδομένα έχουν λιγότερα δείγματα, γεμίζουμε τις υπόλοιπες θέσεις με μηδενικά (zero padding).

Όπως έγινε αντιληπτό κατά την εκτέλεση των πειραμάτων τα δεδομένα φάνηκαν ανεπαρκή στο πλήθος τους, προκειμένου να εκπαιδευτεί ένα μοντέλο παραγωγής νέων ήχων με ικανοποιητικά αποτελέσματα. Έτσι κρίθηκε σκόπιμο να εφαρμοστεί η τεχνική του data augmentation, που αποτελεί μια διαδικασία τροποποίησης των ήδη υπαρχόντων δεδομένων με σκοπό να εμπλουτίσουμε το σύνολο δεδομένων με νέα δεδομένα. Στη δική μας περίπτωση αποφασίσαμε να εφαρμόσουμε τις εξής τροποποιήσεις στα ηχητικά κλιπ:

- Μετακίνηση στο χρόνο (roll),
- Αύξηση ταχύτητας σε 1.25 (time stretch),
- Μείωση ταχύτητας σε 0.75 (time stretch),
- Αλλαγή τόνου (pitch shift)

Με τον τρόπο αυτό καταφέραμε να τετραπλασιάσουμε τα δεδομένα μας από 8.732 σε 34.928. Η βελτίωση των αποτελεσμάτων χάρη σε αυτή την αύξηση των δεδομένων παρουσιάζεται στο Κεφάλαιο 5, στα αποτελέσματα των πειραμάτων μας. Θα μπορούσε να γίνει χρήση και Data Generator, με αποτέλεσμα να έχουμε συνεχώς νέα τροποποιημένα δεδομένα και αυτό μάλιστα αποτελεί μια από τις πιθανές βελτιώσεις όπως αυτές θα παρουσιαστούν στο Κεφάλαιο 6.

Τα δεδομένα μας είναι πλέον έτοιμα να τροφοδοτηθούν σε ένα νευρωνικό δίκτυο. Παρόλα αυτά η τροφοδοσία τους σε ένα δίκτυο με τη μορφή μιας ακολουθίας 64.000 αριθμητικών τιμών δεν είναι και η πιο βολική για τον σκοπό που θέλουμε να πετύχουμε. Έτσι προχωράμε στη μετατροπή των δεδομένων σε αναπαραστάσεις στο πεδίο της συχνότητας. Λεπτομέρειες πάνω σε αυτό παρουσιάζονται στην συνέχεια.

4.1.3 Τρόποι αναπαράστασης των δεδομένων

Όπως αναφέρθηκε και στις ενότητες 2.5 και 3.1 ένα σημαντικό στοιχείο της εργασίας ήταν η επιλογή της κατάλληλης αναπαράστασης των δεδομένων. Τελικώς, όπως φάνηκε από την βιβλιογραφία, η καταλληλότητα μιας αναπαράστασης είναι άμεσα συνδεδεμένη με την εργασία (task) που θέλουμε να φέρουμε εις πέρας. Μιας και το ζητούμενο που μελετάμε στην παρούσα εργασία και με το συγκεκριμένο θορυβώδες σύνολο δεδομένων δεν έχει ξαναμελετηθεί, καλούμαστε να αποφασίσουμε εμείς ποια είναι η αναπαράσταση των δεδομένων που θα δώσει τα καλύτερα αποτελέσματα. Για το σκοπό αυτό μετατρέψαμε το δεδομένα σε 3 διαφορετικές αναπαραστάσεις που έχουν δώσει καλά αποτελέσματα στην κατηγοριοποίηση και παραγωγή ήχου (φωνή ή μουσική) με τη χρήση νευρωνικών δικτύων. Οι αναπαραστάσεις αυτές είναι οι ίδιες που παρουσιάστηκαν και στην ενότητα 2.5, στα πλαίσια του θεωρητικού υπόβαθρου της ψηφιακής επεξεργασίας σήματος, δηλαδή:

- Φασματογράφημα (Spectrogram)
- Φασματογράφημα Mel (Mel-Spectrogram)
- MFCCs

Χάρη στην βοήθεια της βιβλιοθήκης librosa η εξαγωγή αυτών των αναπαραστάσεων είναι πολύ εύκολη, με απλή χρήση έτοιμων συναρτήσεων της βιβλιοθήκης και δεν απαιτεί τη χρήση των μαθηματικών εξισώσεων της ενότητας 2.5. Παρόλα αυτά επηρεάζει άμεσα το αποτέλεσμα η δική μας επιλογή ως προς τις παραμέτρους αυτών των συναρτήσεων. Πάμε να δούμε αναλυτικά τις επιλογές που κάναμε

Φασματογράφημα (Spectrogram)

Η εξαγωγή του STFT ενός σήματος ήχου είναι εφικτή μέσω της συνάρτησης “librosa.core.stft”. Η συγκεκριμένη συνάρτηση δέχεται πλήθος παραμέτρων με τις σημαντικότερες να αφορούν το μήκος του παραθύρου FFT (n_fft), το μήκος μετακίνησης του παραθύρου (hop_length) και τον τύπο του παραθύρου (window). Οι τιμές που εμείς επιλέξαμε για τις παραμέτρους αυτές είναι:

- n_fft = 1024
- hop_length = n_fft / 2 = 512
- window = ‘hann’

Γενικά από την αρχή της αβεβαιότητας του μετασχηματισμού Fourier γνωρίζουμε ότι μπορούμε να έχουμε καλύτερη ακρίβεια στη συχνότητα ή καλύτερη ακρίβεια στο χρόνο, αλλά όχι και τα δύο μαζί. Βάση της βιβλιογραφίας πάνω στο συγκεκριμένο σύνολο δεδομένων, αλλά και λόγω των δικτύων που μπορούμε να υλοποιήσουμε με τους υπολογιστικούς πόρους που έχουμε στην διάθεση μας, επιλέξαμε αυτές τις παραμέτρους. Η έξοδος της συνάρτησης "librosa.core.stft" είναι ένας πίνακας numpy δύο διαστάσεων που περιλαμβάνει μιγαδικούς αριθμούς. Προκείμενου να δημιουργήσουμε μια κατάλληλη είσοδο για το δίκτυο μας παίρνουμε το πλάτος (magnitude) του STFT και μετατρέπουμε τις μονάδες σε decibel (dB). Αυτό το επιτύχαμε κάνοντας χρήση των συναρτήσεων "numpy.abs" και "librosa.amplitude_to_db" αντίστοιχα. Στη συνέχεια καλούμαστε να κανονικοποιήσουμε το φασματογράφημα σε ένα εύρος τιμών που θα είναι κατάλληλο για τροφοδοσία σε ένα νευρωνικό δίκτυο. Η συνάρτηση "librosa.amplitude_to_db" παίρνει το μέτρο του STFT και το μετάτρεπε σε decibel (dB), με τιμή αναφοράς το 0 και ελάχιστη τιμή το -80. Εφόσον όλες οι τιμές βρίσκονται πλέον στο διάστημα $[-80,0]$ είναι αρκετά εύκολα αν τις μετασχηματίσουμε ώστε να ανήκουν στο διάσημα $[-1,1]$ μέσω της εξίσωσης:

$$2\frac{(x_i + 80)}{80} - 1$$

Όπου x_i κάθε τιμή του φασματογραφήματος.

Φασματογράφημα Mel (Mel-Spectrogram)

Προκειμένου να λάβουμε το φασματογράφημα Mel και να το φέρουμε σε μορφή κατάλληλη ώστε να γίνει η είσοδος στα δίκτυα που κατασκευάσαμε, ακολουθήσαμε την διαδικασία που περιγράφεται στη συνέχεια. Αρχικά κάναμε χρήση της έτοιμης συνάρτησης που προσφέρει η βιβλιοθήκη librosa, "librosa.feature.melspectrogram" προκειμένου να λάβουμε το φασματογράφημα Mel ενός ηχητικού σήματος. Οι παράμετροι της συνάρτησης που δώσαμε είναι οι εξής:

- n_fft = 1024
- n_mels = 128

Στη συνέχεια μετατρέπουμε το φάσμα που προκύπτει σε μονάδες decibel (dB) μέσω της έτοιμης συνάρτησης "librosa.power_to_db". Η συνάρτηση αυτή θεωρεί ως σημείο αναφοράς την μέγιστη τιμή του φασματογραφήματος mel και άρα την αντιστοιχίζει στα $0\,\text{dB}$, με όλες τις υπόλοιπες τιμές να αντιστοιχίζονται σε αρνητικές τιμές dB. Ύστερα από ανάλυση και παρατήρηση των δεδομένων παρατηρήσαμε πως ελάχιστες τιμές των φασμάτων βρίσκονται κάτω από τα -70dB ($< 1\,\%$) και έτσι μπορούμε να τις αποκόψουμε χωρίς να χάσουμε σημαντική πληροφορία. Όλες οι τιμές βρίσκονται πλέον στο διάστημα $[-70,0]$. Τέλος για να κανονικοποιήσουμε στο διάστημα $[-1,1]$ χρησιμοποιούμε την εξίσωση:

$$2\frac{(x_i + 70)}{70} - 1$$

Όπου x_i κάθε τιμή του φασματογραφήματος Mel.

MFCCs

Σε αυτή την περίπτωση κάνουμε χρήση της συνάρτησης “librosa.feature.mfcc”. Οι παράμετροι οι οποίες δέχεται η συγκεκριμένη συνάρτηση αφορούν τον ρυθμό δειγματοληψίας και το πλήθος των MFCCs. Η επιλογή που εμείς κάναμε για τον πλήθος των MFCCs είναι να κρατήσουμε τους πρώτους 40 συντελεστές. Ενδεχομένως αυτή η επιλογή να είναι υπερβολική, αλλά η διαθέσιμη μνήμη και επεξεργαστική ισχύς μας επιτρέπει να κάνουμε αυτή την επιλογή χωρίς κάποιο αντάλλαγμα σε απόδοση.

Για την κανονικοποίηση των MFCCs επιλέγουμε να κανονικοποιήσουμε τα δεδομένα γύρω από το μηδέν και με τυπική απόκλιση ίση με τη μονάδα.

4.2 Διαθέσιμοι υπολογιστικοί πόροι

Ο κώδικας της εργασίας είναι εξ ολοκλήρου γραμμένος σε Python (3.7) και κάνει χρήση τη βιβλιοθήκης Tensorflow 2, καθώς και της ενσωματωμένης έκδοσης του Keras API που παρέχεται στο πακέτο του TensorFlow 2. Δεδομένου ότι χρησιμοποιούμε βαθιά νευρωνικά δίκτυα, οι υπολογιστικοί πόροι που χρειαζόμαστε είναι αναγκαίο να περιλαμβάνουν μια τουλάχιστον GPU (κάρτα γραφικών). Επίσης έχουμε ανάγκη και για περίπου 15 Gb μνήμης, η οποία είναι και η μνήμη που καταλαμβάνουν τα δεδομένα στην μορφή κανονικοποιημένου STFT.

Κατά την έναρξη της εκπόνησης της παρούσας εργασίας (Δεκέμβριος 2019) το TensorFlow 2 είχε μόλις κυκλοφορήσει. Οι δυνατότητες του καθώς και η ευχρηστία του ήταν κατά πολύ βελτιωμένη σε σχέση με τις προηγούμενες εκδόσεις. Ακόμη, όπως συμβαίνει με τις νέες τεχνολογίες, είχε ξεκινήσει μια διαδικασία μετάβασης και οι χρήστες παρακινούνταν να χρησιμοποιούν το TensorFlow 2 για κάθε νέα εργασία. Έτσι έγινε η επιλογή να χρησιμοποιηθεί και στην παρούσα εργασία. Παρόλα αυτά το εργαστήριο “Τεχνητής νοημοσύνης και συστημάτων μάθησης” (AILS) του Εθνικού Μέτσοβου Πολυτεχνείου, λόγω της νεότητας του TensorFlow 2, δεν είχε αναβαθμισμένους τις GPUs των διακομιστών (GPU drivers, CUDA Toolkit, cuDNN) και χωρίς δικαιώματα root ήταν αδύνατο να γίνουν οι απαραίτητες αλλαγές. Έτσι η εναλλακτική λύση ήταν η χρήση κάποιας διαδικτυακής πλατφόρμας όπως το Google Colab ή το Kaggle. Οι συγκεκριμένες πλατφόρμες προσφέρουν αρκετά ικανές GPUs (Tesla K80 και TeslaP100 αντίστοιχα). Η τελική επιλογή για την εκτέλεση των πειραμάτων ήταν το περιβάλλον του Kaggle. Έτσι οι τελικοί διαθέσιμοι πόροι διαμορφώθηκαν ως εξής:

- CPU: 2-core Intel Xeon
- RAM: 13Gb
- GPU : 1 × Tesla P100 16Gb (~40hrs/εβδομάδα)
- HDD: 20Gb

Πέρα από τον περιορισμό στον χρόνο χρήσης της GPU, ο οποίος δεν ήταν τόσο σοβαρό εμπόδιο αφού τα πειράματα μας είχαν διάρκεια το πολύ κάποιες ώρες, ο πιο σημαντικός περιορισμός ήταν η διαθέσιμη μνήμη, καθώς όπως αναφέραμε τα δεδομένα στη μορφή STFT απαιτούσαν για την αποθήκευση τους ~15Gb. Ακόμη η ευελιξία που μας προσφέρει το περιβάλλον του Kaggle σε ό,τι αφορά τη διαχείριση των δεδομένων στη μνήμη είναι αρκετά περιορισμένη. Ο τρόπος με τον οποίο αντιμετωπίστηκε αυτό το εμπόδιο περιγράφεται στην επόμενη ενότητα (CVAE + STFT).

4.3 Πειραματικές διατάξεις

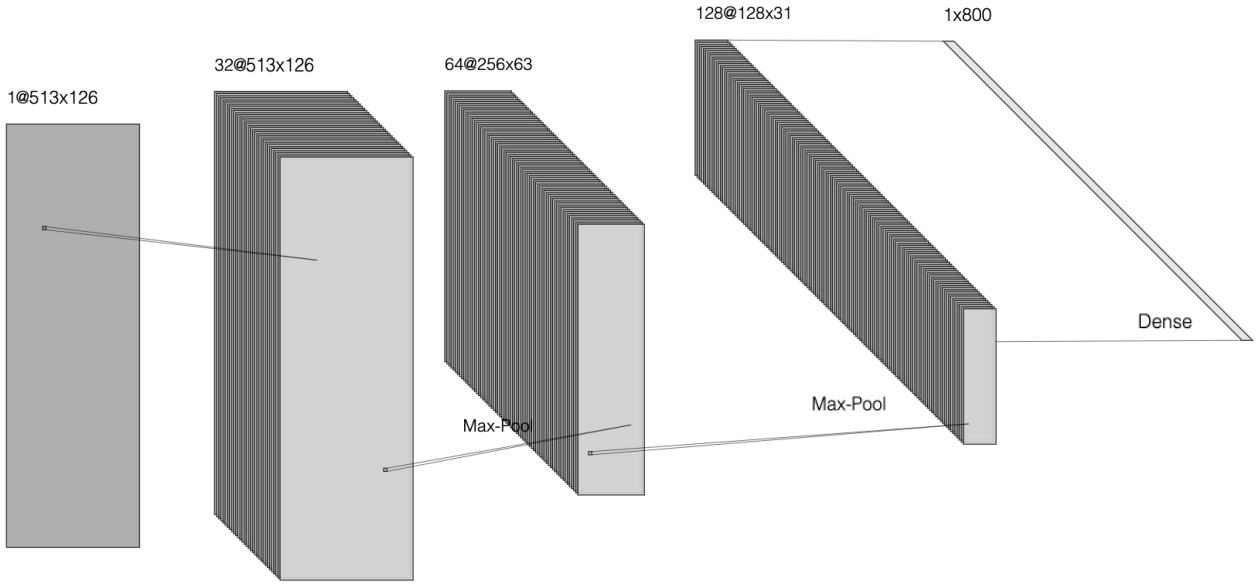
Οι πειραματικές διατάξεις που υλοποιήθηκαν χωρίζονται σε δύο κατηγορίες. Αρχικά έχουμε τις αρχιτεκτονικές που έκαναν χρήση του CVAE, όπως αυτός παρουσιάστηκε στην ενότητα 3.2. Στη δεύτερη κατηγορία κάναμε χρήση ενός πλήρους συνδεδεμένου VAE. Σε αυτή την περίπτωση το μοντέλο υλοποιήθηκε μόνο για είσοδο που αποτελείται από MFCCs. Ο λόγος για αυτή την επιλογή είναι πως η χρήση πλήρως συνδεδεμένου δικτύου με είσοδο STFTs είχε τεράστιες απαιτήσεις σε μνήμη και υπολογιστική ισχύ, καθώς το δίκτυο θα αποτελούνταν από εκατοντάδες εκατομμύρια παραμέτρους. Στη συνέχεια παρουσιάζονται πιο αναλυτικά οι αρχιτεκτονικές οι οποίες υλοποιήθηκαν.

CVAE + Spectrogram

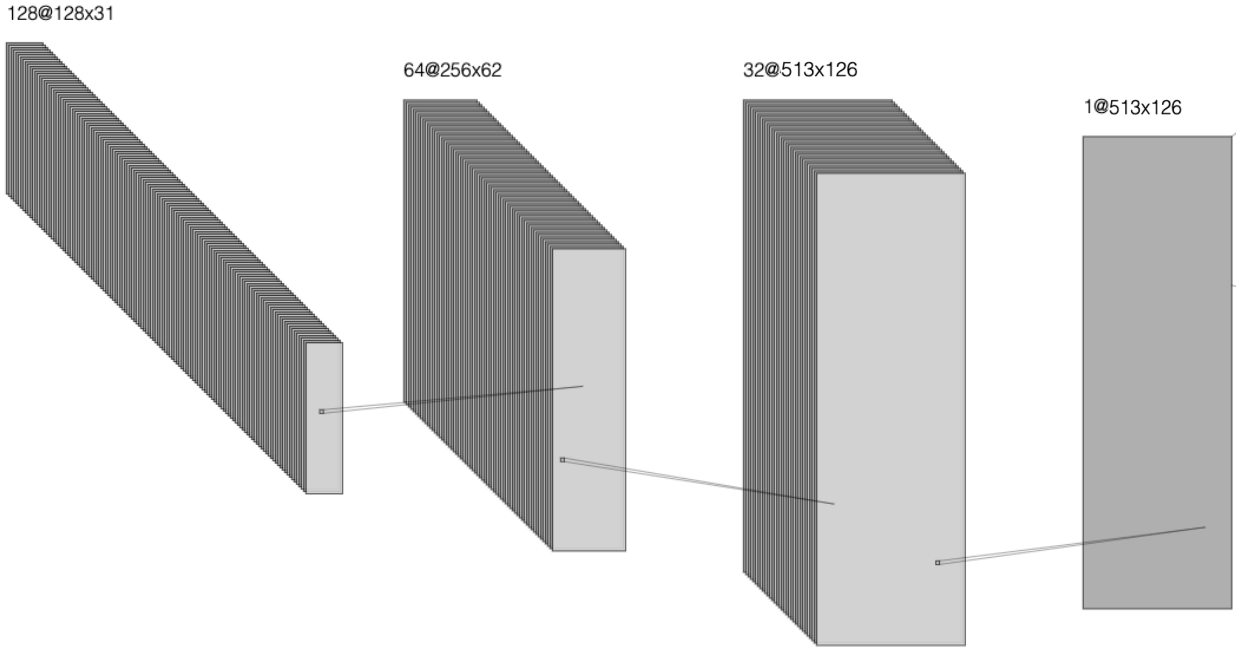
Το πρώτο μοντέλο που εξετάζουμε αποτελεί και το πιο ογκώδες και απαιτητικό μοντέλο σε ότι αφορά την απαιτούμενη μνήμη. Το μέγεθος των δεδομένων σε συνδυασμό με το ότι η εκπαίδευση και ο έλεγχος του μοντέλου έγιναν στην διαδικτυακή πλατφόρμα του Kaggle (βλ. ενότητα 4.2), μας οδήγησε στο να προβούμε σε έναν αναγκαστικό συμβιβασμό και να μειώσουμε τα δεδομένα μας έτσι ώστε να χωρούν στη μνήμη που προσφέρει το Kaggle. Για το λόγο αυτό επιλέξαμε να μην χρησιμοποιήσουμε τα δεδομένα που παράχθηκαν από την διαδικασία του data augmentation και να πραγματοποιήσουμε την εκπαίδευση και την αξιολόγηση του μοντέλου κάνοντας χρήση των αρχικών 8732 αρχείων ήχου.

Στη συνέχεια παρουσιάζουμε την τελική αρχιτεκτονική του CVAE δικτύου που επιλέξαμε ως την βέλτιστη ύστερα από δοκιμές διαφορετικών υπερπαραμέτρων (Σχ. 4.5). Αρχικά δίνεται ένα κανονικοποιημένο φασματογράφημα διαστάσεων 513×126. Στη συνέχεια ακολουθεί μια σειρά από 3 συνελικτικά στρώματα τα οποία αποτελούνται από 32, 64 και 128 φίλτρα αντίστοιχα. Το μέγεθος του δεκτικού πεδίου κάθε συνελικτικού στρώματος είναι 3×3 και το βήμα ίσο με 1. Μεταξύ των συνελικτικών στρωμάτων παρεμβάλλονται στρώματα συγκέντρωσης μεγίστου με δεκτικό πεδίο μεγέθους 2×2, κάτι που σημαίνει πως το μέγεθος της εισόδου κάθε φορά υποδιπλασιάζεται. Τέλος, ύστερα από τα συνελικτικά στρώματα ο κωδικοποιητής περιλαμβάνει ένα πλήρως συνδεδεμένο στρώμα μεγέθους $400 + 400$, το οποίο περιγράφει την μέση τιμή και τον λογάριθμο της διακύμανσης αντίστοιχα, για 400 χαρακτηριστικά της εισόδου. Η συνάρτηση ενεργοποίησης που χρησιμοποιήθηκε για τα συνελικτικά στρώμα τόσο του κωδικοποιητή όσο και του αποκωδικοποιητή, είναι η συνάρτηση selu.

Ο αποκωδικοποιητής του CVAE είναι το κομμάτι του νευρωνικού μας δικτύου το οποίο είναι υπεύθυνο για την παραγωγή νέων δεδομένων. Στην αρχιτεκτονική ενός VAE, ο αποκωδικοποιητής αποτελεί έναν καθρεφτισμό του κωδικοποιητή, με μικρές διαφοροποιήσεις. Αρχικά έχουμε ένα πλήρως συνδεδεμένο στρώμα μεγέθους $128 \cdot 31 \cdot 128$. Το στρώμα αυτό δέχεται την είσοδο από την ενδιάμεση αναπαράσταση μεγέθους $1 \times (400 + 400)$ και την τροφοδοτεί σε ένα στρώμα που την ανασχηματίζει ώστε να έχει διαστάσεις 128×31. Στη συνέχεια, ο αποκωδικοποιητής αποτελείται από 3 στρώματα αποσυνέλιξης, τα οποία αποτελούνται από 128, 64 και 32 φίλτρα αντίστοιχα. Το δεκτικό πεδίο (πυρήνας, kernel) μετακινείται με βήμα (stride) ίσο με 2 και έτσι το μέγεθος της εισόδου ενός στρώματος αποσυνέλιξης διπλασιάζεται στην έξοδο του. Η έξοδος του αποκωδικοποιητή είναι ένα νέο κανονικοποιημένο φασματογράφημα Mel με διστάσεις ίδιες με τις εισόδου του κωδικοποιητή, δηλαδή 513×126.



(α)



(β)

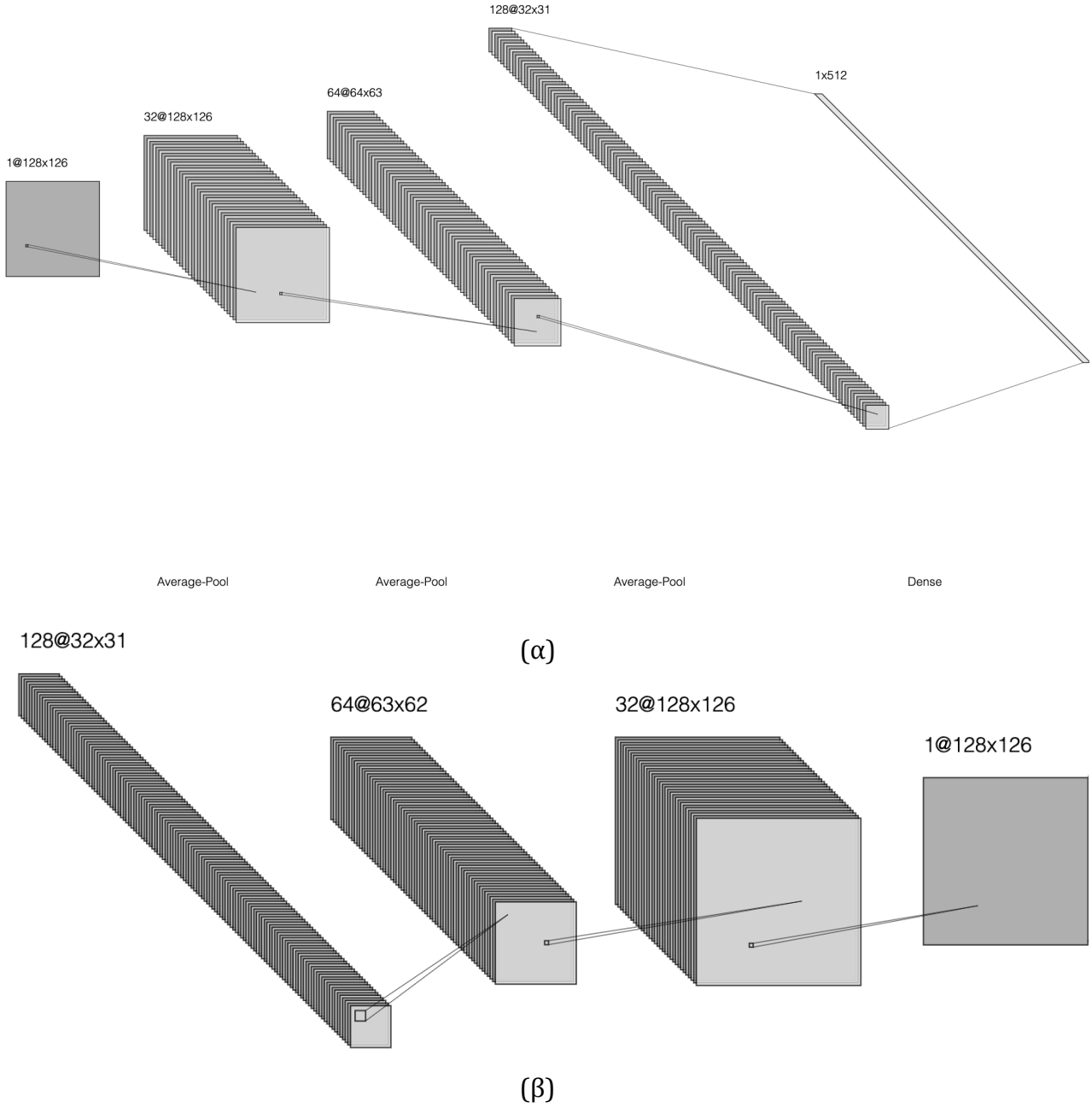
Σχήμα 4.5: Δίκτυο CVAE με είσοδο Spectrogram.
(α) Κωδικοποιητής. (β) Αποκωδικοποιητής.

Η συνάρτηση κόστους χρησιμοποιήθηκε που είναι αυτή που παρουσιάστηκε στην ενότητα 2.3.2. Η τιμή της υπερπαραμέτρου στάθμισης του σφάλματος ανακατασκευής, R_LOSS_FACTOR την επιλέξαμε ίση με 5000. Ο αλγόριθμος βελτιστοποίησης που χρησιμοποιήθηκε είναι ο αλγόριθμος adam, με ρυθμό εκμάθησης ίσο με 10^{-6}. Το batch size επιλέχθηκε ίσο με 128.

CVAE + Mel-Spectrogram

Το μέγεθος των δεδομένων στη μορφή του κανονικοποιημένου φασματογραφήματος Mel, μας καθιστά επιτρεπτό να τα χρησιμοποιήσουμε όλα τα για την εκπαίδευση του μοντέλου μας. Στη συνέχεια παρουσιάζουμε την αρχιτεκτονική του CVAE, η οποία μας έδωσε τα καλύτερα ακουστικά αποτελέσματα (Σχ. 4.6), ύστερα από δοκιμή αρκετών διαφορετικών τιμών για τις διαθέσιμες υπερπαραμέτρους.

Αρχικά δίνουμε για είσοδο κανονικοποιημένα φασματογραφήματα Mel μεγέθους 128×126. Η είσοδος αυτή στη συνέχεια τροφοδοτείται σε ένα συνελικτικό στρώμα με μέγεθος πυρήνα 3×3. Προσθέτουμε πως για όλα τα επόμενα συνελικτικά στρώματα το μέγεθος πυρήνα είναι το ίδιο και ίσο με 3×3. Ο κωδικοποιητής αποτελείται από 3 συνελικτικά στρώματα τα οποία αποτελούνται



Σχήμα 4.6: Δίκτυο CVAE με είσοδο Mel-Spectrogram.
(α) Κωδικοποιητής. (β) Αποκωδικοποιητής.

από 32, 64 και 128 φίλτρα αντίστοιχα. Μεταξύ των συνελικτικών στρωμάτων παρεμβάλλονται στρώματα συγκέντρωσης μέσου όρου με δεκτικό πεδίο μεγέθους 2×2, κάτι που σημαίνει πως το μέγεθος της εισόδου κάθε φορά υποδιπλασιάζεται (για να βρούμε το ακριβές μέγεθος κάθε επόμενης εισόδου συνυπολογίζουμε και το padding που εφαρμόζεται). Ύστερα από τα συνελικτικά στρώματα ο κωδικοποιητής περιλαμβάνει ένα πλήρως συνδεδεμένο στρώμα μεγέθους $256 + 256$ το οποίο περιγράφει την μέση τιμή και την διακύμανση αντίστοιχα. Η συνάρτηση ενεργοποίησης που χρησιμοποιήθηκε για τα συνελικτικά στρώμα τόσο του κωδικοποιητή όσο και του αποκωδικοποιητή, είναι η συνάρτηση selu.

Ο αποκωδικοποιητής αποτελείται από ένα πλήρως συνδεδεμένο στρώμα αρχικά μεγέθους $32 \cdot 31 \cdot 128$ το οποίο στη συνέχεια ανασχηματίζεται ώστε να έχουμε βάθος 128, ύψος 32 και πλάτος 31. Αυτή η είσοδος τροφοδοτείται σε μία σειρά στρωμάτων αποσυνέλιξης με 128, 64 και 32 φίλτρα αντίστοιχα. Ο πυρήνας μετατοπίζεται με βήμα 2 άρα το μέγεθος της εισόδου ενός στρώματος αποσυνέλιξης διπλασιάζεται στην έξοδο. Η έξοδος του αποκωδικοποιητή είναι ένα νέο κανονικοποιημένο φασματογράφημα Mel με διστάσεις ίδιες με τις εισόδου του κωδικοποιητή, δηλαδή 128×126.

Σαν συνάρτηση χρησιμοποιήθηκε η συνάρτηση κόστους του VAE που παρουσιάστηκε στην ενότητα $2.3.2$. Η τιμή της υπερπαραμέτρου στάθμισης του σφάλματος ανακατασκευής, R_LOSS_FACTOR την επιλέξαμε ίση με 5000. Ο αλγόριθμος βελτιστοποίησης που χρησιμοποιήθηκε είναι ο αλγόριθμος adam, με ρυθμό εκμάθησης ίσο με 10^{-5}. Το batch size ήταν ίσο με 128.

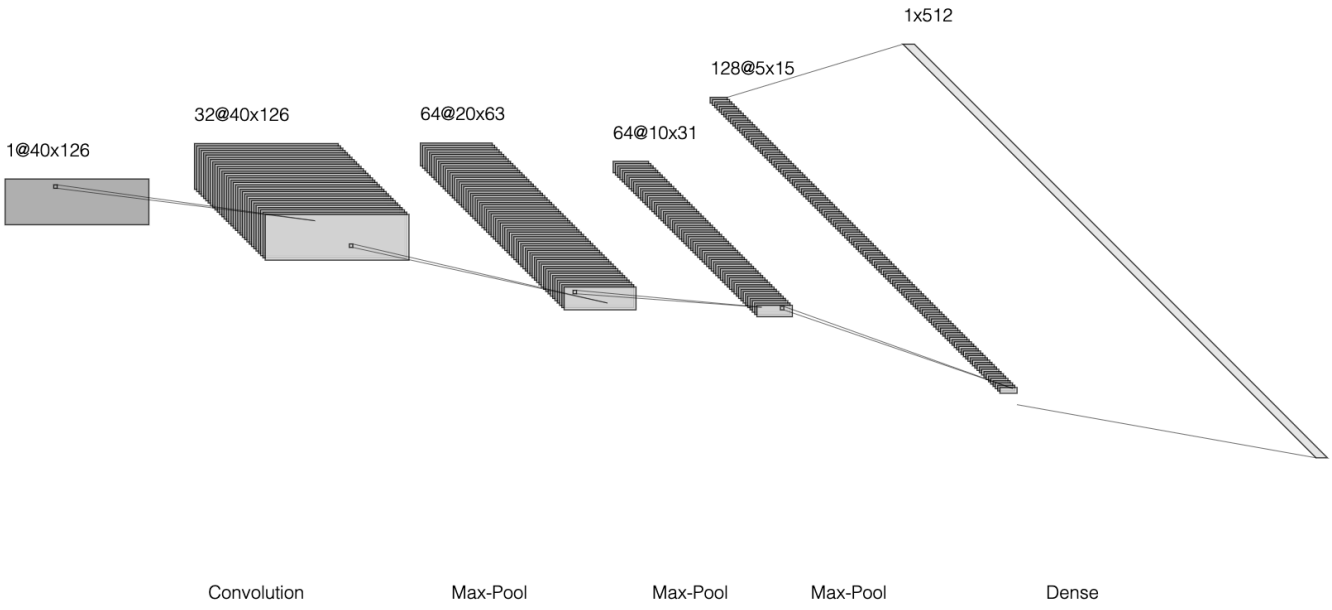
CVAE + MFCCs

Στην περίπτωση αυτή, όπως και στην προηγούμενη, ήταν δυνατό να αξιοποιήσομε όλα τα δεδομένα που προέκυψαν από την διαδικασία του data augmentation, μιας και το συνολικό μέγεθος τους ήταν κατά πολύ μικρότερο από ότι στις προηγούμενες περιπτώσεις. Το μοντέλο που θα παρουσιάσουμε στην συνέχεια προέκυψε ύστερα από μεγάλο πλήθος δοκιμών διαφορετικών παραμέτρων για τα στρώματα του CVAE. Τελικώς ο συνδυασμός των παραμέτρων που παρουσιάζουμε φάνηκε να δίνει τα βέλτιστα αποτελέσματα για αυτή την αναπαράσταση των δεδομένων. Στις εικόνες που ακολουθούν παρουσιάζουμε σχηματικά τον κωδικοποιητή και τον αποκωδικοποιητή της εν λόγω αναπαράστασης και στη συνέχεια προχωράμε σε μια πιο αναλυτική παρουσίαση των παραμέτρων τους.

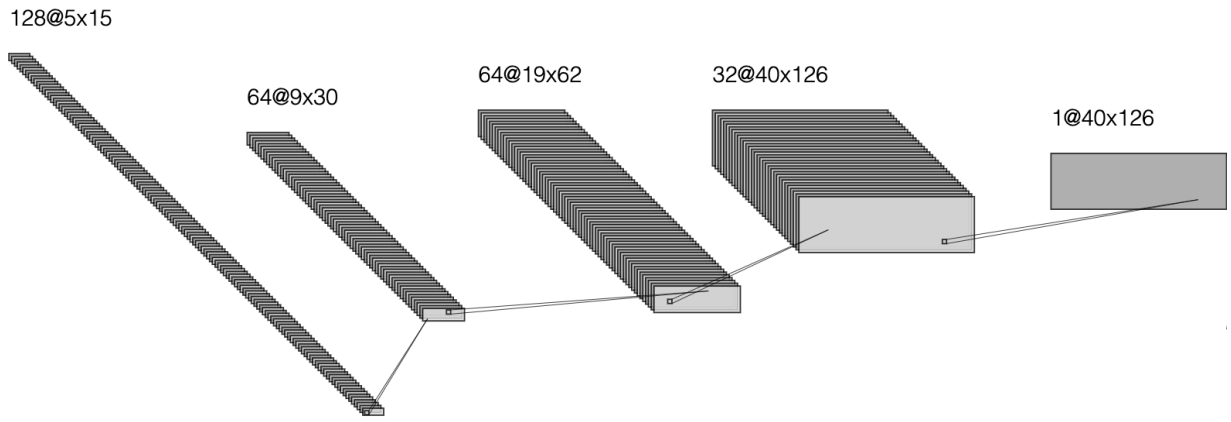
Όπως βλέπουμε (Σχ. 4.7) ξεκινάμε με μια είσοδο μεγέθους 40×126 η οποία στη συνέχεια τροφοδοτείται σε ένα συνελικτικό στρώμα με μέγεθος πυρήνα 3×3. Για τα στρώματα που ακολουθούν δεν αναφέρουμε το μέγεθος του πυρήνα, διότι για όλα το μέγεθος είναι το ίδιο δηλαδή 3×3. Αυτή η επιλογή έγινε αφού δοκιμάστηκαν και διάφορα άλλα μεγέθη πυρήνα χωρίς να διαφοροποιούν σημαντικά το αποτέλεσμα. Το πλήθος των φίλτρων που διαθέτει κάθε συνελικτικό στρώμα φαίνεται στο σχήμα 4.7. Στον κωδικοποιητή μεταξύ Η συνάρτηση ενεργοποίησης που χρησιμοποιήθηκε για τα συνελικτικά στρώμα τόσο του κωδικοποιητή όσο και του αποκωδικοποιητή, είναι η συνάρτηση selu. Δοκιμάστηκε και η συνάρτηση relu, η οποία όμως τελικώς είχε παρόμοια αποτελέσματα. Εκτός από τα συνελικτικά στρώματα, έχουμε και δύο πλήρως συνδεδεμένα στρώματα. Το ένα αποτελεί το τελευταίο στρώμα του κωδικοποιητή, έχει μέγεθος 512 και αποτελείται ουσιαστικά από δυο διανύσματα μεγέθους 256, τα οποία αναπαριστούν τις μέσες τιμές και τον λογάριθμο των διακυμάνσεων των διάφορων χαρακτηριστικών που κωδικοποιούνται στην ενδιάμεση κωδικοποίηση της αρχιτεκτονικής. Το

άλλο πλήρως συνδεδεμένο στρώμα αποτελεί την είσοδο του αποκωδικοποιητή η οποία ανασχηματίζεται προκειμένου να τροφοδοτηθεί στα στρώματα αποσυνέλιξης.

Η συνάρτηση κόστους που χρησιμοποιήθηκε είναι η γνωστή συνάρτηση κόστους του VAE, όπως αυτή έχει παρουσιαστεί, με την της παραέτρου R_LOSS_FACTOR να είναι ίση με 1000. Ο αλγόριθμος βελτιστοποίησης που χρησιμοποιήθηκε είναι ο αλγόριθμος adam, με ρυθμό εκμάθησης ίσο με 10^{-4}. Το μοντέλο εκπαιδεύτηκε για 150 εποχές με batch size 128.

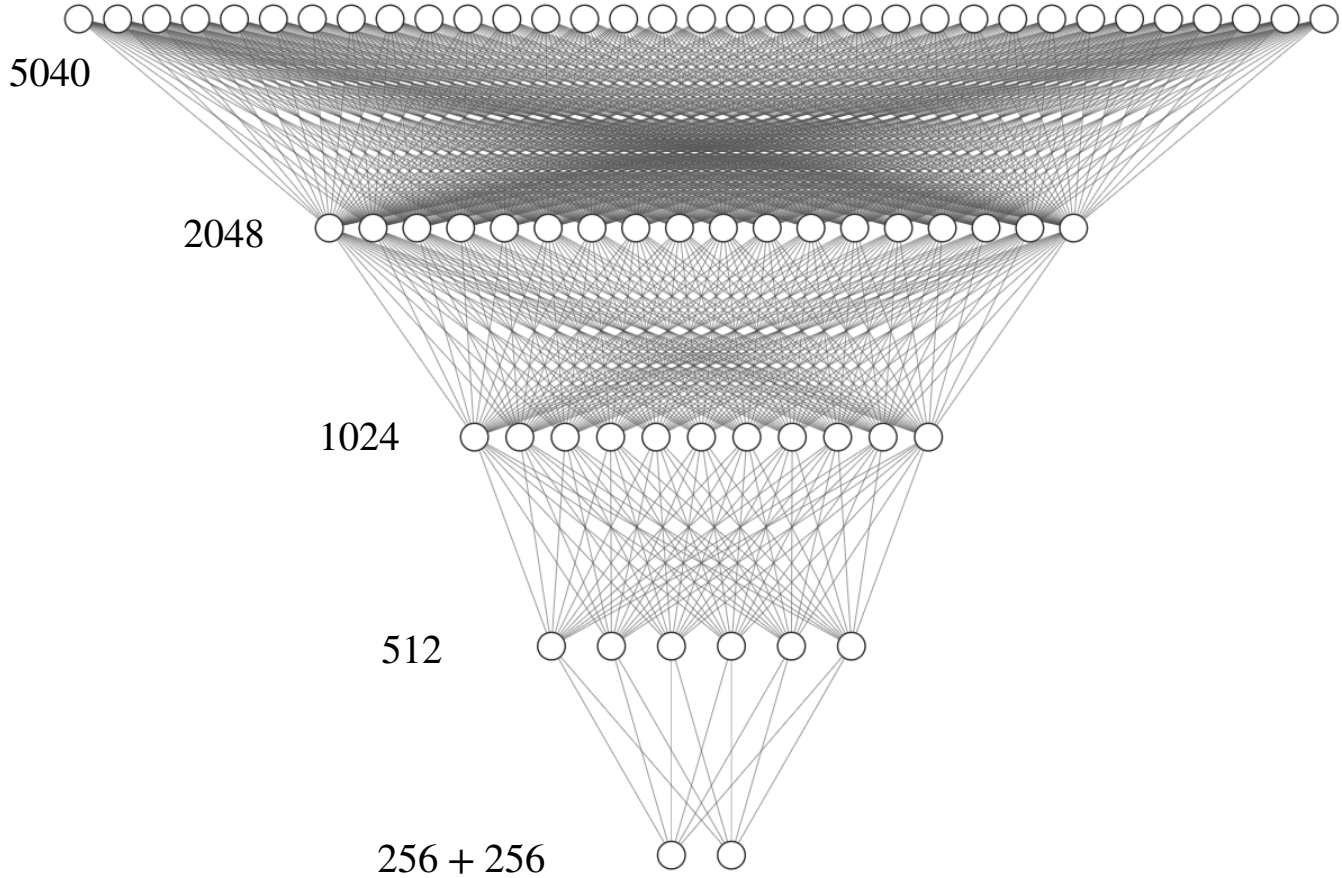


(α)

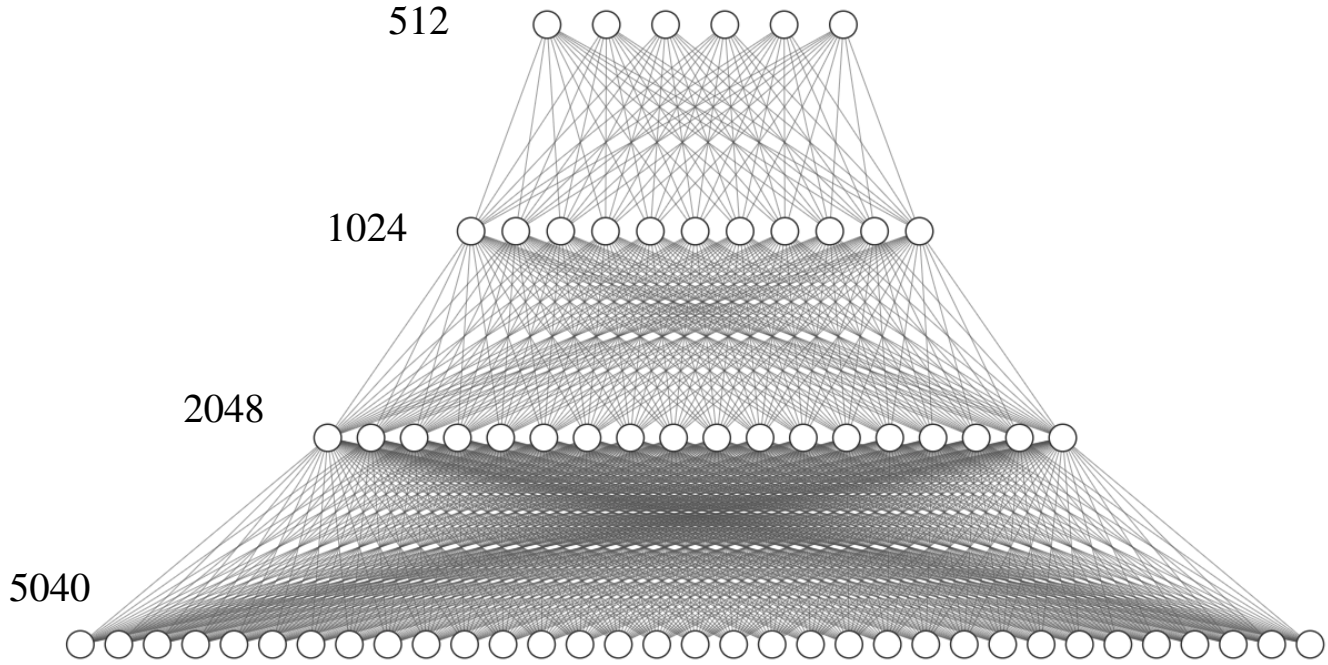


(β)

Σχήμα 4.7: Δίκτυο CVAE με είσοδο MFCCs.
(α) Κωδικοποιητής. (β) Αποκωδικοποιητής.



(α)



(β)

Σχήμα 4.8: Δίκτυο FC-VAE με είσοδο MFCCs.
(α) Κωδικοποιητής. (β) Αποκωδικοποιητής.

Πλήρως συνδεδεμένος (FC) VAE + MFCCs

Το μέγεθος της εισόδου όταν αναπαριστούμε τον ήχο σαν MFCCs μας επιτρέπει να δοκιμάσουμε έναν VAE στον οποίο ο κωδικοποιητής και ο αποκωδικοποιητής αποτελούν πλήρως συνδεδεμένα δίκτυα. Η αρχιτεκτονική δικτύου στην οποία καταλήξαμε και παρουσιάζεται στη συνέχεια (Σχ. 4.8), προέκυψε ύστερα από δοκιμές διαφορετικών τιμών σε ότι αφορά το πλήθος των στρωμάτων, το πλήθος των νευρώνων κάθε στρώματος και το μέγεθος της ενδιάμεσης αναπαράστασης. Η τελική αρχιτεκτονική φαίνεται να παράγει τα βέλτιστα πιθανά αποτελέσματα που μπορεί να προσφέρει ένα πλήρως συνδεδεμένο δίκτυο για το πρόβλημα που εξετάζουμε και τα δεδομένα που έχουμε στη διάθεση μας. Περισσότερα σχετικά με την απόδοση του μοντέλου (όπως και των μοντέλων που παρουσιάστηκαν προηγουμένως) αναλύονται στο επόμενο κεφάλαιο.

Το τελικό μοντέλο αποτελείται από έναν κωδικοποιητή και έναν αποκωδικοποιητή, τεσσάρων στρωμάτων ο καθένας. Η ενδιάμεση αναπαράσταση έχει μέγεθος 256×2 (512), όπως και στην περίπτωση CVAE + MFCCs. Μεγαλύτερες τιμές της ενδιάμεσης αναπαράστασης δεν φάνηκε να προσφέρει καλύτερα αποτελέσματα. Οι συνάρτηση ενεργοποίησης για όλα τα στρώματα εκτός από το ενδιάμεσο στρώμα το VAE και την έξοδο του αποκωδικοποιητή είναι η μη γραμμική συνάρτηση selu. Η συνάρτηση relu δοκιμάστηκε αλλά ήταν αδύνατο να εκπαιδευτεί το μοντέλο και κολλούσε σε πολύ υψηλές τιμές της συνάρτησης κόστους.

Οι υπόλοιπες παράμετροι του μοντέλου είναι ίδιες με την περίπτωση CVAE + MFCCs.

4.3 Ανακατασκευή φάσης

Ένα σημαντικό κομμάτι της εργασίας είναι η ανακατασκευή της εξόδου των νευρικών μας δικτύων, ώστε να παραχθεί κάποιος ήχος. Σε όλα τα πειράματα που πραγματοποιήσαμε, προκειμένου να έχουμε είσοδο στα δίκτυα μας που αποτελείται μόνο από πραγματικούς αριθμούς, αγνοήσαμε την φάση και κρατήσαμε μόνο το πλάτος του φάσματος κάθε ήχου. Έτσι κατά την διάρκεια της παραγωγής νέων ήχων κληθήκαμε να βρούμε έναν τρόπο ώστε να ανακατασκευάσουμε την φάση για ένα παραχθέν φάσμα. Ο τρόπος που επιλέξαμε να το κάνουμε αυτό είναι κάνοντας χρήση του γνωστότερου ίσως αλγορίθμου για την ανακατασκευή φάσης, ο οποίος είναι γνωστός ως αλγόριθμος Griffin-Lim.

Ο αλγόριθμος Griffin-Lim αρχικά δημιουργεί έναν μιγαδικό πίνακα και προσθέτει σε αυτόν το πλάτος του φάσματος ως το πραγματικό στρώμα. Στη συνέχεια τοποθετεί ως φανταστικό στρώμα ομοιόμορφο θόρυβο. Από αυτό το σημείο ξεκινά μια επαναληπτική διαδικασία με σκοπό την ανακατασκευή της φάσης του ηχητικού σήματος. Με είσοδο αυτόν τον μιγαδικό πίνακα υπολογίζεται ο αντίστροφος STFT και το αποτέλεσμα είναι μια χρονοσειρά. Πάνω σε αυτή τη χρονοσειρά πραγματοποιείται ο ευθύς μετασχηματισμός STFT. Σε αυτό το σημείο αντικαθιστούμε το πραγματικό στρώμα με το αρχικό πλάτος του φάσματος και με αυτόν τον τρόπο εξάγουμε λίγη πληροφορία σχετικά με την φάση του σήματος. Η διαδικασία αυτή επαναλαμβάνεται συνήθως εκατοντάδες ή και χιλιάδες φορές μέχρι το αποτέλεσμα να είναι ικανοποιητικό.

Γενικά ο αλγόριθμος Griffin-Lim είναι αρκετά αξιόπιστος και γρήγορος. Από την δημοσίευση του το 1984 έχουν δημοσιευτεί αρκετές εναλλακτικές του, με κύριο σκοπό την περαιτέρω επιτάχυνση του και την βελτίωση της αποδοτικότητας του. Η βιβλιοθήκη librosa έχει έτοιμη μια γρήγορη υλοποίηση του αλγορίθμου Griffin-Lim που δημοσιεύτηκε το 2013.

5

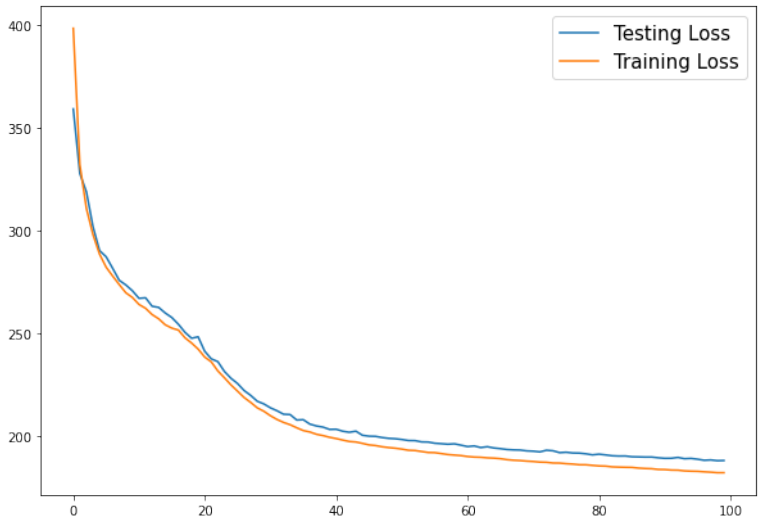
5. Αποτελέσματα

5.1 Αποτελέσματα πειραμάτων

CVAE + Spectrogram

Προτού παρουσιάσουμε τα αποτελέσματα αυτού του μοντέλου, είναι αναγκαίο να επαναλάβουμε πως λόγω των υψηλών απαιτήσεων υλικού από το μοντέλο, δεν ήμασταν σε θέση να αξιοποιήσουμε όλα τα διαθέσιμα δεδομένα για την εκπαίδευση και την αξιολόγηση του. Έτσι η σύγκριση του με τα άλλα μοντέλα δεν γίνεται κάτω από ίδιες συνθήκες. Παρόλα αυτά, το γεγονός ότι αυτό το συγκεκριμένο μοντέλο δεν είχε καθόλου καλά αποτελέσματα μας επιτρέπει να το μελετήσουμε και να το κατατάξουμε έναντι των άλλων χωρίς μεγάλη δυσκολία.

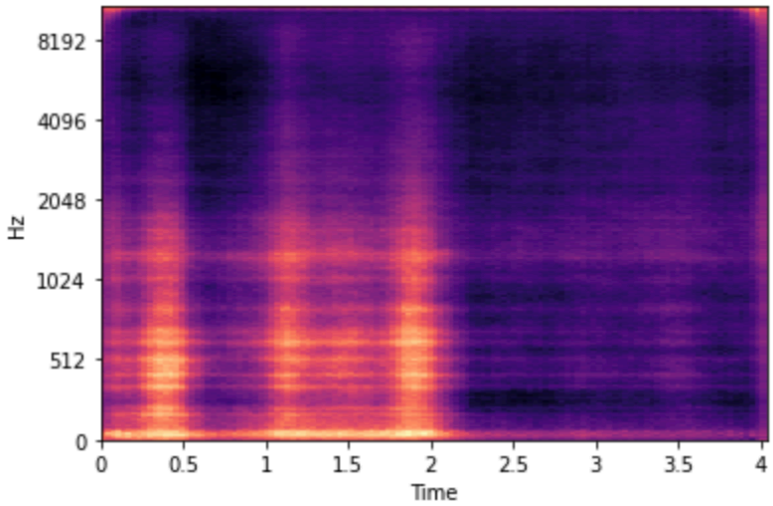
Αρχικά το μοντέλο εκπαιδεύτηκε για 100 εποχές και είχε τελική τιμή συνάρτησης κόστους ίση με 187 (Σχ. 5.1). Αυτή η πληροφορία από μόνη της δεν φανερώνει κάποιο στοιχείο, αφού η τιμή



Σχήμα 5.1: Συνάρτηση κόστους για "CVAE + Spectrogram" με όλα τα δεδομένα διαθέσιμα.

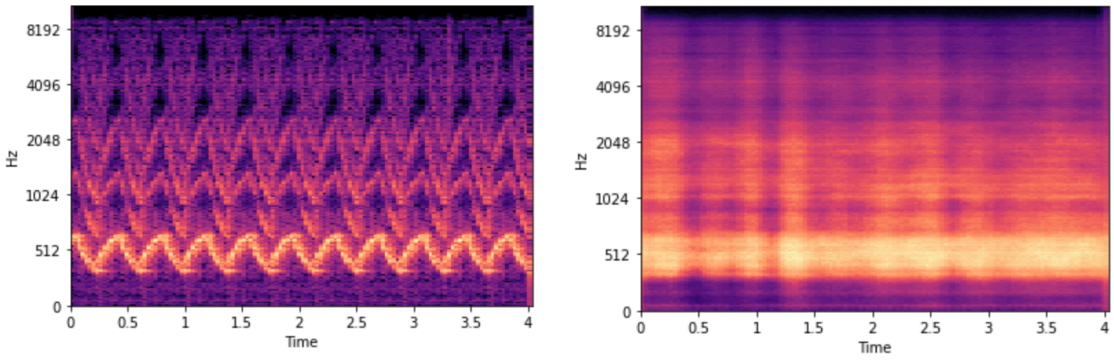
της συνάρτησης κόστους εξαρτάται και από άλλους παράγοντες όπως η μεταβλητή στάθμισης σφάλματος ανακατασκευής R_LOSS_FACTOR. Έτσι είναι αναγκαίο να ελέγξουμε τις δυνατότητες του μοντέλου στην πράξη.

Αρχικά, κατά την διαδικασία παράγωγης νέων δεδομένων και ύστερα από πολλές προσπάθειες, το μοντέλο παρήγαγε αποτελέσματα όπως αυτό του σχήματος 5.2.



Σχήμα 5.2: Φασματογράφημα από γάβγισμα.

Το συγκεκριμένο φασματογράφημα φαίνεται να έχει τα χαρακτηριστικά ενός ήχου όπως το γάβγισμα σκύλου (dog_bark). Παρά το γεγονός αυτό, το ακουστικό αποτέλεσμα, ύστερα και από την ανακατασκευή της φάσης, δεν ήταν καθόλου καλό. Συγκεκριμένα ο ήχος ήταν θόρυβος με απλές αυξομειώσεις της έντασης, χωρίς να είναι δυνατό στο ανθρώπινο αυτί να ξεχωρίσει κάποια άλλα χαρακτηριστικά του ήχου. Έτσι έγινε εμφανές ότι δεν θα ήταν δυνατή η παραγωγή νέων ήχων για δύο βασικούς λόγους. Αρχικά τα δεδομένα μας ήταν λίγα στο πλήθος και ακόμη και αυτά τα δεδομένα που διαθέταμε ήταν αρκετά θορυβώδη αφού πολλοί ήχοι μοιάζουν αρκετά μεταξύ τους. Για να αποδείξουμε ότι τα δεδομένα μας δεν είναι διαχωρίσιμα προχωρήσαμε στην κατασκευή ενός πλήρως συνδεδεμένου ταξινομιτή, με σκοπό να δούμε πόσο καλά μπορεί το δίκτυο μας να ξεχωρίσει τα δεδομένα στην ενδιάμεση αναπαράσταση τους. Τα αποτελέσματα του ταξινομιτή έδωσαν ακρίβεια (precision) ίση με 35.53 % , ανάκληση (recall) ίση με 54.1 % και F1-score ίσο με 42.84 % , πάνω στο σύνολο δεδομένων ελέγχου. Είναι λοιπόν προφανές, χωρίς περαιτέρω έρευνα, πως το μοντέλο μας αδυνατεί να ξεχωρίσει τα δεδομένα και άρα είναι αδύνατο να διακρίνει τα χαρακτηριστικά τους και να παράγει νέα παρόμοια δεδομένα. Ενδεικτικά στο σχήμα 5.3 παρουσιάζουμε το φασματογράφημα ενός ήχου σειρήνας και την αντίστοιχη ανακατασκευή που πέτυχε το μοντέλο μας.

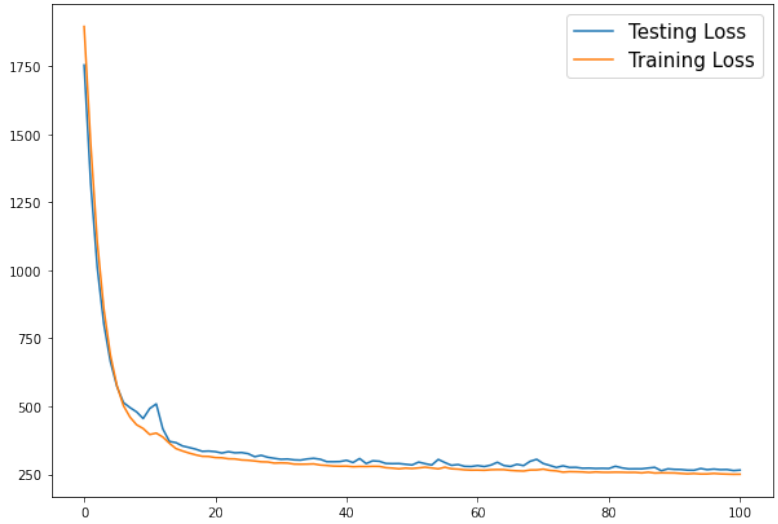


(α) (β)

Σχήμα 5.3: Φασματογραφήματα (α) Αρχικής σειρήνας.
(β) Σειρήνας όπως την ανακατασκεύασε το μοντέλο μας.

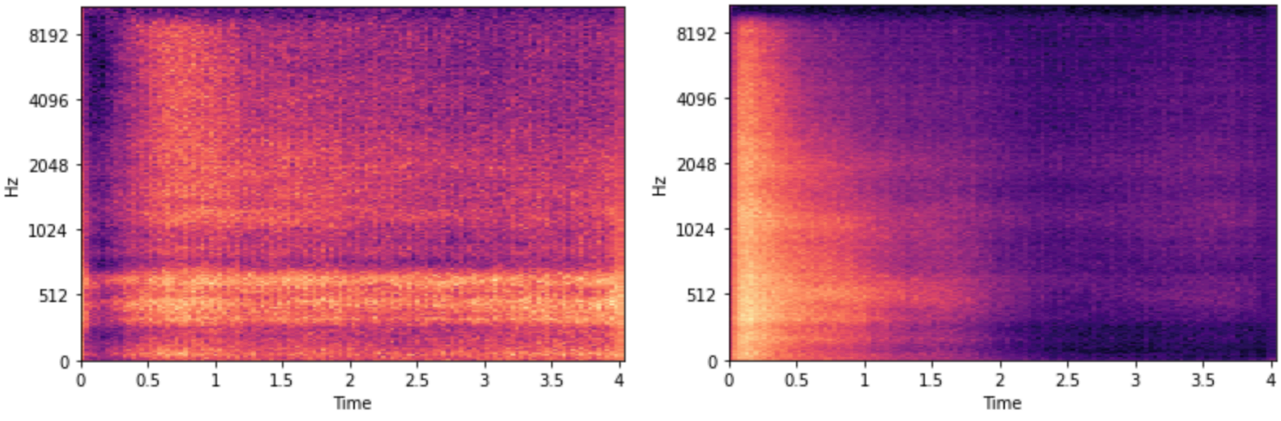
Είναι προφανές πως αυτό δεν είναι ένα ικανοποιητικό αποτέλεσμα. Έτσι αποφασίσαμε να διαχωρίσουμε κάποιους ήχους με διακριτά χαρακτηριστικά και να επαναλάβουμε την διαδικασία εκπαίδευσης μόνο με αυτούς τους ήχους. Οι κατηγορίες ήχων που επιλέξαμε είναι οι ήχοι σειρήνας (siren) και οι ήχοι πυροβολισμού (gun_shot). Αυτό είναι και ένα πείραμα που αποφασίσαμε να επαναλάβουμε για όλα τα μοντέλα που ακολουθούν, έτσι ώστε να διευκολυνθεί η αξιολόγηση και η σύγκριση τους.

Στο δεύτερο αυτό πείραμα εκπαιδεύσαμε το μοντέλο μας για 100 εποχές με τελική τιμή συνάρτησης κόστους ίση με 265 (Σχ. 5.4). Το μοντέλο αυτό ήταν σε θέση να παράγει



Σχήμα 5.4: Συνάρτηση κόστους για "CVAE + Spectrogram" για ήχους σειρήνας και πυροβολισμού.

αποτελέσματα που είχαν κατανοητό ακουστικό αποτέλεσμα. Η παραγωγή ήχων πυροβολισμού ήταν πιο εύκολη, καθώς είναι ένα ήχος με λιγότερη λεπτομέρεια στο φασματογράφημα του. Οι ήχοι σειρήνας, πάρα το γεγονός ότι καταφέραμε να τους δημιουργήσουμε, δεν ήταν οι βέλτιστοι ποιοτικά, αφού το φασματογράφημα τους, όπως φαίνεται και στο σχήμα 5.5α, απαιτεί σημαντική λεπτομέρεια για να έχουμε ποιοτικό και καθαρό ακουστικό αποτέλεσμα. Ακόμη κατά τη διαδικασία παραγωγής ήταν πολλές οι φορές που το αποτέλεσμα ήταν θόρυβος και δεν ανήκε σε καμία από τις δύο κατηγορίες ήχων, γεγονός που δείχνει ότι η κατανομή που δημιούργησε ο CVAE στην κρυφή του αναπαράσταση δεν είχε την συνέχεια που θέλαμε να επιτύχουμε. Αυτό σημαίνει

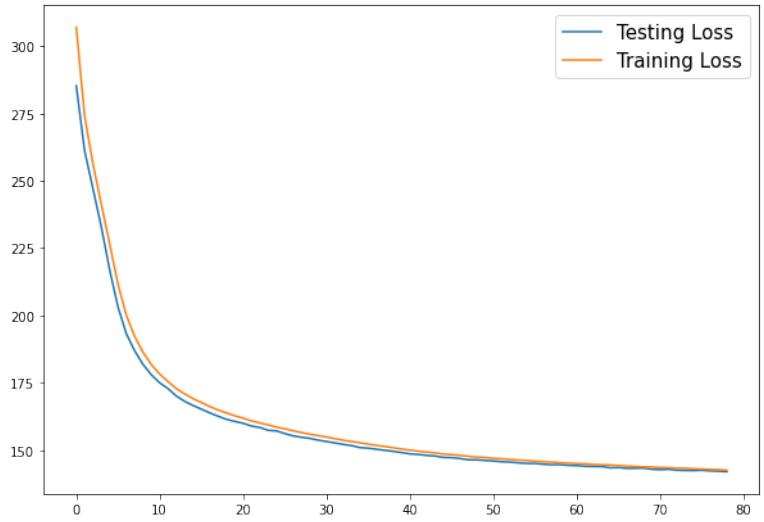


Σχήμα 5.5: Παραχθέντα φασματογραφήματα (α) Σειρήνας. (β) Πυροβολισμού.

πως υπήρχαν αρκετά σημεία του χώρου που τα δείγματα που λαμβάναμε ήταν τυχαία και δεν είχαν τα χαρακτηριστικά των ήχων που θα θέλαμε. Στο σχήμα 5.5 παρουσιάζονται δυο φασματογραφήματα, ένα για κάθε κατηγορία που καταφέραμε να παράγουμε.

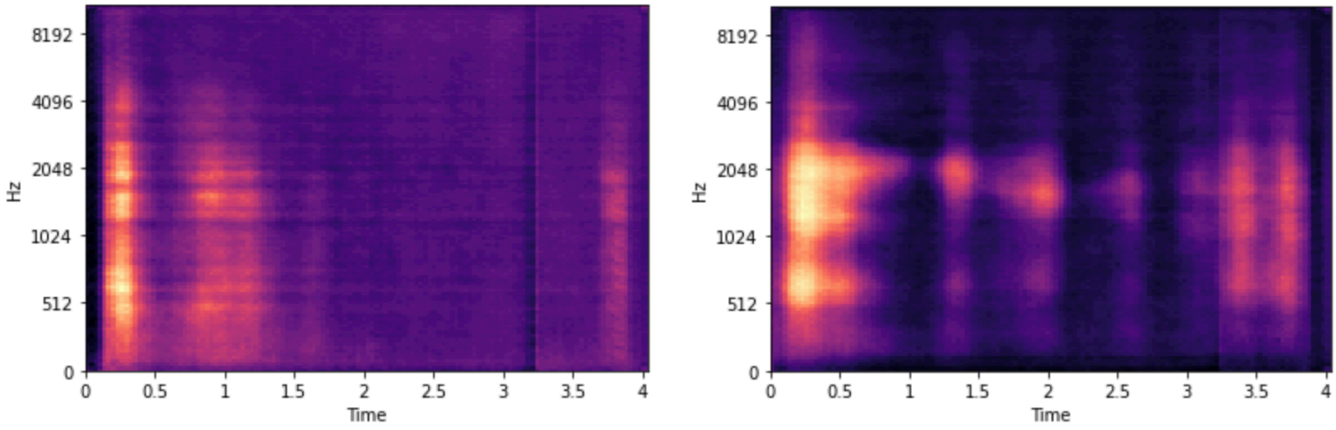
CVAE + Mel-Spectrogram

Ξεκινώντας με όλα τα δεδομένα στη διάθεση μας εκπαιδεύσαμε το μοντέλο για 80 εποχές με τελική τιμή συνάρτησης κόστους ίση με 142. Για περισσότερες εποχές το μοντέλο δεν φάνηκε να παρουσιάζει περαιτέρω βελτίωση (Σχ. 5.6)



Σχήμα 5.6: Συνάρτηση κόστους για "CVAE + Mel-Spectrogram" με όλα τα δεδομένα διαθέσιμα.

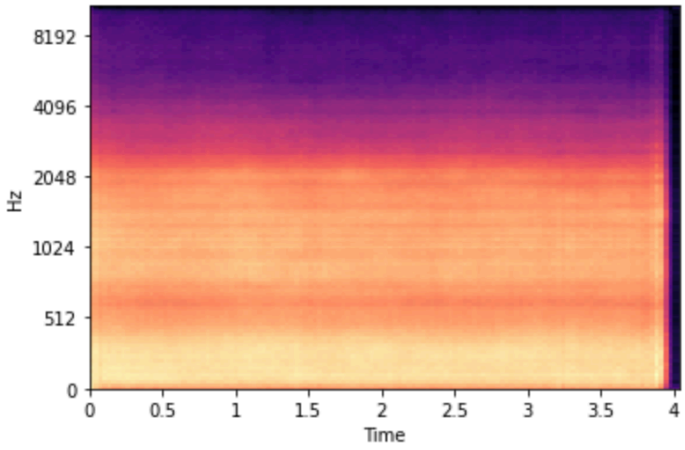
Σε αυτή την περίπτωση η διαδικασία της παραγωγής νέων ήχων φάνηκε να λειτουργεί μόνο για κατηγορίες ήχων με πολύ διαφορετικό και ξεχωριστό φασματογράφημα Mel. Συγκεκριμένα αυτές οι κατηγορίες είναι οι ήχοι γαβγίσματος σκύλου (dog_bark) και πυροβολισμών (gun_shot). Στη συνέχεια παρουσιάζονται δυο ήχοι που παράχθηκαν και το ακουστικό τους αποτέλεσμα αφορούσε γάβγισμα σκύλου (Σχ. 5.7).



(α) (β)

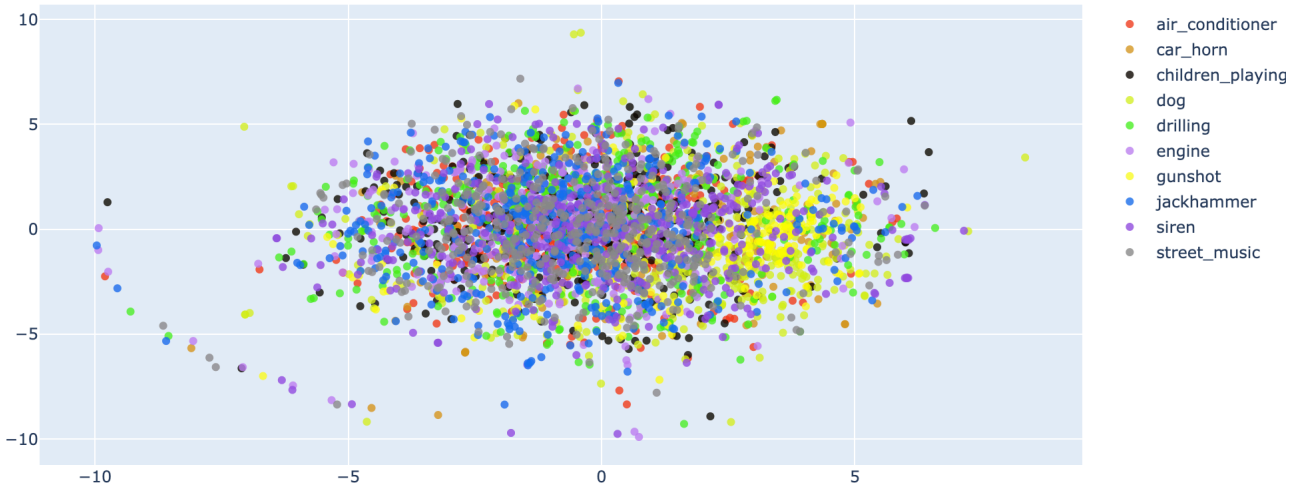
Σχήμα 5.7: Παραχθέντα φασματογραφήματα Mel.
(α) Γάβγισμα 1. (β) Γάβγισμα 2.

Επειδή οι υπόλοιποι ήχοι έχουν αρκετά παρόμοια φασματογραφήματα Mel το σύστημα μας ήταν πολύ δύσκολο να καταφέρει να τα διαχωρίσει στην ενδιάμεση αναπαράσταση βάση κάποιων διακριτών χαρακτηριστικών τους, όπως συνέβη στην περίπτωση των ήχων γαβγίσματος και πυροβολισμού που διαφοροποιούνται σημαντικά. Έτσι στη γενική περίπτωση οι ήχοι που παράγονταν είχαν την εικόνα του σχήματος 5.8. Το ακουστικό αποτέλεσμα είναι θόρυβος και σπανίως μπορεί να ακουστούν τα χαρακτηριστικά κάποιου ήχου, όπως μια σειρήνα (siren).



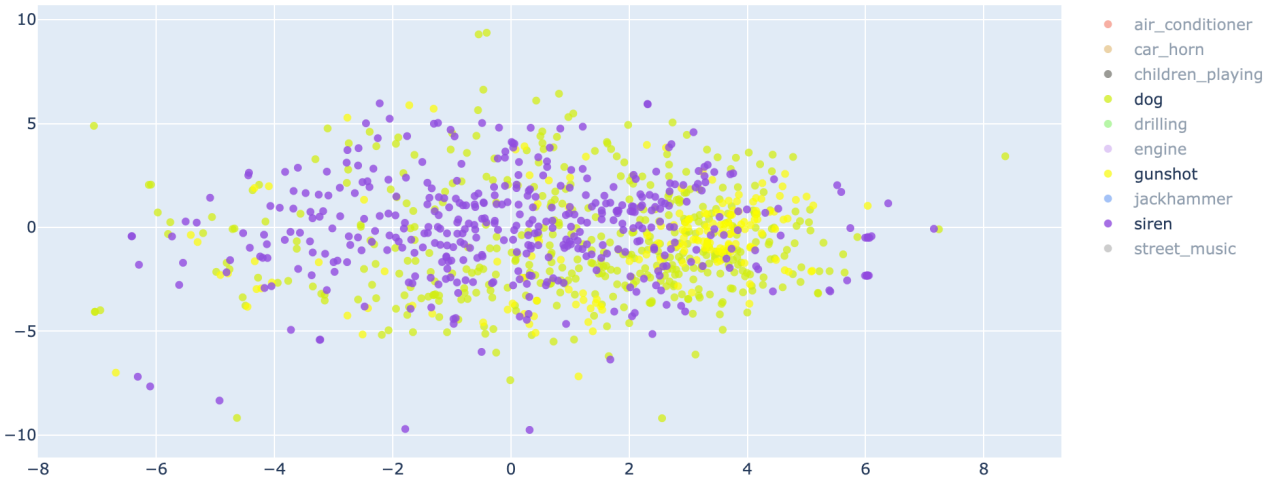
Σχήμα 5.8: Παραχθέν φασματογράφημα Mel ήχου σειρήνας.

Προκειμένου να γίνει ξεκάθαρο ότι στην ενδιάμεση αναπαράσταση τα δεδομένα δεν ήταν διαχωρίσιμα, τα μετατρέψαμε όλα στην αντίστοιχη ενδιάμεση αναπαράσταση τους, τα χωρίσαμε σε σύνολα εκπαίδευσης και ελέγχου και εκπαιδεύσαμε έναν ταξινομητή πάνω σε αυτά. Τα αποτελέσματα έδωσαν ακρίβεια (precision) ίση με 41.77 % , ανάκληση (recall) ίση με 65.01 % και F1-score ίσο με 51.69 % , πάνω στο σύνολο δεδομένων ελέγχου. Κάνοντας χρήση της τεχνικής t-SNE για την μη γραμμική μείωση της διαστατικότητας των δεδομένων πήραμε την παρακάτω εικόνα, η οποία κάνει πιο ξεκάθαρη την σύγχυση που υπάρχει μεταξύ των διαφορετικών κατηγοριών (Σχ. 5.9).



Σχήμα 5.9: Απεικόνιση ανάλυσης t-SNE δυο διαστάσεων για όλα τα δεδομένα.

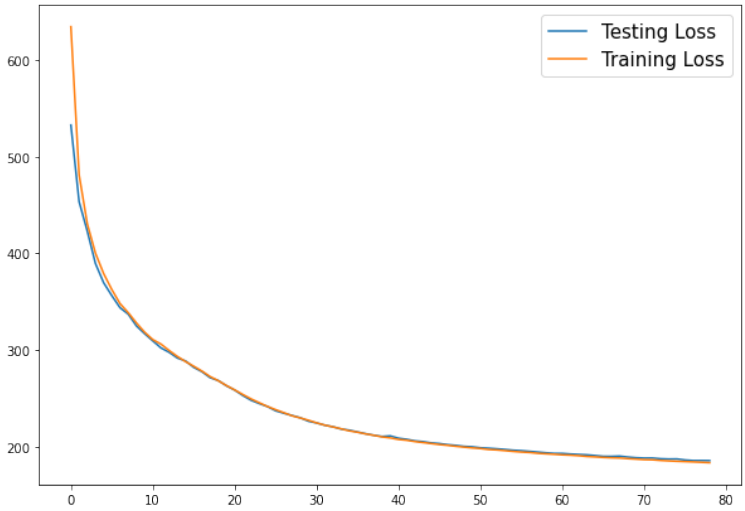
Με βάση αυτή την εικόνα προχωρήσαμε στο δεύτερο σκέλος αξιολόγησης των παραγωγικών ικανοτήτων του μοντέλου μας. Στο σκέλος αυτό κρατάμε μόνο εύκολα διαχωρισμούς ήχους ώστε να αξιολογήσουμε το μοντέλο αποκλειστικά για τις δυνατότητες παραγωγής νέων ήχων. Οι κατηγορίες που επιλέξαμε αφορούν τους ήχους σειρήνας (siren), πυροβολισμού (gun_shot) και γαβγίσματος (dog_bark). Στο σχήμα που ακολουθεί έχουμε ξεχωρίσει αυτούς τους ήχους από τους υπόλοιπους (Σχ. 5.10).



Σχήμα 5.10: Απεικόνιση ανάλυσης t-SNE δυο διαστάσεων για δεδομένα των κατηγοριών σειρήνα, γάβγισμα, πυροβολισμός.

Οι ήχοι, δεδομένου ότι χρησιμοποιήσαμε την τεχνική t-SNE, φαίνεται να είναι διαχωρισμοί με τους ήχους πυροβολισμού και γαβγίσματος να έχουν κάποια επικάλυψη.

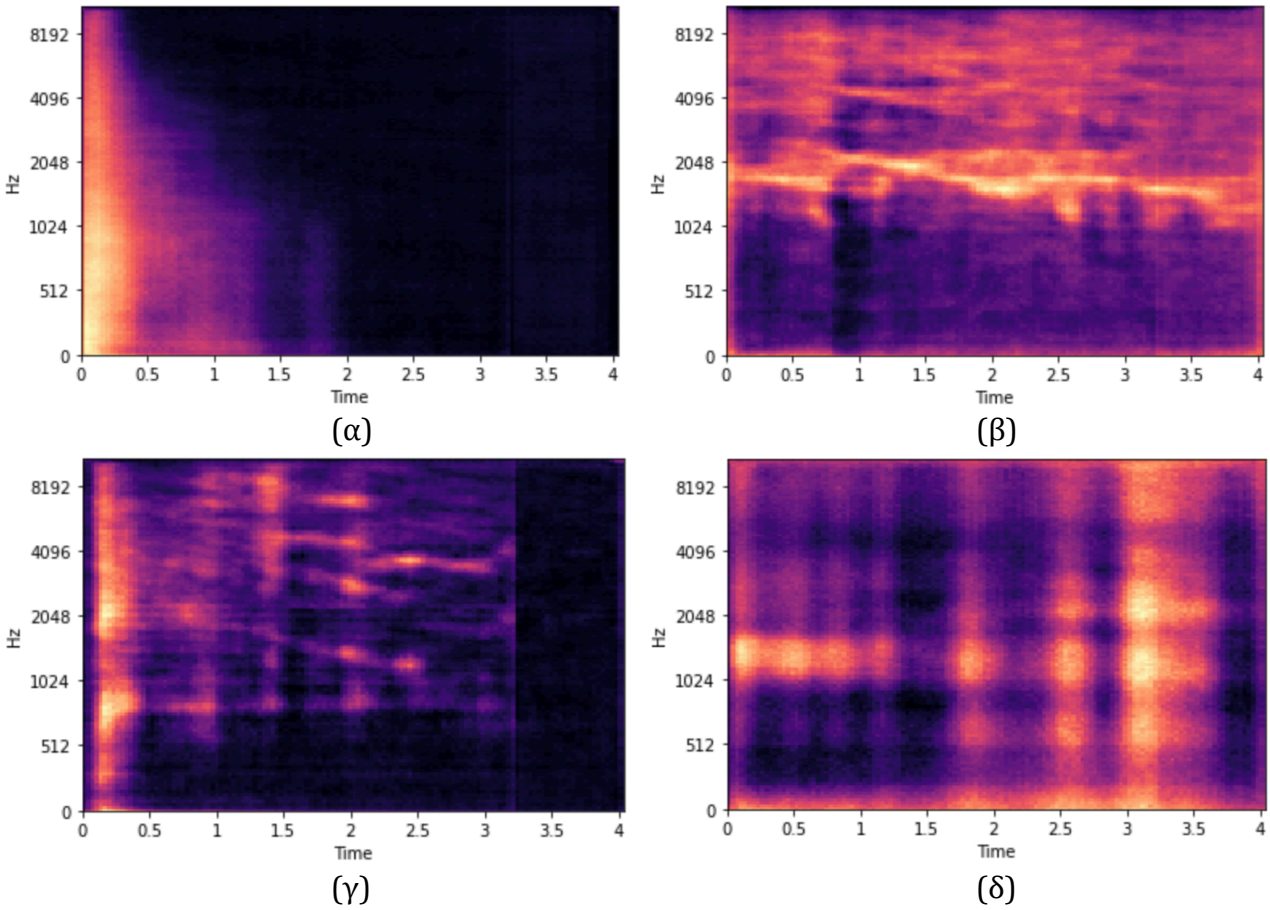
Το νέο μοντέλο εκπαιδεύτηκε για 80 εποχές με τελική τιμή συνάρτησης κόστους ίση με 185.



Σχήμα 5.11: Συνάρτηση κόστους για "Mel-Spectrogram" ήχους σειρήνας, γαβγίσματος και πυροβολισμού.

Καταφέραμε να παράγουμε ήχους σε κάθε μια από τις κατηγορίες. Το πιο ενδιαφέρον αποτέλεσμα αποτελεί η παραγωγή κάποιων ήχων που αποτελούν ανάμειξη 2 κατηγοριών. Συγκεκριμένα, καταφέραμε να παράγουμε ήχο που αποτελεί συνδυασμό του ήχου πυροβολισμού με τον ήχο

σειρήνας και ήχο που αποτελεί συνδυασμό του ήχου γαβγίσματος με τον ήχο σειρήνας. Στις εικόνες που ακολουθούν (Σχ. 5.12) παρουσιάζομαι τα φασματογραφήματα Mel για τους ήχους που καταφέραμε να παράγουμε.



Σχήμα 5.12: Παραχθέν φασματογραφήματα Mel (α) Πυροβολισμός. (β) Σειρήνα. (γ) Πυροβολισμός + Σειρήνα. (δ) Γάβγισμα + Σειρήνα.

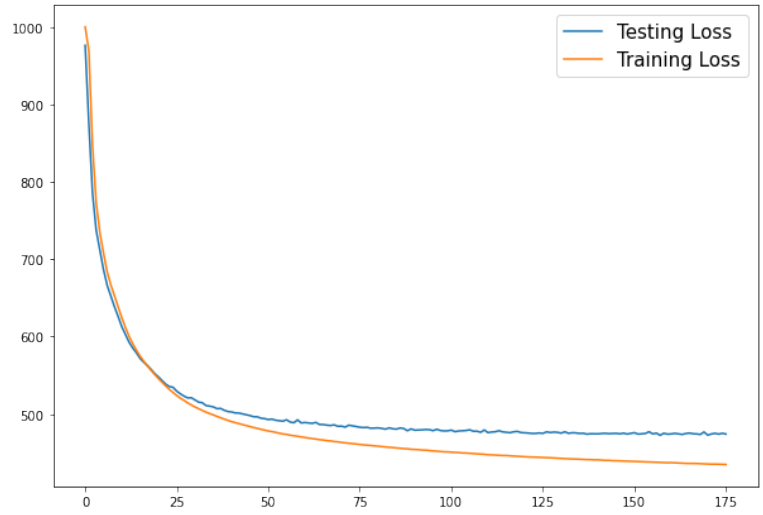
Το γεγονός ότι μπορούμε να παράγουμε ήχους, οι οποίοι αποτελούν συνδυασμό δυο κατηγοριών σημαίνει πως ο VAE έχει δουλέψει σωστά και έχει δημιουργήσει μια συνεχή κατανομή στην ενδιάμεση αναπαράσταση. Από όποιο σημείο αυτής της κατανομής και αν πάρουμε ένα δείγμα, θα λάβουμε ένα ορθό ακουστικό αποτέλεσμα, ακόμη και αν αυτό είναι μια μίξη διαφορετικών ήχων.

Η ποιότητα των ήχων δεν είναι η βέλτιστη. Παρόλα αυτά είναι εύκολα κατανοητοί και διακριτοί μεταξύ τους . Σχετικά με τη βελτίωση της ποιότητας, παρουσιάζουμε τους τρόπους με τους οποίους αυτή μπορεί να επιτευχθεί στο κεφάλαιο 6.

CVAE + MFCCs

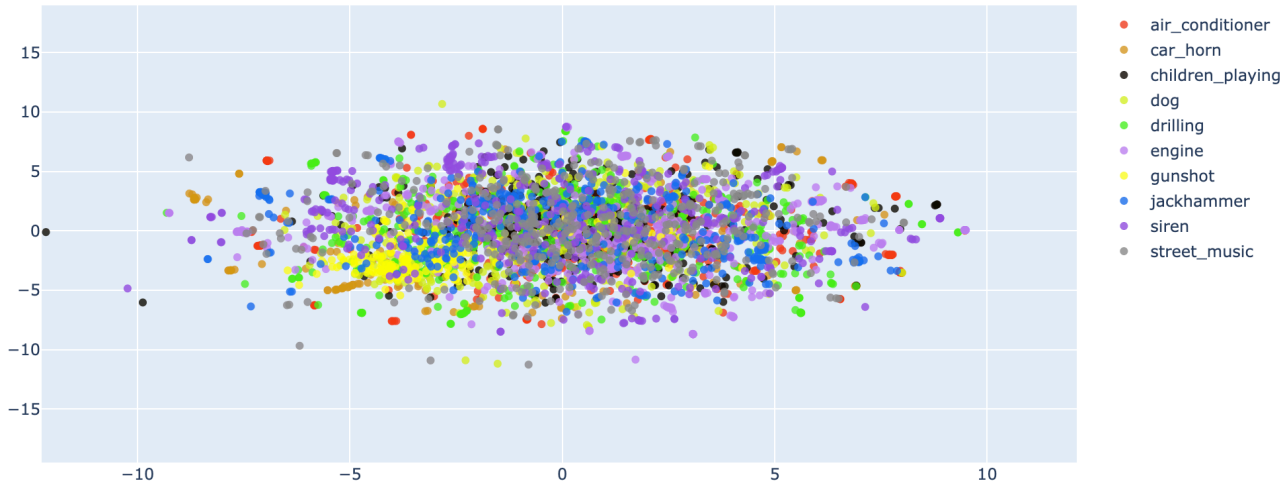
Αρχικά για την εκπαίδευση του συγκεκριμένου μοντέλου χρησιμοποιήθηκαν όλα τα δεδομένα. Στην περίπτωση αυτή και ύστερα από 175 εποχές η τιμή της συνάρτησης κόστους ήταν ίση με 473. Από εκείνο το σημείο και ύστερα φάνηκε να πραγματοποιείται υπερπροσαρμογή στα δεδομένα εκπαίδευσης.

Αυτή η τιμή της συνάρτησης κόστους σε κάποιες περιπτώσεις μπορεί να είναι ικανοποιητική ώστε ένα μοντέλο να παράγει αληθοφανείς ήχους. Κάτι τέτοιο δεν συνέβη με το συγκεκριμένο



Σχήμα 5.13: Συνάρτηση κόστους για "CVAE + MFCCs" με όλα τα δεδομένα.

μοντέλο, το οποίο παρήγαγε ήχους που έμοιαζαν απλά με θόρυβο. Η αδυναμία αυτή οφείλεται στο γεγονός ότι ο κωδικοποιητής του CVAE δεν ήταν σε θέση να διαχωρίσει τους ήχους στην ενδιάμεση αναπαράσταση. Τα πειράματα επαναλήφθηκαν αυξάνοντας το μέγεθος της ενδιάμεσης αναπαράστασης ($300 \cdot 2, 400 \cdot 2, 512 \cdot 2$), κάτι που δεν βελτίωσε τον διαχωρισμό των δεδομένων στην ενδιάμεση αναπαράσταση. Μάλιστα για να αποδείξουμε ότι ο διαχωρισμός των δεδομένων ήταν αδύνατος εκπαιδεύσαμε ένα απλό πλήρως συνδεδεμένο μοντέλο ταξινόμησης πάνω στις ενδιάμεσες αναπαραστάσεις των δεδομένων. Τα αποτελέσματα έδωσαν ακρίβεια (precision) ίση με 63.13 % , ανάκληση (recall) ίση με 77.68 % και F1-score ίσο με 68.21 % , πάνω στο σύνολο δεδομένων ελέγχου. Αυτό δείχνει την αδυναμία του μοντέλου να διαχωρίσει τα δεδομένα και κατ' επέκταση να είναι είναι σε θέση να παράγει νέα δεδομένα. Κάνοντας χρήση της τεχνικής t-SNE για την μη γραμμική μείωση της διαστατικότητας των δεδομένων πήραμε την παρακάτω εικόνα, η



Σχήμα 5.14: Απεικόνιση ανάλυσης t-SNE δυο διαστάσεων για όλα τα δεδομένα.

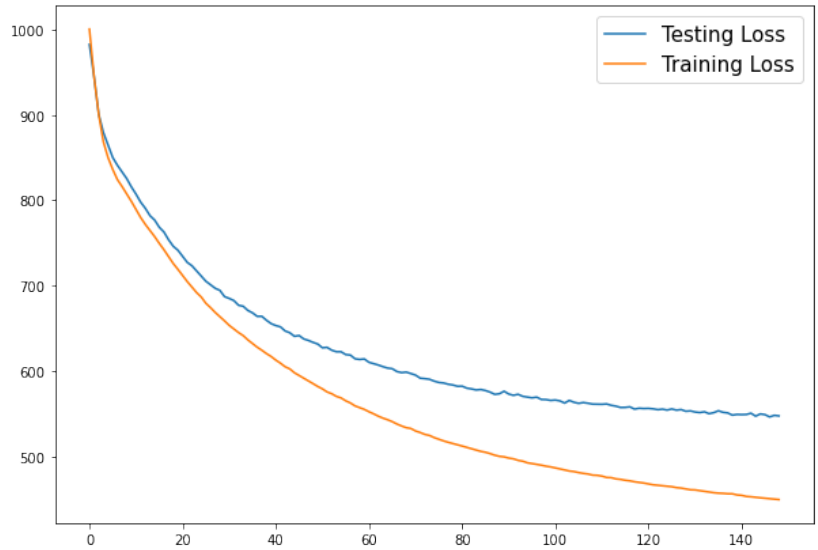
οποία κάνει πιο ξεκάθαρη την σύγχυση που υπάρχει μεταξύ των διαφορετικών κατηγοριών.

Στη συνέχεια αφού παρατηρήσαμε ότι το πλήρες σύνολο των δεδομένων δημιουργούσε σύγχυση στο μοντέλο μας, με αποτέλεσμα αυτό να μην μπορεί να παράγει ορθά αποτελέσματα που

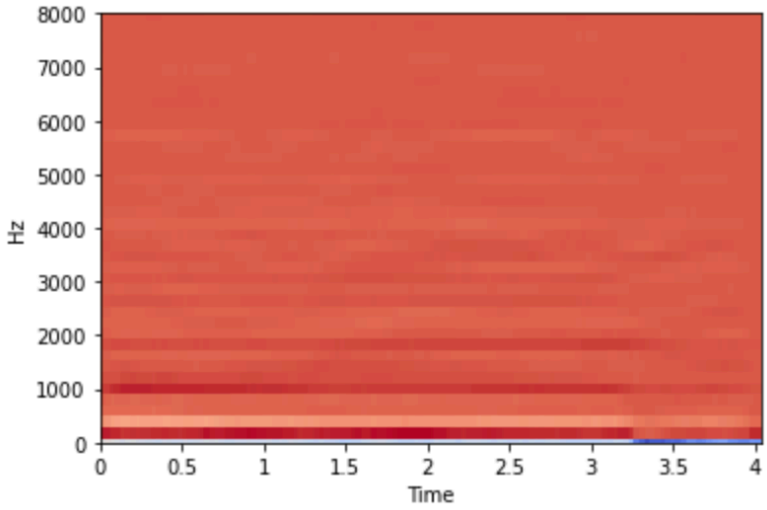
να προσομοιώνουν πραγματικούς ήχους, αποφασίσαμε να κρατήσουμε μόνο κάποια δεδομένα που το μοντέλο μας εύκολα θα μπορούσε να διαχωρίσει. Όπως φαίνεται και στο σχήμα 5.15 δύο κατηγορίες που ικανοποιούν αυτή την προϋπόθεση είναι οι ήχοι σειρήνας (siren) και πυροβολισμού (gunshot). Βάση λοιπόν του ότι στην ενδιάμεση αναπαράσταση ο διαχωρισμός των δύο αυτών κατηγοριών είναι ξεκάθαρος αναμένουμε να έχουμε καλύτερο ακουστικό αποτέλεσμα.

Σχήμα 5.15: Απεικόνιση ανάλυσης t-SNE δυο διαστάσεων για δεδομένα των κατηγοριών σειρήνα, πυροβολισμός.

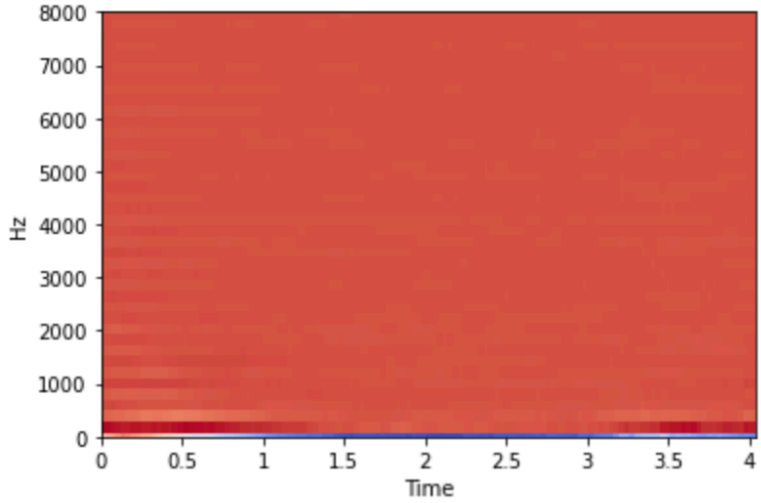
Η δεύτερη εκδοχή του μοντέλου αυτού εκπαιδεύτηκε για 150 εποχές έχοντας τελικά τιμή συνάρτησης κόστους ίση με 547. Ακόμη στις 150 εποχές φάνηκε να έχουμε υπερπροσαρμογή του παραγωγικού μας μοντέλου (Σχ. 5.16). Το γεγονός αυτό όμως στα πλαίσια ενός παραγωγικού μοντέλου δεν έχει την αρνητική σημασία που έχει στην επιβλεπόμενη μάθηση και υποδηλώνει άπλα ότι οι παραγόμενοι ήχοι μπορεί να μοιάζουν περισσότερο με τους ήχους του συνόλου εκπαίδευσης. Στην περίπτωση μας αυτό αυτό δεν οδηγεί σε καμία ουσιαστική διαφορά στους παραγόμενους ήχους, οι οποίοι αποτελούν απλά νέους ήχους από σειρήνες και πυροβολισμούς.



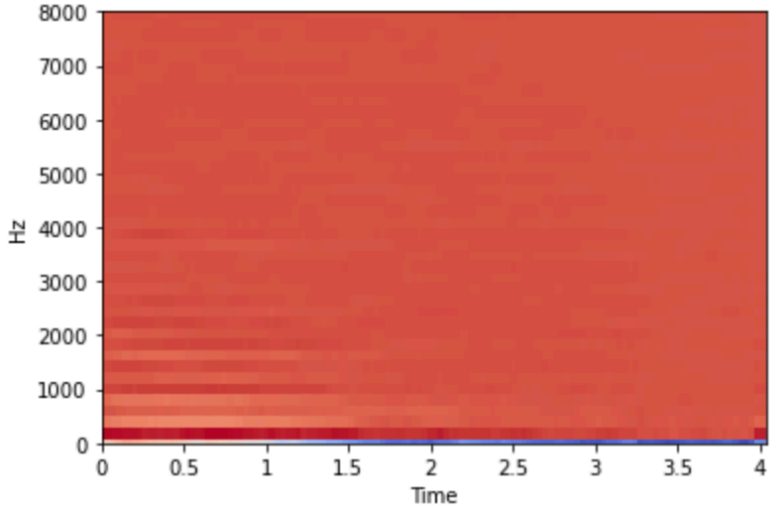
Σχήμα 5.16: Συνάρτηση κόστους για “CVAE + MFCCs” ήχους σειρήνας και πυροβολισμού.



(α)



(β)



(γ)

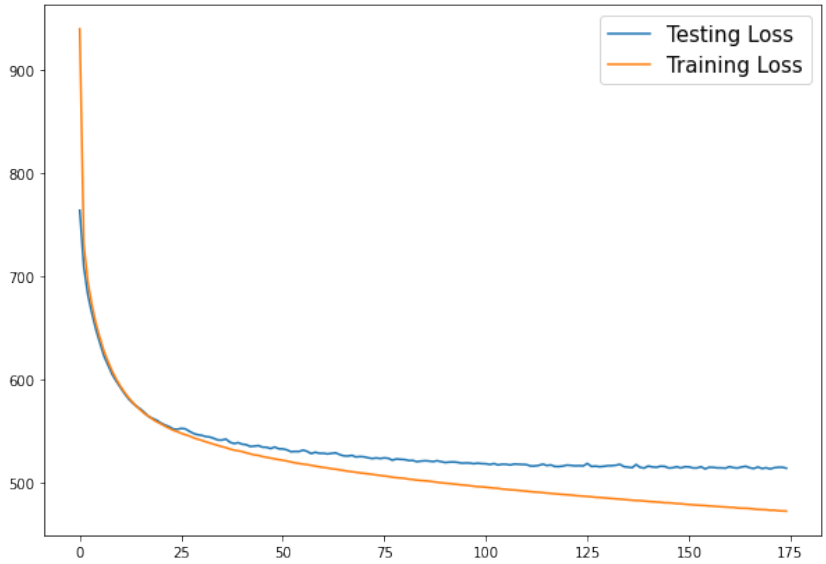
Σχήμα 5.17: Παραχθείσες MFCCs (α) Σειρήνας. (β) Πυροβολισμού. (γ) Σειρήνας + Πυροβολισμού.

Το μοντέλο κατάφερε να παράγει ήχους σε κάθε μια από τις δύο κατηγορίες ήχων με τις οποίες τροφοδοτήθηκε. Μάλιστα στο δείγμα των ήχων που πάρθηκε προκειμένου να αποφανθούμε ότι το μοντέλο λειτουργεί αποτελεσματικά, παρατηρήθηκε ότι αρκετές φορές παράγονταν ήχοι που ήταν κάτι ανάμεσα σε σειρήνα και πυροβολισμό. Πιο συγκεκριμένα οι ήχοι αυτοί είχαν την μικρή διάρκεια των ήχων πυροβολισμού αλλά παράλληλα και την ταλάντωση στη συχνότητα που είναι χαρακτηριστικό των ήχων σειρήνας. Αυτό είναι και ένα στοιχείο που μας δείχνει ότι έχει δημιουργηθεί επιτυχώς μια συνεχής κατανομή στην ενδιάμεση αναπαράσταση του CVAE, κάτι το οποίο είναι και ο τελικός σκοπός ενός VAE. Στη συνέχεια παρουσιάζονται τα φάσματα MFCC τριών ήχων που παράχθηκαν από το εκπαιδευμένο μοντέλο μας (Σχ. 5.17).

Πλήρως συνδεδεμένος (FC) VAE + MFCCs

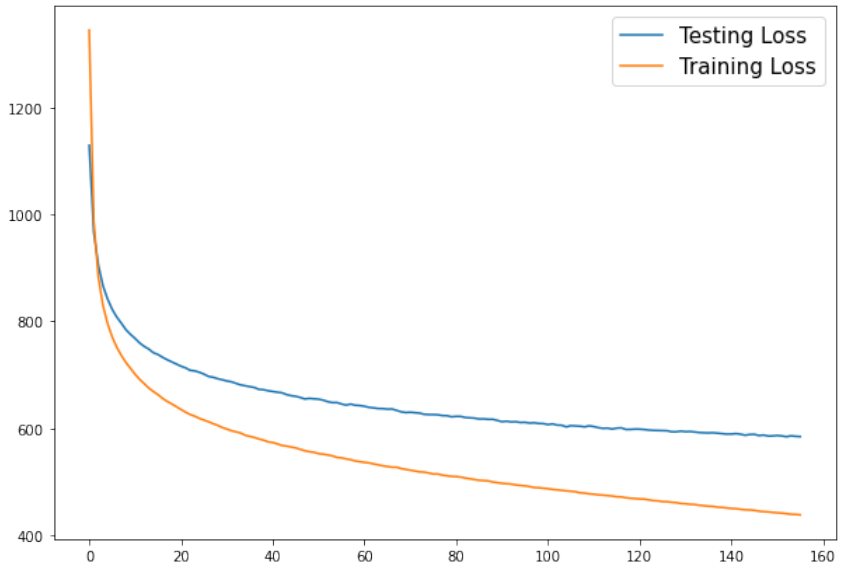
Το μοντέλο αυτό όπως και η αμέσως προηγούμενη περίπτωση δοκιμάστηκε σε δύο διαφορετικά σενάρια. Τα σενάρια αυτά είναι, αρχικά κάνοντας χρήση όλων των δεδομένων και στη συνέχεια χρησιμοποιώντας δυο διαχωρίσιμες κατηγορίες δεδομένων όπως οι ήχοι σειρήνας και πυροβολισμών. Ο σκοπός της δοκιμής των σεναρίων αυτών είναι προκειμένου να συγκριθούν άμεσα τα δύο μοντέλα που κάνουν χρήση της αναπαράστασης MFCC. Η σύγκριση των μοντέλων ακολουθεί στη ενότητα 5.2.

Αρχικά η απόδοση το μοντέλου εκτιμήθηκε στην περίπτωση που τα δεδομένα όλων των κατηγοριών ήταν διαθέσιμα. Στην περίπτωση αυτή το μοντέλο εκπαιδεύτηκε για 175 εποχές με τελική τιμή συνάρτησης κόστους ίση με 514 (Σχ. 5.18).

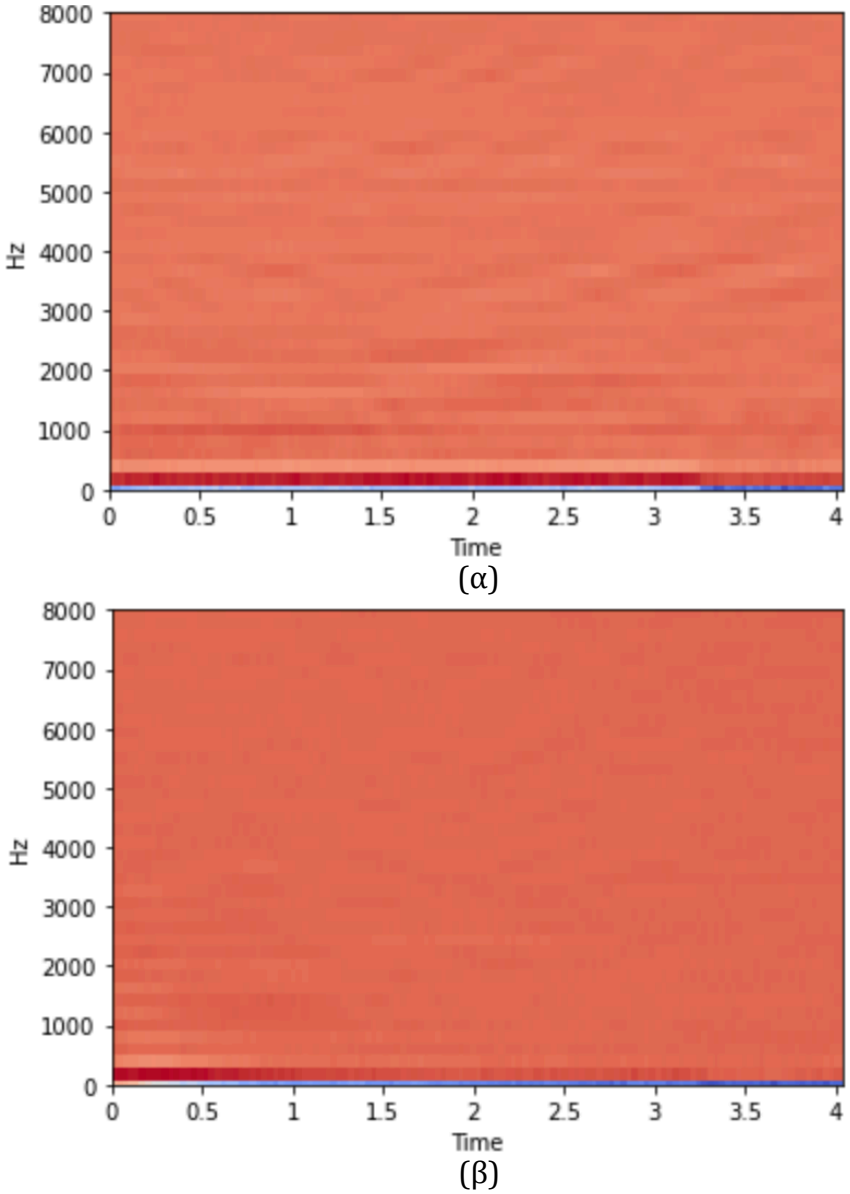


Σχήμα 5.18: Συνάρτηση κόστους για "FC-VAE + MFCCs" με όλα τα δεδομένα.

Επειδή και σε αυτή την περίπτωση τα δεδομένα δεν ήταν εφικτό να διαχωριστούν από το μοντέλο στην ενδιάμεση αναπαράσταση (όπως και στη περίπτωση CVAE + MFCCs χρησιμοποιήσαμε ταξινομητή), οι ήχοι που παρήγαγε το εκπαιδευμένο μοντέλο αναλογούσαν πιο πολύ σε θόρυβο με εξαίρεση κάποιους ήχους σειρήνας που μπορούν να ακουστούν σε μερικά από τα δείγματα που παράχθηκαν. Συγκεκριμένα τα αποτελέσματα του ταξινομητή σε αυτή την περίπτωση έδωσαν ακρίβεια (precision) ίση με 60.90 % , ανάκληση (recall) ίση με 74.48 % και F1-score ίσο με 66.96 % , πάνω στο σύνολο δεδομένων ελέγχου. Έτσι προχωρήσαμε στην διαδικασία



Σχήμα 5.19: Συνάρτηση κόστους για "FC-VAE + MFCCs" για ήχους σειρήνας, πυροβολισμού.



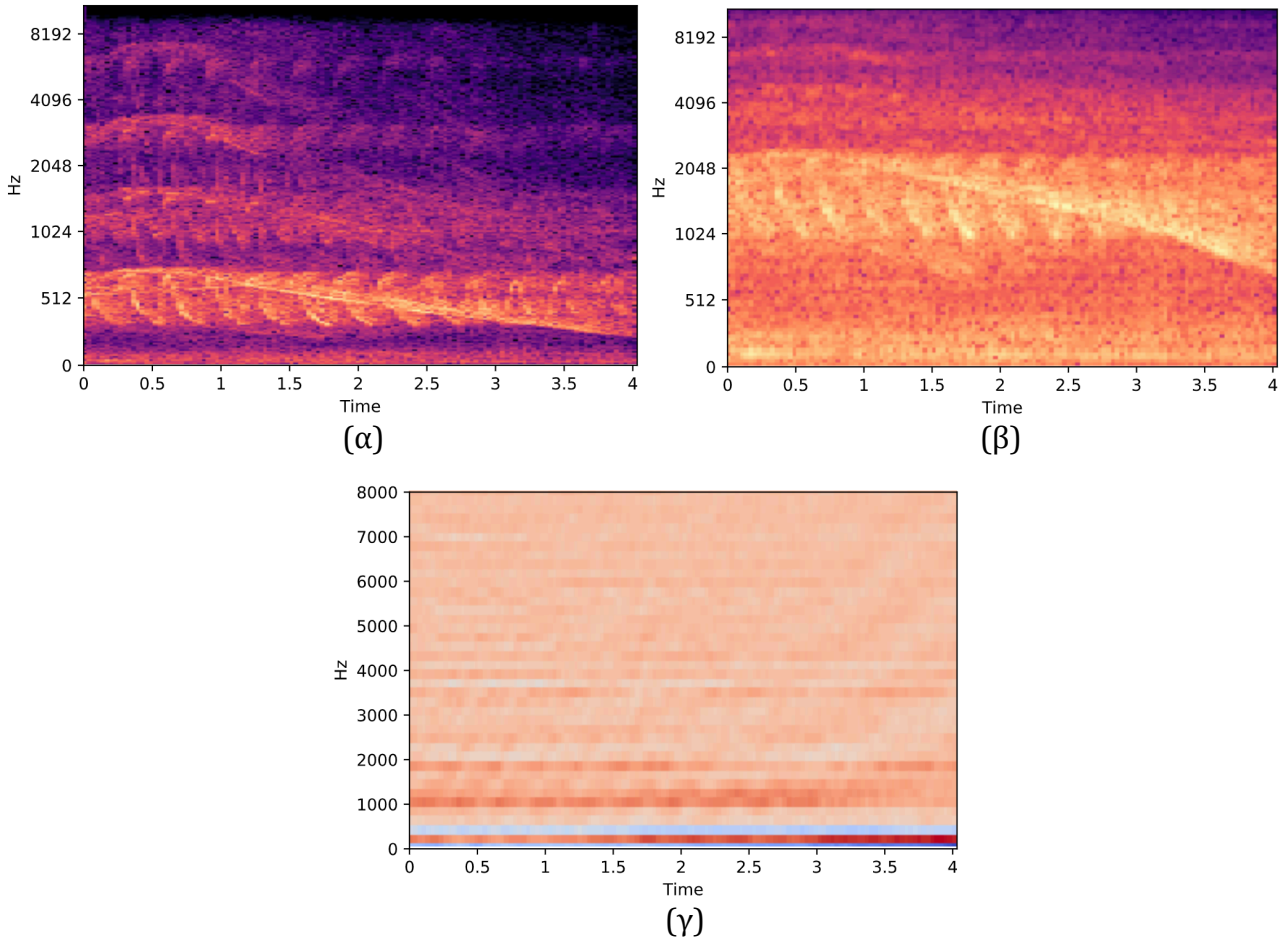
Σχήμα 5.20: Παραχθείσες MFCCs (α) Σειρήνας. (β) Πυροβολισμού.

υλοποίησης του δεύτερου σεναρίου, η οποία παρουσιάζεται στη συνέχεια.

Στο δεύτερο σενάριο εκπαίδευσης, απομονώσαμε και εκπαιδεύσαμε το μοντέλο πάνω στα δεδομένα που ανήκαν στις κατηγορίες των ήχων σειρήνας και πυροβολισμού. Στην περίπτωση αυτή το μοντέλο ύστερα από 150 εποχές είχε τιμή συνάρτησης κόστους ίση με 585. Η καμπύλη (Σχ. 5.19) φάνηκε να είναι ιδιαίτερα απότομη και το μοντέλο πολύ γρήγορα, ύστερα από ~10 εποχές άρχισε να υπερπροσαρμόζεται στα δεδομένα εκπαίδευσης. Το γεγονός αυτό όπως και στην περίπτωση του προηγούμενου μοντέλου δεν είναι κάτι που επηρεάζει τα τελικά αποτελέσματα του μοντέλου, τα οποία έχουν απλά μια παραπάνω ομοιότητα με τα δεδομένα του συνόλου εκπαίδευσης χωρίς αυτό να σημαίνει πως πρόκειται για απλή ανακατασκευή τους. Τα φάσματα MFCC κάποιων ενδεικτικών ήχων που παράχθηκαν παρουσιάζονται στο σχήμα 5.20.

5.2 Σύγκριση και σχολιασμός αποτελεσμάτων

Το μοντέλο “CVAE + Spectrogram” φάνηκε να είναι το πιο αδύναμο από όλα τα μοντέλα σε ότι αφορά τις παραγωγικές του δυνατότητες. Όπως αναφέρθηκε και σε προηγούμενα κεφάλαια, το γεγονός ότι δεν χρησιμοποιήσαμε όλα τα δεδομένα για την εκπαίδευση του σίγουρα έπαιξε κάποιο ρόλο στην ποιότητα των αποτελεσμάτων. Παρόλα αυτά η επίδραση αυτή δεν μπορεί να θεωρηθεί ο βασικός παράγοντας για την κακή ποιότητα των αποτελεσμάτων. Ο κύριος λόγος είναι η πολυπλοκότητα της εισόδου. Δηλαδή το απλό φασματογράφημα αποδίδει πολύ περισσότερη λεπτομέρεια σε κάθε σημείο του χώρου. Έτσι προκειμένου να παραχθεί κάποιο άρτιο ακουστικό



Σχήμα 5.21: Αναπαράσταση σειρήνας (α) Φασματογράφημα. (β) Φασματογράφημα Mel. (γ) MFCCs.

αποτέλεσμα, απαιτείται από το μοντέλο μας να παράγει μια πολύ πιο λεπτομερή εικόνα, σε σχέση με τα άλλα μοντέλα (Σχ. 5.21). Ακόμη και στην περίπτωση που αφαιρέσαμε κάποια δεδομένα και κρατήσαμε μόνο δεδομένα από εύκολα διαχωρίσιμες κατηγορίες, ενώ το μοντέλο ήταν σε θέση να παράγει κατανοητό ακουστικό αποτέλεσμα αυτό ήταν χειρότερο από τα άλλα μοντέλα που χρησιμοποιήσαμε.

Καταλήγουμε λοιπόν στο ότι η αρχιτεκτονική του μοντέλου "CVAE + Spectrogram" φαίνεται να τείνει στο bottleneck, γεγονός που οφείλεται στα δεδομένα τα οποία δέχεται και όχι στο βάθος του δικτύου ή τις τιμές άλλων υπερπαραμέτρων του.

Δεδομένων όσων ήδη αναφέρθηκαν, το μοντέλο "CVAE + Mel-Spectrogram" αναμένεται να βελτιώσει το πρόβλημα που αντιμετωπίστηκε με το μοντέλο "CVAE + Spectrogram" και τα δεδομένα εισόδου αυτου. Πράγματι, η είσοδος με φασματογράφημα Mel φάνηκε να δίνει καλύτερα αποτελέσματα σε όλες τις περιπτώσεις, ενώ η αρχιτεκτονική του μοντέλου παρέμεινε σχεδόν ίδια, με μικρές διαφορές (βλ. ενότητα 4.3). Το μοντέλο αυτό ήταν και το μόνο, το οποίο κατά την εκπαίδευση με όλα τα δεδομένα κατάφερε να παράγει κάποιο ακουστικό αποτέλεσμα για εύκολα διακριτές κατηγορίες όπως το γάβγισμα σκύλου. Βέβαια για τις υπόλοιπες κατηγορίες η έξοδος δεν ήταν το ίδιο καλή. Το γεγονός αυτό δεν οφείλεται όμως, στην αδυναμία του δικτύου μας ή στην αναπαράσταση της εισόδου όπως στην περίπτωση "CVAE + Spectrogram". Ο λόγος που το μοντέλο παρουσιάζει αδυναμία στην παραγωγή νέων ήχων όταν δέχεται σαν είσοδο τα επαυξημένα δεδομένα όλων των κατηγοριών είναι ο εξής και παρατίθεται ευθύς αμέσως. Αρχικά τα δεδομένα από την αρχή συμπεριλάμβαναν κατηγορίες που πολύ δύσκολα ήταν διαχωρίσιμες μεταξύ τους. Το πιο χαρακτηριστικό παράδειγμα αφορά τις κατηγορίες κλιματιστικό (air_conditioner), τρυπάνι (drilling), κινητήρας σε λειτουργία (engine_idling) και κομπρεσέρ (jackhammer). Οι συγκεκριμένες κατηγορίες περιλαμβάνουν όλες ήχους που μοιάζουν αρκετά με θόρυβο και είναι παρόμοιοι μεταξύ τους. Σε αντίθεση με το προηγούμενο μοντέλο, εδώ επιλέξαμε να πραγματοποιήσουμε μια πιο εκτενή ανάλυση και να καταλήξουμε στο να κρατήσουμε μόνο κατηγορίες ήχων που δεν είχαν ιδιαίτερη επικάλυψη στα χαρακτηριστικά τους (siren, dog_bark, gun_shot). Πράγματι το μοντέλο μας κατάφερε να εκπαιδευτεί και να παράγει νέους ήχους για κάθε μια από αυτές τις κατηγορίες. Μάλιστα το πιο εντυπωσιακό αποτέλεσμα ήταν ότι καταφέραμε να παράγουμε και ήχους, οι οποίοι αποτελούσαν μίξη περισσότερων από μιας κατηγορίας από αυτές. Αυτό αποδεικνύει πως, η κατανομή που θέλουμε να πετύχουμε στον VAE και αναλύσαμε θεωρητικά στην ενότητα 2.3.2, επιτεύχθηκε με ιδανικό τρόπο με αυτόν τον τρόπο και αυτό το υποσύνολο δεδομένων. Η ποιότητα των δεδομένων σίγουρα επιδέχεται βελτίωσης, αλλά το συμπέρασμα είναι πως το συγκεκριμένο μοντέλο αποτελεί μια πολλά υποσχόμενη αφετηρία στην παραγωγή νέων ήχων κάνοντας χρήση αρχιτεκτονικών CVAE.

Προχωράμε στην σύγκριση των μοντέλων που κάνουν χρήση της αναπαράστασης των δεδομένων ως MFCC, δηλαδή "CVAE + MFCCs" και "Πλήρως συνδεδεμένος VAE + MFCCs". Στην περίπτωση αυτή, όπως και στις προηγούμενες, συγκρίνουμε τις επιδόσεις των μοντέλων πάνω σε δύο σενάρια όπως αυτά ορίστηκαν στις αντίστοιχες ενότητες παρουσίασης των αποτελεσμάτων των μοντέλων. Στην περίπτωση που κάναμε χρήση όλων των δεδομένων, το μοντέλο του CVAE φάνηκε να έχει ελαφρώς καλύτερη επίδοση από εκείνη του πλήρως συνδεδεμένου VAE. Παρόλα αυτά η επίδοση αυτή δεν ήταν τόσο καλύτερη ώστε να είναι σε θέση να παράγει αληθοφανείς ήχους στην έξοδο. Καταλήξαμε πως κανένα από τα δύο μοντέλα ανεξάρτητα από τις τιμές των υπερπαραμέτρων δεν μπορεί να βελτιωθεί τόσο ώστε να παράγει νέους ήχους, όταν στην είσοδο δεχόταν δεδομένα και από τις 10 κατηγορίες. Το γεγονός αυτό δεν οφείλεται όμως, στην αδυναμία των δικτύων μας σε ότι αφορά την αρχιτεκτονική τους. Το μεγαλύτερο ρόλο σε αυτό τον παίζει η ανεπάρκεια της αναπαράστασης MFCC, να αποδώσει αποδοτικά όλη την πληροφορία του κάθε ήχου και ιδιαίτερα τα στοιχεία που διακρίνουν τις κατηγορίες που συγχέονται πιο εύκολα μεταξύ τους. Κάτι τέτοιο βέβαια μπορεί να μην είναι και δυνατόν με καμία μορφή αναπαράστασης, βάση των αποτελεσμάτων που έχουμε ήδη αναλύσει. Στο συμπέρασμα αυτό καταλήξαμε ύστερα από

αρκετές διαφοροποιήσεις των δικτύων, οι οποίες υπέδειξαν πως τελικά το bottleneck του δικτύου δεν βρίσκεται στην αρχιτεκτονική αλλά στα ίδια τα δεδομένα και στον τρόπο αναπαράστασης τους. Έτσι αναλύοντας για άλλη μια φορά τις κατηγορίες δεδομένων που είναι εύκολα διαχωρίσιμες στην ενδιάμεση αναπαράσταση από το δίκτυο μας κρατήσαμε μόνο κάποια καλά διαχωρίσιμα δεδομένα όπως οι σειρήνες (siren) και οι πυροβολισμοί (gun_shot). Με τον τρόπο αυτό καταφέραμε να παράγουμε νέους ήχους και να δείξουμε πως εφόσον τα δεδομένα έχουν αρκετά διακριτά χαρακτηριστικά μεταξύ τους, η αναπαράσταση MFCC και οι αρχιτεκτονικές μας είναι επαρκείς. Σε ότι έχει να κάνει με την σύγκριση μεταξύ των μοντέλων που κάνουν χρήση της αναπαράστασης MFCC, το μοντέλο CVAE παρήγαγε καλύτερα ακουστικά αποτελέσματα από το μοντέλο πλήρως συνδεδεμένου VAE. Μάλιστα στην περίπτωση του CVAE καταφέραμε να παράγουμε έναν υβριδικό ήχο με χαρακτηριστικά και των δύο κατηγοριών, κάτι που δεν καταφέραμε να πάρουμε στον πλήρως συνδεδεμένο VAE όσες δοκιμές και αν κάναμε. Το αποτέλεσμα αυτό είναι κάτι που μας δείχνει πως η οργάνωση των δεδομένων στην ενδιάμεση αναπαράσταση και η κατανομή από την οποία παίρνουμε τα νέα δείγματα έχει καλύτερη συνέχεια στην περίπτωση CVAE.

6

6. Συμπεράσματα και μελλοντική δουλειά

6.1 Συμπεράσματα

Ο τρόπος παραγωγής και η φύση των αποτελεσμάτων της πειραματικής διαδικασίας δεν επιτρέπει την ποσοτική και μαθηματική κατηγοριοποίηση των αποτελεσμάτων μας (βλ. ενότητα 3.4). Παρόλα αυτά η ανθρώπινη αντίληψη μας οδηγεί στο να επιλέξουμε ως το αποδοτικότερο μοντέλο το μοντέλο “CVAE + Mel-Spectrogram”. Ο λόγος για αυτό είναι πως πέρα από την ποιότητα των ήχων, το μοντέλο αυτό κατάφερε να δίνει με μεγαλύτερη συχνότητα ήχους που περιείχαν μίξη περισσότερων από μιας κατηγορίας. Αυτό δείχνει, όπως έχει αναφερθεί σε προηγούμενο κεφάλαιο, ότι η εργασία που προσπαθεί να πετύχει ένας VAE έχει ολοκληρωθεί με επιτυχία. Αυτό αποδεικνύει την καταλληλότητα του μοντέλου μας για την χρήση του ως ένας VAE, που είναι σε θέση να παράγει νέους ήχους μέσω της χρήσης συνελικτικού κωδικοποιητή και αποκωδικοποιητή (CVAE), ο οποίος ήταν και ο άμεσος σκοπός αυτής της εργασίας.

Ένα άλλο πολύ βασικό εύρημα της εργασίας είναι η καταλληλότητα των αντίστοιχων αναπαραστάσεων για τον σκοπό που θέλουμε να πετύχουμε. Από την μια το απλό φασματογράφημα περιέχει πάρα πολύ λεπτομέρεια, πράγμα που το καθιστά ακατάλληλο για να είναι το προϊόν παραγωγής ενός παραγωγικού δικτύου. Από την άλλη η αναπαράσταση με MFCCs, αν και έδωσε ικανοποιητικά αποτελέσματα, λόγω της απλότητας της και της συμπυκνωμένης πληροφορίας που διαθέτει, φαίνεται πως για πιο σύνθετους ήχους οι δυνατότητες που προσφέρει είναι περιορισμένες και έτσι θα καθορίζει αυτή το bottleneck του δικτύου και όχι το ίδιο το δίκτυο. Η αναπαράσταση με φασματογράφημα Mel φαίνεται να κρατάει μια ισορροπία μεταξύ των άλλων δυο αναπαραστάσεων. Διαθέτει μεν αρκετή λεπτομέρεια, περισσότερη από την αναπαράσταση MFCC, έτσι ώστε να είναι σε θέση να ανακατασκευάσει και πιο σύνθετους ήχους , αλλά όχι τόση ώστε να την καθιστά περίπλοκη όπως το απλό φασματογράφημα. Φαίνεται πως δεν φτάσαμε στα όρια της συγκεκριμένης αναπαράστασης και πως ενδεχομένως υπάρχουν ακόμη καλύτερα και περισσότερα αποτελέσματα να μας δώσει.

Η συγκεκριμένη εργασία είχε ως απώτερο σκοπό να αποφανθεί για το αν υπάρχει πεδίο για έρευνα σε ότι αφορά την παραγωγή νέων ήχων, με υψηλό βαθμό τυχαιότητας, όπως αυτοί ενός αστικού περιβάλλοντος κάνοντας χρήση VAE και συγκεκριμένα αρχιτεκτονικές CVAE. Το συμπέρασμα είναι πως πράγματι υπάρχει δυνατότητα να επιτευχθεί αυτό, αλλά τα αποτελέσματα δείχνουν πως δεν μπορεί να γίνει χωρίς αυτό να έχει αντίκτυπο στην ποιότητα των ήχων που θα

παραχθούν. Αυτό βέβαια είναι κάπως αναμενόμενο αφού είναι αδύνατο να αναπαραστήσουμε τον ήχο σε μια οπτική αναπαράσταση χωρίς να χαθεί καθόλου πληροφορία. Παρόλα αυτά, ακόμη και αυτό το εμπόδιο υπάρχει πιθανότητα να ξεπεραστεί με τη χρήση διαφορετικών καναλιών που θα επεξεργάζονται διαφορετικές αναπαραστάσεις του ήχου και θα διατηρούν ξεχωριστή και σημαντική πληροφορία. Αυτό βέβαια είναι ένα αρκετά πιο δύσκολο έργο. Κάποιο ες ακόμη βελτιώσεις που θα μπορούσαν να γίνουν στο μέλλον πάνω στην παρούσα εργασία παρουσιάζονται στην επόμενη ενότητα.

6.2 Μελλοντική δουλειά

Η εργασία περιέχει πολλά σημεία στα οποία θα μπορούσαν να γίνουν διαφορετικές επιλογές. Στις περισσότερες περιπτώσεις αυτές οι εναλλακτικές δοκιμάστηκαν και δεν έδωσαν καλύτερα αποτελέσματα από αυτά που παρουσιάστηκαν. Υπάρχουν, παρόλα αυτά, κάποιες επιλογές που θα μπορούσαν να δοκιμαστούν σε μελλοντικές εκδοχές της εργασίας και να βελτιώσουν πιθανώς τα αποτελέσματα αυτής.

Αρχικά το περιεχόμενο της εργασίας θα μπορούσε να δοκιμαστεί πάνω σε κάποιο άλλο σύνολο δεδομένων. Όπως αναφέρθηκε σε πολλά σημεία της εργασίας, το σύνολο δεδομένων που χρησιμοποιήσαμε αποδείχτηκε πως δεν ήταν ιδανικό. Αυτό είναι ίσως και αναμενόμενο αφού δημιουργήθηκε έχοντας κατά νου την ταξινόμηση ήχων και όχι την δημιουργία ενός μοντέλου που θα παρήγαγε νέους ήχους. Αποτελεί βέβαια ένα καλά οργανωμένο σύνολο και προσφέρει την επιλογή να απομονώσουμε κατηγορίες ήχων και να εργαστούμε πάνω σε αυτές, όπως τελικά κάναμε και πήραμε ακουστικά αποτελέσματα.

Ένα άλλο κομμάτι πιθανής βελτίωσης αποτελεί η κανονικοποίηση που εφαρμόστηκε στα δεδομένα. Πιο συγκεκριμένα, η μεγαλύτερη βελτίωση μπορεί να υπάρξει στην αναπαράσταση MFCC, όπου εφαρμόσαμε μια καθολική κανονικοποίηση μηδενικού μέσου και μοναδιαίας απόκλισης πάνω σε όλες τις κατηγορίες. Σε αυτή τη περίπτωση μπορεί αν δοκιμαστεί το σενάριο να κανονικοποιηθεί κάθε κατηγορία ξεχωριστά ή να δοκιμαστεί και κάποια άλλη μορφή κανονικοποίησης.

Μια ακόμη πιθανή βελτίωση αποτελεί η μέθοδος με την οποία πραγματοποιείτε η ανακατασκευή της φάσης. Πέρα από την μέθοδο Griffin-Lim, που χρησιμοποιήσαμε στην παρούσα εργασία και αποτελεί την πιο διαδεδομένη μέθοδο ανακατασκευής της φάσης, έχουν αναπτυχθεί και άλλες πιο σύγχρονοι μέθοδοι που ενδεχομένως να είναι πιο αποδοτικές. Στην εργασία δεν αφιερώθηκε χρόνος στο συγκεκριμένο κομμάτι, μιας και η ανακατασκευή της φάσης άφορα την τελειοποίηση του τελικού ακουστικού αποτελέσματος και όχι την διαδικασία παραγωγής νέων δεδομένων από το νευρωνικό μας δίκτυο. Παρόλα αυτά, αφορά ένα κομμάτι που σίγουρα π

Ένα σενάριο το οποίο δοκιμάστηκε και δεν φάνηκε να αποδίδει, είναι να δίνουμε στην ενδιάμεση αναπαράσταση 10 επιπλέον θέσεις που δηλώνουν την κατηγορία κάθε δεδομένου εκπαίδευσης, με σκοπό να δίνουμε πρόσθετη πληροφορία που μπορεί να διευκολύνει ή και να εμπλουτίσει την παραγωγική διαδικασία. Αυτό ήταν ένα σενάριο που δεν φάνηκε να δουλεύει και να προσφέρει κάτι στα μοντέλα μας. Κάτι που θα μπορούσε να προσφέρει όμως είναι η δημιουργία και η εκπαίδευση διαφορετικού αποκωδικοποιητή για κάθε κατηγορία ήχου, δηλαδή να έχουμε έναν κοινό κωδικοποιητή και 10 αποκωδικοποιητές (για το συγκεκριμένο σύνολο δεδομένων που αποτελείται από 10 κατηγορίες). Η δοκιμή αυτού του σεναρίου ήταν αδύνατη με τους πόρους που είχαμε στη διάθεση μας, αλλά φαίνεται πολλά υποσχόμενη αν βασιστούμε στα πειράματα που κάναμε χρησιμοποιώντας μόνο δύο κατηγορίες ήχων.

Τέλος, μια σύγχρονη τεχνική που παρουσιάζει πολλά και καλά αποτελέσματα είναι η χρήση vector quantization σε συνδυασμό με VAE. Μια πρόσφατη εργασία πάνω στη παραγωγή μουσικής

της εταιρίας OpenAI, με όνομα Jukebox (Dhariwal et al, 2020), κάνει χρήση ακριβώς αυτής της τεχνικής και επιδεικνύει εντυπωσιακά και πρωτοφανή αποτελέσματα. Η εφαρμογή αυτής της τεχνικής στην παρούσα εργασία είναι ένα πολύ μεγάλο έργο, που όμως αφήνει μεγάλα περιθώρια βελτίωσης και προσδοκιών για το μέλλον της παραγωγής ήχων αστικού περιβάλλοντος και γενικότερα φυσικών ήχων, κάνοντας χρήση βαθιάς μηχανικής μάθησης.

Παράρτημα Α: Εξαγωγή ELBO

Για την εξαγωγή του ELBO ξεκινάμε την ανάλυση μας από τον τύπο της απόκλισης KL:

$$D_{KL}\left(q_\theta(z|x_i)||p(z|x_i)\right) = -\int q_\theta(z|x_i)\log\left(\frac{p(z|x_i)}{q_\theta(z|x_i)}\right)dz \geq 0$$

Βάση του θεωρήματος Bayes έχουμε:

$$D_{KL}\left(q_\theta(z|x_i)||p(z|x_i)\right) = -\int q_\theta(z|x_i)\log\left(\frac{p_\phi(x_i|z)p(z)}{q_\theta(z|x_i)p(x_i)}\right)dz \geq 0$$

Κάνοντας χρήση του κανόνα των λογαρίθμων έχουμε:

$$D_{KL}\left(q_\theta(z|x_i)||p(z|x_i)\right) = -\int q_\theta(z|x_i)\left[\log\left(\frac{p_\phi(x_i|z)p(z)}{q_\theta(z|x_i)}\right) - \log p(x_i)\right]dz \geq 0$$

Επιμερίζοντας το ολοκλήρωμα η εξίσωση γίνεται:

$$-\int q_\theta(z|x_i)\log\left(\frac{p_\phi(x_i|z)p(z)}{q_\theta(z|x_i)}\right)dz + \int q_\theta(z|x_i)\log p(x_i)dz \geq 0$$

Στην παραπάνω εξίσωση παρατηρούμε ότι ο λογάριθμος του δεύτερου ολοκληρώματος είναι μια σταθερά και άρα μπορεί αν βγει έξω από αυτό:

$$-\int q_\theta(z|x_i)\log\left(\frac{p_\phi(x_i|z)p(z)}{q_\theta(z|x_i)}\right)dz + \log p(x_i)\int q_\theta(z|x_i)dz \geq 0$$

Και εφόσον μέσα στο ολοκλήρωμα μένει μόνο μια κατανομή πιθανότητας αυτό το ολοκλήρωμα έχει αποτέλεσμα ίσο με τη μονάδα, άρα:

$$-\int q_\theta(z|x_i)\log\left(\frac{p_\phi(x_i|z)p(z)}{q_\theta(z|x_i)}\right)dz + \log p(x_i) \geq 0.$$

Μεταφέροντας το ολοκλήρωμα στο άλλο μέλος και κάνοντας χρήση του κάνονα των λογαρίθμων τελικά λαμβάνουμε:

$$\log p(x_i) \geq \int q_\theta(z|x_i)\Big[\log p_\phi(x_i|z) + \log p(z) - \log q_\theta(z|x_i)\Big]dz.$$

Εκφράζοντας το δεξί μέλος της ανισότητας ως την αναμενόμενη τιμή:

$$\log p(x_i) \geq E_{\sim q_\theta(z|x_i)}\Big[\log p(x_i, z) - \log q_\theta(z|x_i)\Big]$$

Τελικά αναλύοντας την τελευταία εξίσωση λαμβάνουμε τον τελικό τύπο:

$$\log p(x_i) \geq \int q_\theta(z|x_i) \log \left(\frac{p(z)}{q_\theta(z|x_i)} \right) dz + \int q_\theta(z|x_i) \log p_\phi(x_i|z) dz$$

$$\log p(x_i) \geq -D_{KL}(q_\theta(z|x_i)||p(z)) + E_{\sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)]$$

Το δεξί μέλος αυτής της ανισότητας είναι το ELBO.

Παράρτημα Β: Εξαγωγή συνάρτησης κόστους VAE

Αφού έχουμε επιλέξει γκαουσιανές κατανομές για το μοντέλο μας, ο όρος της απόκλισης KL (όρος κανονικοποίησης) γράφεται ως:

$$-D_{KL}(q_\theta(z|x_i)||p(z)) =$$

$$\int \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right) \log\left(\frac{\frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right)}\right) dz$$

Υπολογίζοντας τις παραστάσεις μέσα στο λογάριθμο η εξίσωση γίνεται:

$$\int \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right) \times$$
$$\left\{-\frac{1}{2}\log(2\pi) - \log(\sigma_p) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{1}{2}\log(2\pi) + \log(\sigma_q) + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right\} dz.$$

Με επιπλέον απλοποιήσεις λαμβάνουμε:

$$\frac{1}{\sqrt{2\pi\sigma_q^2}} \int \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right) \left\{-\log(\sigma_p) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \log(\sigma_q) + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right\} dz,$$
$$= \frac{1}{\sqrt{2\pi\sigma_q^2}} \int \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right) \left\{\log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right\} dz.$$

Εκφράζοντας το παραπάνω σαν αναμενόμενη τιμή:

$$\begin{aligned} -D_{KL}(q_\theta(z|x_i)||p(z)) &= E_q\left\{\log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right\} \\ &= \log\left(\frac{\sigma_q}{\sigma_p}\right) + E_q\left\{-\frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right\} \\ &= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_p)^2\right\} + \frac{1}{2\sigma_q^2}E_q\left\{(x-\mu_q)^2\right\} \end{aligned}$$

Εφόσον η διακύμανση σ^2 είναι η αναμενόμενη τιμή του τετραγώνου της απόστασης από τη μέση τιμή,

$$\sigma_q^2 = E_q\left\{(x-\mu_q)^2\right\},$$

Εχουμε τελικά:

$$\begin{aligned}
-D_{KL}(q_\theta(z|x_i)||p(z)) &= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_p)^2\right\} + \frac{\sigma_q^2}{2\sigma_q^2} \\
&= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_p)^2\right\} + \frac{1}{2} \\
&= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_q+\mu_q-\mu_p)^2\right\} + \frac{1}{2} \\
&= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(\underbrace{x-\mu_q}_{a}+\underbrace{\mu_q-\mu_p}_{b})^2\right\} + \frac{1}{2}
\end{aligned}$$

Κάνοντας χρήση της ταυτότητας $(\alpha+\beta)^2 = \alpha^2 + 2\alpha\beta + \beta^2$ έχουμε:

$$\begin{aligned}
-D_{KL}(q_\theta(z|x_i)||p(z)) &= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(\underbrace{x-\mu_q}_{a}+\underbrace{\mu_q-\mu_p}_{b})^2\right\} + \frac{1}{2} \\
&= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_q)^2 + 2(x-\mu_q)(\mu_q-\mu_p) + (\mu_q-\mu_p)^2\right\} + \frac{1}{2} \\
&= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_q)^2 + 2(x-\mu_q)(\mu_q-\mu_p) + (\mu_q-\mu_p)^2\right\} + \frac{1}{2} \\
&= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}\left[E_q\left\{(x-\mu_q)^2\right\} + 2E_q\left\{(x-\mu_q)(\mu_q-\mu_p)\right\} + E_q\left\{(\mu_q-\mu_p)^2\right\}\right] + \frac{1}{2} \\
&= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}\left[\sigma_q^2 + 2*0*(\mu_q-\mu_p) + (\mu_q-\mu_p)^2\right] + \frac{1}{2} \\
&= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{\sigma_q^2 + (\mu_q-\mu_p)^2}{2\sigma_p^2} + \frac{1}{2}
\end{aligned}$$

και αν επιλέξουμε $\sigma_p = 1$ και $\mu_p = 0$:

$$\begin{aligned} -D_{KL}(q_\theta(z|x_i)||p(z)) &= \log(\sigma_q) - \frac{\sigma_q^2 + \mu_q^2}{2} + \frac{1}{2} \\ &= \frac{1}{2}\log(\sigma_q^2) - \frac{\sigma_q^2 + \mu_q^2}{2} + \frac{1}{2} \\ &= \frac{1}{2}\left[1 + \log(\sigma_q^2) - \sigma_q^2 - \mu_q^2\right] \end{aligned}$$

Με βάση την εξίσωση (5) της ενότητας 2.3.2 (σελ. 37), έχουμε ότι για ένα δεδομένο x_i η ποσότητα που θέλουμε να μεγιστοποιήσουμε γίνεται:

$$\frac{1}{2}\left[1 + \log(\sigma_j^2) - \sigma_j^2 - \mu_j^2\right] + E_{\sim q_\theta(z|x_i)}\left[\log p_\phi(x_i|z)\right]$$

Όπου σ_j^2 και μ_j είναι παράμετροι της μέσα στην κατανομή q και j είναι ο δείκτης της ενδιάμεσης αναπαράστασης z. Για ένα πακέτο δεδομένων (batch):

$$\mathcal{G} = \sum_{j=1}^{J} \frac{1}{2}\left[1 + \log(\sigma_i^2) - \sigma_i^2 - \mu_i^2\right] + \frac{1}{L}\sum_{l} E_{\sim q_\theta(z|x_i)}\left[\log p(x_i|z^{(i,l)})\right]$$

Όπου J η διάσταση της λανθάνουσας ή ενδιάμεσης αναπαράστασης και L, το πλήθος των δειγμάτων που έχουμε δειγματοληπτηθεί, βάση πάντα του τεχνάσματος της αναπαραμετροποίησης (reparametrization trick). Εφόσον η προηγούμενη συνάρτηση είναι μια συνάρτηση προς μεγιστοποίηση, προκειμένου να λάβουμε την συνάρτηση κόστους αρκεί να πάρουμε το αρνητικό της προς μεγιστοποίηση συνάρτησης. Τελικά η συνάρτηση κόστους του VAE είναι:

$$\mathscr{L} = -\sum_{j=1}^{J} \frac{1}{2}\left[1 + \log(\sigma_i^2) - \sigma_i^2 - \mu_i^2\right] - \frac{1}{L}\sum_{l} E_{\sim q_\theta(z|x_i)}\left[\log p(x_i|z^{(i,l)})\right]$$

Και προκειμένου να εκπαιδεύσουμε τον VAE, ψάχνουμε τις βέλτιστες παραμέτρους (θ^*, φ^*) που ελαχιστοποιούν την συνάρτηση $\mathscr{L}$:

$$(\theta^*, \phi^*) = argmin_{(\theta,\phi)}\mathcal{L}(\theta, \phi)$$

Βιβλιογραφία

Blei, D. M., Kucukelbir, A., & Mcauliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association, 112*(518), 859-877. doi: 10.1080/01621459.2017.1285773

Carnap, R., D. Hilbert, W., & B.. Russell, A. (1970, January 01). A logical calculus of the ideas immanent in nervous activity. Retrieved from https://link.springer.com/article/10.1007/BF02478259

Chen, L., Dai, S., Pu, Y., Li, C., Su, Q., & Carin, L. (2017, October 19). Symmetric Variational Autoencoder and Connections to Adversarial Learning. Retrieved from https://arxiv.org/abs/1709.01846

Choi, K., Fazekas, G., Cho, K., & Sandler, M. (2018, June 03). A Comparison of Audio Signal Preprocessing Methods for Deep Neural Networks on Music Tagging. Retrieved from https://arxiv.org/abs/1709.01922

Christensen, C. B., Lauridsen, H., Christensen-Dalsgaard, J., Pedersen, M., & Madsen, P. T. (2015). Better than fish on land? Hearing across metamorphosis in salamanders. *Proceedings of the Royal Society B: Biological Sciences,282*(1802), 20141943. doi:10.1098/rspb.2014.1943

Dhariwal, P., Jun, H., Payne, C., Kim, J., Radford, A., & Sutskever, I. (2020, April 30). Jukebox: A Generative Model for Music. Retrieved from https://arxiv.org/abs/2005.00341

Doersch, C. (2016, August 13). Tutorial on Variational Autoencoders. Retrieved from https://arxiv.org/abs/1606.05908

Foster, D. (2019). *Generative deep learning: Teaching machines to paint, write, compose, and play*. Beijing ; Boston ; Farnham ; Sebastopol ; Tokyo: O'Reilly.

Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep learning*. Cambridge, MA: MIT Press.

Griffin, D., & Lim, J. (1983). Signal estimation from modified short-time Fourier transform. *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*. doi:10.1109/icassp.1983.1172092

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology, 195*(1), 215-243. doi:10.1113/jphysiol.1968.sp008455

Huzaifah, M. (2017, June 22). Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks. Retrieved from https://arxiv.org/abs/1706.07156

Jordan, M. I. (1998). *An introduction to variational methods for graphical models* (pp. 183-233). Berkeley, CA: University of California, Berkeley, Computer Science Division.

Kingma, D., & Welling, M. (2014, May 01). Auto-Encoding Variational Bayes. Retrieved from https://arxiv.org/abs/1312.6114

Kuleshov, V., & Ermon, S. (n.d.). Variational inference. Retrieved from https://ermongroup.github.io/cs228-notes/inference/variational/

Mann, C. R., & Twiss, G. R. (1911). *Physics*. Toronto: Educational Book.

Marr, B. (2019, May 13). What Is The Difference Between Artificial Intelligence And Machine Learning? Retrieved from https://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/

Mehlig, B. (2019, February 01). Artificial Neural Networks. Retrieved July 12, 2020, from https://arxiv.org/abs/1901.05639

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Mitra, S. K. (2001). *Digital signal processing: A computer-based approach*. Boston: McGraw-Hill/Irwin.

Naesseth, C. A. (2018). Pp. 21-30. In *Machine learning using approximate inference: Variational and sequential Monte Carlo methods*. Linköping: Linköping University Electronic Press.

Nordby, J. (2020). Environmental Sound Classification on Microcontrollers using Convolutional Neural Networks. Retrieved from https://paperswithcode.com/paper/environmental-sound-classification-on

Odaibo, S. (2019, July 21). Tutorial: Deriving the Standard Variational Autoencoder (VAE) Loss Function. Retrieved from https://arxiv.org/abs/1907.08956

Perraudin, N., Balazs, P., & Sondergaard, P. L. (2013). A fast Griffin-Lim algorithm. *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. doi:10.1109/waspaa.2013.6701851

Aarhus University. (2015, February 6). Researchers reveal how hearing evolved. *ScienceDaily*. Retrieved from www.sciencedaily.com/releases/2015/02/150206125257.htm

Salamon, J., & Bello, J. (2016, November 28). Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. Retrieved from https://arxiv.org/abs/1608.04363

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016, June 10). Improved Techniques for Training GANs. Retrieved from https://arxiv.org/abs/1606.03498

Schmidhuber, J. (2014, April 30). Deep Learning in Neural Networks: An Overview. Retrieved from https://arxiv.org/abs/1404.7828v1

Sharma, J., Granmo, O., & Goodwin, M. (2020). Environment Sound Classification Using Multiple Feature Channels and Attention Based Deep Convolutional Neural Network. *Interspeech 2020*. doi:10.21437/interspeech.2020-1303

Shuvaev, S., Giaffar, H., & Koulakov, A. (2017, December 08). Representations of Sound in Deep Learning of Audio Features from Music. Retrieved from https://arxiv.org/abs/1712.02898

Zell, A. (2003). *Simulation neuronaler Netze*. München: Oldenbourg.