* Εφόσον φτιάξω την αρχιτεκτονική του μοντέλου εάν έχω όρεξη και υπάρχει νόημα και είναι πραγματοποιίσημο μπορώ να προσθέσω μηχανισμό attention στο CNN όπως στα papers «Speech Emotion Recognition Using Convolutional Neural Networks with Attention Mechanism» και «Attention Based Fully Convolutional Network for

Speech Emotion Recognition» (1806.01506v2.pdf)

* To «Speech emotion recognition with deep convolutional neural networks» (main.pdf) φαίνεται να έχει οκ αποτελέσματα και να είναι πολύ απλό. Από αρνητικά έχει ότι είναι 1d CNN. Θέλω να αντιγράψω ότι βάζει στο CNN για κάθε audio sample πολλούς διαφορετικούς μετασχηματισμούς έτσι ώστε το CNN να μπορεί να μάθει το ίδιο σήμα ήχου από πολλές διαφορετικές οπτικές γωνίες. Όπως κάνουν στα CNN για εικόνες rgb αλλά εδώ να είναι πχ STFT , Μel\_spectrogram και MFCCs.
* Μάλλον θα αντιγράψω την αρχιτεκτονική του «FSER» και θα πειραματιστώ προσθέτοντας είτε extra dropout layers, είτε προσθέτοντας batch normalization layers (πριν την ReLU) , είτε βάζοντας σαν είσοδο με βάθος πολλών μετασχηματισμών, είτε μεταγενέστερα προσθέτοντας το attention μηχανισμό.
* Υπάρχει νόημα να βάλω όλους αυτούς του μετασχηματισμούς στην είσοδο ή η πληροφορία που προσφέρουν επικαλύπτετε πχ για να φτάσουμε στο mel\_spectogram κάνουμε fft και για να πάμε στα MFCCs θέλουμε το mel\_spectogram.
* Είδα ότι τα 2d CNN χρησιμοποιούν mel\_spectogram για emotional\_ recognition και τα 1d CNN MFCCs (και αλλά συμπληρωματικά) όμως τα MFCCs ώς γράφημα προς τον χρόνο γίνονται σαν mel\_spectrogram (που είναι x = χρόνος και y = mel συχνότητες) το MFCCs ως προς τον χρόνο θα είναι σαν (x = χρόνος και y = mfccs ) . Θα είναι επομένως πιο συμπυκνωμένες με την ίδια πληροφορία και επομένως οι διαστάσεις του input και των feature maps θα είναι μικρότερες, άρα λιγότερο max pooling και λιγότεροι υπολογισμοί.
* Το CNN είτε θα έχει είσοδο μεταβλητού μεγέθους (δύσκολο) είτε θα παίρνει σαν είσοδο το median time + ένα ποσό που θα βρούμε πειραματικά των audio samples και τα μικρότερα θα συμπληρώνονται με zero padding και τα μεγαλύτερα θα κόβονται.
* Το mel\_spectrogram μπορώ να το επεξεργαστώ και σαν rgb εικόνα «FSER» είτε ίσως σαν grayscale εικόνα, το rgb είναι καλήτερο φαντάζομαι αλλά αυξάνει το βάθος του input.
* Πιστεύετε πως εάν έκανα την αναφορά στα αγγλικά θα μπορούσα να βάλω την διπλωματική στο βιογραφικό και μετά να κάνω Erasmus και πρακτική για να πάω να κάνω πρακτική σε εταιρία στο εξωτερικό, το εαρινό εξάμηνο που έρχετε;;