

# Αναγνώριση Νοηματικής Γλώσσας σε Πραγματικό Χρόνο με Τεχνικές Βαθιάς Μάθησης

Σπύρος Σάββας

Πρόγραμμα Μεταπτυχιακών Σπουδών (Π.Μ.Σ.)

Τμήμα Ψηφιακών Συστημάτων

Μεγάλα Δεδομένα και Αναλυτική

Πανεπιστήμιο Πειραιώς

Πειραιάς, Ελλάδα

spyros\_savvas@hotmail.com

## ABSTRACT

Η δημιουργία ενός μοντέλου Αναγνώρισης Νοηματικής Γλώσσας με Τεχνικές Βαθιάς Μάθησης είναι απαραίτητη για διάφορους λόγους. Πρότον, επιτρέπει την γεφύρωση του επικοινωνιακού χάσματος μεταξύ της κοινότητας των κωφών και βαρήκοων ατόμων με τον υπόλοιπο πληθυσμό, προάγοντας την κοινωνική ενσωμάτωση. Ένα τέτοιο μοντέλο μπορεί να διευκολύνει την άμεση μετάφραση της νοηματικής γλώσσας σε προφορικό ή γραπτό λόγο, επιτρέποντας την άμεση κατανόηση και επικοινωνία σε πραγματικό χρόνο. Επιπλέον, η αυτοματοποιημένη αναγνώριση της νοηματικής γλώσσας μπορεί να χρησιμοποιηθεί σε διάφορες εφαρμογές, όπως η εκπαίδευση, η εργασία καθώς και σε καθημερινές εφαρμογές βελτιώνοντας την ποιότητα ζωής των ατόμων που χρησιμοποιούν τη νοηματική γλώσσα. Στην παρούσα εργασία το μοντέλο νοηματικής γλώσσας σε πραγματικό χρόνο αναπτύσσεται για την αναγνώριση νοημάτων της Αμερικανικής Νοηματικής Γλώσσας. Για την αναγνώριση των νοημάτων και για την δημιουργία του συνόλου δεδομένων μας χρησιμοποιούμε τις βιβλιοθήκες OpenCV και στην συνέχεια την MediaPipe που συλλέγει βασικά σημεία (landmarks) των χεριών και του προσώπου. Στην συνέχεια εκπαιδεύουμε το νευρωνικό μοντέλο με την χρήση TensorFlow, Keras και LSTM. Τέλος, το μοντέλο μπορεί να δοκιμαστεί σε πραγματικό χρόνο με τη λήγη ζωντανής μετάδοσης από την κάμερα.

## I. ΕΙΣΑΓΩΓΗ

Οι νοηματικές χρησιμοποιούνται ως έναν διαφορετικό τρόπο επικοινωνίας από αυτόν που έχουμε συνηθίσει στην καθημερινή μας ζωή. Οι νοηματικές γλώσσες

χρησιμοποιούν τη φυσική και οπτική διαδρομή που παράγεται από το σώμα και γίνεται αντιληπτή από τα μάτια του δέκτη. Αυτό φυσικά δεν ισχύει για τις ομιλούμενες γλώσσες. Οι ομιλούμενες γλώσσες χρησιμοποιούν τη φωνητική - ακουστική οδό, δηλαδή ο ομιλητής αρθρώνει προφορικά τις λέξεις του και αυτές γίνονται αντιληπτές από το αυτί του δέκτη. Δυστυχώς οι νοηματικές γλώσσες δεν είναι διεθνείς, οπότε όλοι οι κωφοί στον κόσμο δεν μπορούν να επικοινωνήσουν μεταξύ τους κοινή γνώση της νοηματικής γλώσσας. Οι νοηματικές γλώσσες αναπτύσσονται όταν ομάδες κωφών συγκεντρώνονται και επικοινωνούν μεταξύ τους. Κάθε χώρα αναπτύσσει τη νοηματική της γλώσσα με διαφορετικές μεθόδους τις οποίες θα συζητήσουμε παρακάτω.

Στην πραγματικότητα τα νοήματα που χρησιμοποιούν οι Κωφοί έχουν παρόμοια δομή με τον τρόπο που προφέρονται οι προφορικές λέξεις. αν πάρουμε ως παράδειγμα την αγγλική γλώσσα και πόσες από τις λέξεις της παράγονται στον προφορικό λόγο θα βρούμε πολλές αντιστοιχίες στη δημιουργία των λέξεων στη νοηματική γλώσσα. Όπως οι προφορικές λέξεις παράγονται από έναν μικρό αριθμό διαφορετικών ήχων, έτσι και τα νοήματα παράγονται από πολλές χειρονομίες. Έτσι, τα νοήματα δεν είναι ολιστικές χειρονομίες αλλά αναλύσιμες χειρονομίες. Δεν είναι απλώς ένα σύνολο χειρονομιών, αλλά ένα σύνολο χειρονομιών που μπορούν να αναλυθούν.

Πριν από την ανάπτυξη της βαθιάς μάθησης, οι επιστήμονες εργάζονταν κυρίως σε μεμονωμένες γλώσσες και χειρονομίες, καθώς η ταξινόμηση γινόταν μόνο μεταξύ διαφορετικών λέξεων και στατικών πλαισίων χειρονομιών. Με τη βαθιά μάθηση, είναι πλέον δυνατή η εκπαίδευση ενός συστήματος μέσω κινούμενων εικόνων, όπως βίντεο, ενώ στο παρελθόν οι μόνες πληροφορίες που μπορούσαν να δοθούν σε ένα σύστημα ήταν με τη μορφή εικόνας. Η αναγνώριση της νοηματικής γλώσσας συνδέεται άμεσα με την ώραση υπολογιστών και, ως εκ τούτου, οι περισσότερες προσεγγίσεις που ασχολούνται με την αναγνώριση της

νοηματικής γλώσσας έχουν προσαρμοστεί προς αυτή την κατεύθυνση.

Στην παρούσα εργασία, ο χρήστης θα μπορεί να δημιουργήσει το δικό του σύνολο δεδομένων και να το αποθηκεύσει σε οποιονδήποτε επιθυμητό αποθηκευτικό χώρο. Θα μπορεί να καταγράψει κινήσεις που υποδηλώνουν λέξεις, φράσεις ή χειρονομίες σε μορφή βίντεο. Στη συνέχεια, όλα τα αποθηκευμένα καρέ θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου. Το μοντέλο θα προβλέπει τα αποτελέσματα που πιστεύει σύμφωνα με τις χειρονομίες που εκτελεί ο χρήστης. Για την υλοποίηση θα χρησιμοποιηθούν ορισμένες βασικές βιβλιοθήκες όπως MediaPipe/Holistic της Google, Tensorflow, OpenCV για την καταγραφή και την Matplotlib για την οπτικοποίηση.

## II. ΜΕΘΟΔΟΛΟΓΙΑ

### A. OpenCV

Η βιβλιοθήκη OpenCV (Open Source Computer Vision Library) είναι μια ισχυρή και ευρέως χρησιμοποιούμενη βιβλιοθήκη ανοιχτού κώδικα για επεξεργασία εικόνας και μηχανική όραση. Αναπτύχθηκε αρχικά από την Intel και τώρα υποστηρίζεται από την κοινότητα της ανοιχτής πηγής. Περιλαμβάνει περισσότερες από 2500 αλγορίθμους που επιτρέπουν την εκτέλεση ενός ευρέος φάσματος εργασιών επεξεργασίας εικόνας, όπως ανίχνευση αντικειμένων, αναγνώριση προσώπου και επεξεργασία βίντεο. Η OpenCV υποστηρίζει πολλαπλές γλώσσες προγραμματισμού, όπως Python, C++, Java και MATLAB, καθιστώντας την ευέλικτη και προσιτή σε προγραμματιστές με διαφορετικά υπόβαθρα. Εφαρμόζεται σε διάφορους τομείς, από την ιατρική απεικόνιση και την παρακολούθηση ασφαλείας έως την επαυξημένη πραγματικότητα και την ρομποτική, καθιστώντας την μια από τις βασικές εργαλειοθήκες για επαγγελματίες και ερευνητές στον τομέα της υπολογιστικής όρασης.

### B. MediaPipe-Holistic

Η βιβλιοθήκη MediaPipe - Holistic είναι ένα ισχυρό εργαλείο ανοιχτού κώδικα από την Google που επιτρέπει την ολοκληρωμένη ανάλυση και παρακολούθηση ανθρώπινων κινήσεων μέσω βίντεο ή πραγματικού χρόνου δεδομένων. Συνδυάζει πολλαπλά υποσυστήματα ανίχνευσης, όπως ανίχνευση προσώπου, ανίχνευση χεριών και ανίχνευση σώματος, παρέχοντας μια ολοκληρωμένη λύση για την παρακολούθηση ολόκληρου του σώματος. Το MediaPipe - Holistic χρησιμοποιεί προηγμένους αλγορίθμους μηχανικής μάθησης για την ακριβή αναγνώριση και ανίχνευση βασικών σημείων (landmarks) σε κάθε μέρος του σώματος, όπως τα μάτια, η μύτη, τα

δάχτυλα και οι αρθρώσεις. Αυτή η βιβλιοθήκη είναι ιδανική για εφαρμογές όπως η ανάλυση στάσης, η αναγνώριση χειρονομιών και η βελτιωμένη αλληλεπίδραση ανθρώπου-υπολογιστή. Η υποστήριξη για πολλές πλατφόρμες, όπως Android, iOS και web, καθώς και η εύκολη ενσωμάτωση με άλλες τεχνολογίες της MediaPipe, καθιστούν το Holistic ένα εξαιρετικά ευέλικτο και χρήσιμο εργαλείο για προγραμματιστές και ερευνητές στον τομέα της επεξεργασίας εικόνας και της υπολογιστικής όρασης.

### C. Recurrent Neural Networks (RNNs)

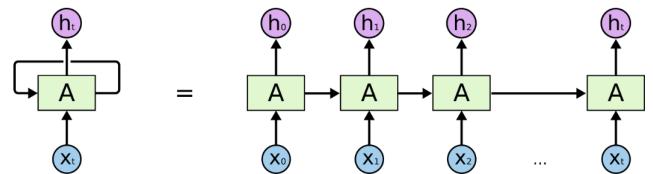
Τα Recurrent Neural Networks (RNNs) ανήκουν στην οικογένεια των Feed-Forward Neural Networks και έχουν ως κύριο γνώρισμα τους τη δυνατότητα να επεκτείνονται σε γειτονικά χρονικά στάδια. Με αυτό τον τρόπο ένα κόμβος του δικτύου έχει τη δυνατότητα σε κάθε χρονική στιγμή να παίρνει τα τρέχοντα δεδομένα εισόδου, παράλληλα με τις τιμές των κρυμμένων κόμβων, συλλέγοντας έτσι πληροφορίες από προηγούμενες χρονικές στιγμές.

Δεδομένης μιας ακολουθίας εισόδου  $x = (x^0, \dots, x^{T-1})$ , τα κρυφά επίπεδα ενός επαναλαμβανόμενου στρώματος  $h = (h^0, \dots, h^{T-1})$  και η έξοδος ενός RNN με ένα κρυφό στρώμα  $y = (y^0, \dots, y^{T-1})$  μπορεί να υπολογιστεί ως εξής:

$$h^t = \sigma(W_{xh}x^t + W_{hh}h^{t-1} + b_h)$$

$$y^t = O(W_{ho}x^t + b_o)$$

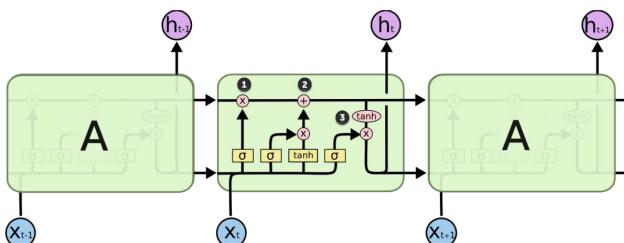
Όπου  $W_{xh}, W_{hh}, W_{ho}$  συμβολίζουν τα βάρη σύνδεσης από το στρώμα εισόδου  $x$  προς το κρυφό στρώμα  $h$ ,  $b_h$  και  $b_o$  δύο διανύσματα bias,  $\sigma()$  και  $O()$  είναι οι συναρτήσεις ενεργοποίησης στο κρυφό στρώμα και στο στρώμα εξόδου αντίστοιχα.



Εικόνα 1: Αρχιτεκτονική RNN

### D. Long-Short Term Memory Neural Networks (LSTMs)

Πέραν των αποδεδειγμένων αποτελεσματικών εφαρμογών των παραπάνω ανατροφοδοτούμενων δικτύων (recurrent networks), παρουσιάζουν και κάποιες αδυναμίες. Οι αδυναμίες αυτές εμφανίζονται στην περίπτωση όπου το χρονικό διάστημα μεταξύ εισόδου και εξόδου είναι υπολογίσιμο. Τα δίκτυα LSTM (Long Short-Term Memory) έρχονται για να λύσουν αυτό το πρόβλημα των χρονικών κενών. Τα LSTM δίκτυα παρουσιάστηκαν από τον Schmidhuber το 1997 και ακολουθούν την αρχιτεκτονική των RNN δικτύων. Η αρχιτεκτονική τους περιλαμβάνει πύλες για τον έλεγχο της πληροφορίας μεταξύ των κελιών. Κύριο ρόλο στην λειτουργία του νευρωνικού παίζει το κελί της μνήμης. Κάθε επαναλαμβανόμενη μονάδα της αλυσίδας του νευρωνικού δικτύου αποτελείται από πύλες και κελιά. Η κατάσταση του κελιού της μνήμης μεταβάλεται από τις πύλες οι οποίες προσθέτουν ή απομακρύνουν πληροφορία από το κελί. Ποιο συγκεκριμένα, οι πύλες που συνθέτουν την δομή είναι η πύλη επιλεκτικής συγκράτησης (forget gate), η πύλη εισόδου και η πύλη εξόδου. Η κάθε πύλη εξαρτάται από μία συνάρτηση ενεργοποίησης. Στην παρακάτω εικόνα παρατηρούμε την αλυσίδα ενός απλού LSTM νευρωνικού. Οι πύλες 1-(forget gate), 2-Input Gate και 3-Output Gate που παρουσιάζονται με συνδυασμούς της σιγμοειδείς και μίας εφαπτομένης συνάρτησης.



Εικόνα 2: Πύλες ενός LSTM δικτύου

Η παραπάνω αλυσίδα περιγράφει την αρχιτεκτονική ενός απλού LSTM νευρωνικού δικτύου.

### E. Αλγόριθμος Backpropagation

Ο αλγόριθμος backpropagation είναι μια μέθοδος εκπαίδευσης τεχνητών νευρωνικών δικτύων, ιδιαίτερα των πολυεπίπεδων νευρωνικών δικτύων (multi-layer neural networks). Η βασική ιδέα του αλγορίθμου είναι η προσαρμογή των βαρών των συνδέσεων μεταξύ των νευρώνων του δικτύου έτσι ώστε να μειώνεται το σφάλμα της εξόδου του δικτύου. Η διαδικασία αυτή πραγματοποιείται σε δύο φάσεις: τη φάση της προώθησης και τη φάση της οπισθοδιάδοσης. Στη φάση της προώθησης, τα δεδομένα εισόδου διέρχονται μέσω των επιπέδων του δικτύου και παράγουν μια έξοδο. Στη συνέχεια, υπολογίζεται το σφάλμα με βάση τη διαφορά μεταξύ της προβλεπόμενης εξόδου και της επιθυμητής

εξόδου. Στη φάση της οπισθοδιάδοσης, το σφάλμα αυτό διαδίδεται προς τα πίσω μέσω του δικτύου, από την έξοδο προς την είσοδο, και τα βάρη των συνδέσεων προσαρμόζονται με βάση την κλίση του σφάλματος σε σχέση με κάθε βάρος. Ο αλγόριθμος χρησιμοποιεί τη μέθοδο του βαθμωτού κατήφορου (gradient descent) για την ελαχιστοποίηση του σφάλματος, κάνοντας μικρές αλλαγές στα βάρη ώστε να μειωθεί η συνολική απόκλιση. Μέσω αυτής της επαναληπτικής διαδικασίας, το νευρωνικό δίκτυο μαθαίνει να αναγνωρίζει μοτίβα και να παρέχει ακριβείς προβλέψεις.

### F. Μετρικές Απόδοσης

Προκειμένου να αξιολογηθεί η αποτελεσματικότητα ενός μοντέλου, πρέπει να εξεταστούν ορισμένες μετρικές απόδοσης. Θα πρέπει να εξετάζονται πολλαπλές μετρικές των μοντέλων, διότι η αξιολόγηση αυτών των τιμών ως ενιαίο κριτήριο επιτυχίας θα ήταν εσφαλμένη. Όλες οι παρατηρήσεις στο σύνολο δεδομένων δοκιμής αντικαθίστανται στο μοντέλο που έχει δημιουργηθεί με το σύνολο δεδομένων εκπαίδευσης και επιτυγχάνονται βαθμολογίες πρόβλεψης ανίχνευσης. Τα αποτελέσματα της σύγκρισης των προβλεπόμενων τιμών με τις πραγματικές τιμές χρησιμοποιούνται για να προσδιοριστεί πόσο καλά προβλέπει αυτό το μοντέλο, καθώς και η επιτυχία και η απόδοσή του. Ο πίνακας σύγχυσης (confusion matrix) συνοψίζει τα αποτελέσματα της ακρίβειας του μοντέλου στην πρόβλεψη, καθώς και τα συμπεράσματα της αξιολόγησης των επιδόσεων του μοντέλου.

Confusion Matrix		Actual Values	
		1	0
Predicted Values	1	True Positive TP	False Positive FP
	0	False Negative FN	True Negative TN

Πίνακας 1: Πίνακας Σύγχυσης (Confusion Matrix)

Αληθώς θετική (TP): Δείχνει ότι οι παρατηρήσεις με πραγματική τιμή κλάσης 1 προβλέφθηκαν σωστά ως 1.

Αληθώς αρνητική (TN): Δείχνει ότι οι παρατηρήσεις με πραγματική τιμή κλάσης 0 προβλέπονται σωστά ως 0.

Ψευδώς αρνητική (FN): Δείχνει ότι οι παρατηρήσεις με πραγματική τιμή κλάσης 1 αξιολογούνται εσφαλμένα ως 0 ως αποτέλεσμα της πρόβλεψης.

Ψευδώς θετική (FP): Δείχνει ότι παρατηρήσεις με πραγματική τιμή κλάσης 0 αξιολογούνται εσφαλμένα ως 1 ως αποτέλεσμα της πρόβλεψης.

Η ακρίβεια (accuracy) ή αλλιώς ορθότητα, είναι το πιο συνηθισμένο μέτρο απόδοσης ενός μοντέλου. Η ορθότητα είναι η αναλογία των παραδειγμάτων για τα οποία έχει γίνει η σωστή πρόβλεψη επί του συνόλου των προβλέψεων και με βάση τις παραπάνω συμβάσεις και ισοδυναμεί με τον δείκτη λάθους (error rate) του μοντέλου και δίνεται από τον τύπο:

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

Η ανάκληση (recall), γνωστή και ως ευαισθησία (sensitivity) ή δείκτης αληθώς θετικών (True Positive Rate – TPR), είναι η αναλογία κάθε θετικού παραδειγμάτος που είναι πραγματικά θετικό. Αφορά την ικανότητα του μοντέλου να αναγνωρίζει ένα παράδειγμα της θετικής κλάσης και υπολογίζεται από τον τύπο:

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

Το Precision, επίσης γνωστό ως δείκτης αληθώς θετικών προβλέψεων, είναι η αναλογία των πραγματικά θετικών παραδειγμάτων μεταξύ όλων των παραδειγμάτων που έχουν προβλεφθεί ως θετικά. Αυτό το μέτρο αναφέρεται στην ικανότητα του μοντέλου να μην κατατάσσει ένα αρνητικό παράδειγμα ως θετικό. Ο τύπος για το Precision είναι:

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

Σε πολλές περιπτώσεις, εάν θέλουμε να συνοψίσουμε την απόδοση του μοντέλου, επιδιώκοντας ένα είδος ισορροπίας μεταξύ της ακρίβειας και την ανάκλησης, χρησιμοποιείται ένα μέτρο απόδοσης που συνδυάζει τα δύο αυτά μέτρα, τον αρμονικό μέσο όρο της ακρίβειας και της ανάκλησης που ονομάζεται F1 αποτέλεσμα (F1 score) και υπολογίζεται από τον τύπο:

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{TP}{TP + \frac{FN+FP}{2}}$$

Η προσδιοριστικότητα (specificity), γνωστή και ως δείκτης αληθώς αρνητικών (True Negative Rate – TNR) μετρά τις ορθά αρνητικές προβλέψεις στο σύνολο των ορθών αρνητικών παραδειγμάτων και υπολογίζεται από τον τύπο:

$$Specificity = TNR = \frac{TN}{TN + FP}$$

Το ψευδώς θετικό ποσοστό (False Positive Rate - FPR) αντιστοιχεί στην αναλογία των αρνητικών παραδειγμάτων που θεωρήθηκαν ως θετικά, σε σχέση με όλα τα αρνητικά παραδείγματα. Όσο μεγαλύτερος είναι ο FPR, πόσα περισσότερα αρνητικά παραδείγματα έχουν ταξινομηθεί λάθος. Ο FPR υπολογίζεται από τον τύπο:

$$FPR = \frac{FP}{FP + TN} = 1 - specificity = 1 - TNR$$

### III. ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΟΥ

#### A. Περιγραφή των Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε για την υλοποίηση του μοντέλου αναγνώρισης νοηματικής γλώσσας σε πραγματικό χρόνο έγινε με τις βιβλιοθήκες OpenCV και MediaPipe. Αρχικά με την βοήθεια της OpenCV μας δίνει την δυνατότητα να ανοίξουμε τη κάμερα του υπολογιστή και στην συνέχεια μέσω της MediaPipe συλλέγουμε τα βασικά σημεία (landmarks) από διάφορα μέρη του σώματος.

Το MediaPipe Holistic παρέχει εκτίμηση της στάσης, του χεριού και των βασικών σημείων του προσώπου, η οποία βοηθά στον εντοπισμό των κινήσεων του χεριού και του σώματος στη νοηματική γλώσσα. Τα δεδομένα θα συλλέγονται σε πραγματικό χρόνο χρησιμοποιώντας ένα προκαθορισμένο σύνολο χειρονομιών της νοηματικής γλώσσας. Στην συνέχεια αποθηκεύονται τα εξαγόμενα βασικά στοιχεία από το MediaPipe - Holistic σε έναν πίνακα Numpy.

Η διαδικασία έχει ως εξής:

**Είσοδος Βίντεο:** Η MediaPipe – Holistic δέχεται το βίντεο που καταγράφεται από την κάμερα του υπολογιστή.

**Εκτίμηση πόζας:** Το πρώτο βήμα με τη χρήσης της MediaPipe είναι η ανίχνευση της στάσης του ατόμου μέσα στο καρέ του βίντεο. Αυτό συνεπάγεται τον εντοπισμό σημαντικών στοιχείων του σώματος, όπως το κεφάλι, οι ώμοι και τα χέρια.

**Εκτίμηση βασικών σημείων Χεριού και Προσώπου:** Αφού εκτιμηθεί η πόζα, το επόμενο

στάδιο είναι ο εντοπισμός των landmarks του προσώπου και του χεριού. Το MediaPipe Holistic εντοπίζει 468 σημεία-κλειδιά στο προσωπείο και 21 σημεία-κλειδιά σε κάθε χέρι.

**Ενσωμάτωση των βασικών σημείων:** Το MediaPipe Holistic συνδυάζει τα σημεία-κλειδιά για να δημιουργήσει μια αναπαράσταση του ατόμου στο καρέ του βίντεο. Συνδυασμός των βασικών σημείων σε έναν ενιαίο χώρο συντεταγμένων επιτρέπει την ακριβή παρακολούθηση των κινήσεων του ατόμου.

**Έξοδος:** Το MediaPipe Holistic παράγει ένα σύνολο βασικών σημείων που αντιπροσωπεύουν τη στάση, το χέρι και το πρόσωπο του ατόμου στο καρέ του βίντεο.



Εικόνα 3: Δείγμα εικόνας κατά την διαδικασία συλλογής του συνόλου δεδομένων

Για κάθε καρέ, συγκεντρώθηκαν συνολικά 1662 σημεία-κλειδιά με βάση τη στάση, τα ορόσημα των χεριών και του προσώπου με τη μορφή ενός ενιαίου χώρου συντεταγμένων. Εάν κάποιο από αυτά τα σημεία δεν παρακολουθούνταν από το μοντέλο, περάσαμε μηδενικές τιμές για να διατηρήσουμε το σχήμα του πίνακα.

Στην παρούσα εργασία το αυτοματοποιημένο μοντέλο αναγνώρισης της νοηματικής γλώσσας εκπαιδεύτηκε να αναγνωρίζει και να διαφοροποιεί τα παρακάτω νοήματα :

- hello
- thanks
- understand

Συλλέξαμε ένα ολοκληρωμένο σύνολο δεδομένων που αποτελείται από 30 βίντεο για κάθε μία από τις χειρονομίες "hello", "thanks" και "understand". Κάθε βίντεο είχε μήκος 30 καρέ. Αυτή η προσέγγιση

εξασφάλισε ότι το μοντέλο θα μπορούσε να μάθει από ένα σύνολο παραδειγμάτων αποτυπώνοντας διαφορετικούς τρόπους με τους οποίους μπορούν να εκτελεστούν αυτές οι χειρονομίες.

### B. Εκπαίδευση Μοντέλου

Η εκπαίδευση του μοντέλου είναι ένα κρίσιμο βήμα στην ανάπτυξη ενός συστήματος για την αναγνώριση της νοηματικής γλώσσας με τη χρήση τεχνικών βαθιάς μάθησης. Σε αυτή την ενότητα θα συζητηθεί η μεθοδολογία που χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου.

Αρχικά, το πρώτο επίπεδο του μοντέλου είναι ένα LSTM με 64 μονάδες, που είναι ο αριθμός των κελιών μνήμης στο στρώμα LSTM. Το σχήμα εισόδου αυτού του στρώματος ορίζεται ως (30, 1662), που σημαίνει ότι τα δεδομένα εισόδου είναι ένας τρισδιάστατος πίνακας με 30 χρονικά βήματα και 1662 χαρακτηριστικά. Η παράμετρος return\_sequences ορίζεται σε True, πράγμα που σημαίνει ότι η έξοδος αυτού του στρώματος θα είναι ένας 3D πίνακας με το ίδιο σχήμα με την είσοδο και μπορεί να χρησιμοποιηθεί ως είσοδος για το επόμενο επίπεδο. Η συνάρτηση ενεργοποίησης που χρησιμοποιείται είναι η «tanh» που χρησιμοποιείται για την εισαγωγή μη γραμμικότητας και για τη διατήρηση των τιμών εξόδου σε ένα εύρος [-1,1] και η συνάρτηση «sigmoid» περιορίζοντας τις τιμές εξόδου μεταξύ 0 και 1.

Το δεύτερο επίπεδο είναι επίσης ένα επίπεδο LSTM με 128 μονάδες και return\_sequences=True, αυτό το επίπεδο θα επεξεργαστεί την έξοδο από το πρώτο επίπεδο LSTM και θα παράγει μια νέα έξοδο.

Στα πρώτα δύο επίπεδα γίνεται η χρήση του Dropout (20%) για την αποφυγή της υπερπροσαρμογής.

Το τρίτο στρώμα είναι επίσης ένα LSTM με 64 μονάδες και return\_sequences=False, αυτό το στρώμα θα επεξεργαστεί την έξοδο από το δεύτερο στρώμα LSTM και θα παράγει έναν πίνακα 2D ως έξοδο.

Το τέταρτο και το πέμπτο στρώμα είναι πλήρως συνδεδεμένα στρώματα (γνωστά και ως πυκνά στρώματα) με 64 και 32 μονάδες αντίστοιχα. Η συνάρτηση ενεργοποίησης που χρησιμοποιείται σε αυτά τα στρώματα είναι η «relu» η οποία εισάγει μη γραμμικότητα στο μοντέλο.

Το τελευταίο στρώμα είναι ένα πυκνό στρώμα με αριθμό μονάδων ίσο με τον αριθμό των ενεργειών και η συνάρτηση ενεργοποίησης έχει οριστεί σε 'softmax', αυτό το στρώμα θα παράγει την τελική έξοδο του μοντέλου.

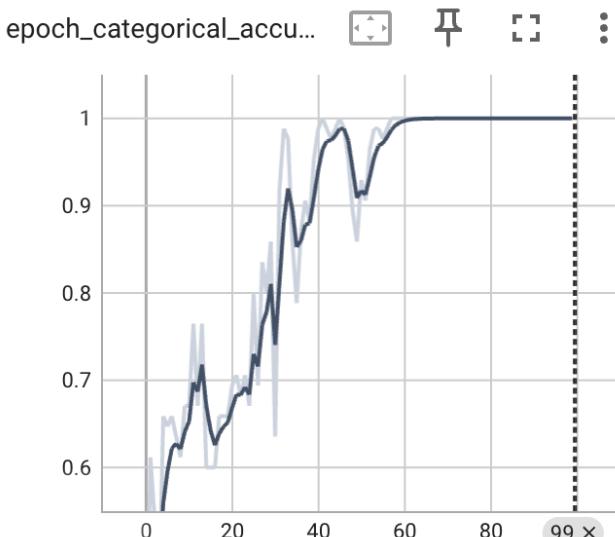
Στη συνέχεια, το μοντέλο καταρτίζεται με τη χρήση του βελτιστοποιητή Adam, ο οποίος είναι μια παραλλαγή της Stochastic Gradient Descent που προσαρμόζει το ρυθμό μάθησης για κάθε παράμετρο. Η συνάρτηση απώλειας που χρησιμοποιείται είναι η «categorical\_crossentropy», η οποία χρησιμοποιείται για προβλήματα ταξινόμησης πολλαπλών κλάσεων, και οι μετρικές που χρησιμοποιείται για την αξιολόγηση του μοντέλου είναι η 'categorical\_accuracy'.

Τέλος, το μοντέλο εκπαιδεύεται χρησιμοποιώντας τη συνάρτηση fit με τα δεδομένα εισόδου, τα δεδομένα εξόδου, τον αριθμό των εποχών και μια συνάρτηση callback. Η callback function είναι το Tensorboard, το οποίο είναι ένα εργαλείο οπτικοποίησης για την παρακολούθηση της προόδου της εκπαίδευσης και της απόδοσης του μοντέλου.

### C. Αξιόλογηση Μοντέλου

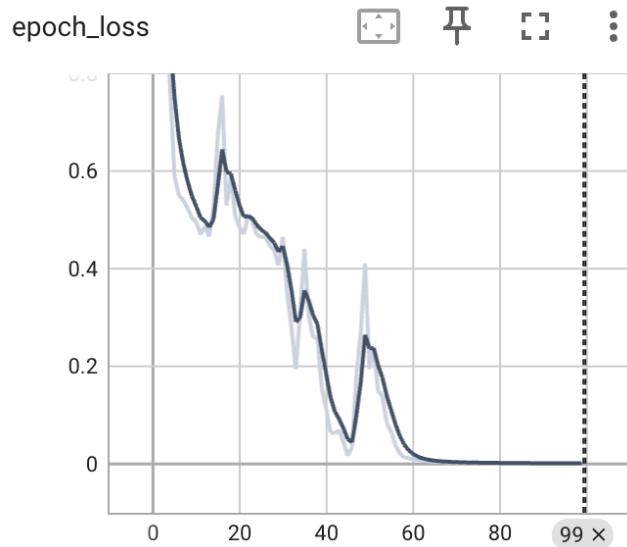
Η αξιολόγηση του μοντέλου αναγνώρισης νοηματικής γλώσσας πραγματοποιήθηκε με βάση την ανάλυση των μετρικών επίδοσης και την παρατήρηση των γραφημάτων εκπαίδευσης. Η αξιολόγηση περιλάμβανε τις εξής μετρικές: accuracy, recall, precision, f1-score, και πίνακα σύγχυσης. Επιπλέον, εξετάστηκαν τα γραφήματα epoch\_categorical\_accuracy και epoch\_loss.

Τα γραφήματα ακρίβειας (epoch\_categorical\_accuracy) και απώλειας (epoch\_loss) που θα δείτε παρακάτω παρουσιάζουν την πορεία εκπαίδευσης του μοντέλου κατά τη διάρκεια των 100 εποχών.



Γράφημα 1: Tensorboard γράφημα ακρίβειας ανά εποχή

Το γράφημα αυτό δείχνει την αύξηση της ακρίβειας του μοντέλου κατά την εκπαίδευση. Η ακρίβεια ξεκίνησε από περίπου 0.6 και σταδιακά αυξήθηκε, φτάνοντας το 1.0 μετά από περίπου 60 εποχές. Αυτή η αύξηση δείχνει ότι το μοντέλο βελτιώθηκε καθώς μάθαινε από τα δεδομένα εκπαίδευσης.



Γράφημα 2: Tensorboard γράφημα απώλειας ανά εποχή

Το γράφημα απώλειας δείχνει τη μείωση της απώλειας κατά την εκπαίδευση του μοντέλου. Η αρχική απώλεια ήταν περίπου 0.7 και σταδιακά μειώθηκε σχεδόν στο 0 μετά από περίπου 60 εποχές. Αυτό υποδεικνύει ότι το μοντέλο κατάφερε να μάθει τα μοτίβα των δεδομένων εκπαίδευσης και να μειώσει τα σφάλματα πρόβλεψης.

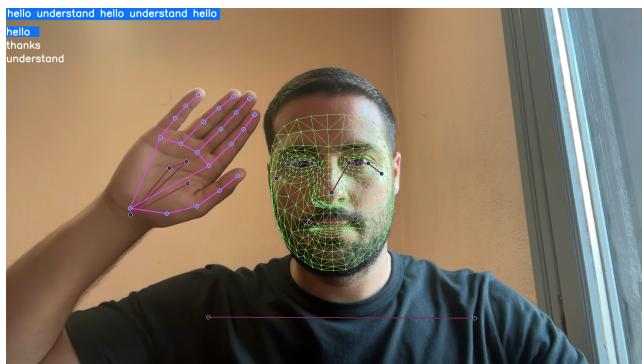
Επίσης οι μετρικές επίδοσης του μοντέλου ήταν εξαιρετικές, με όλες τις μετρικές να φτάνουν το μέγιστο δυνατό, όπως φαίνεται παρακάτω:

Accuracy	1.0
Recall	1.0
Precision	1.0
F1-Score	1.0

## IV. ΕΦΑΡΜΟΓΗ ΣΕ ΠΡΑΓΜΑΤΙΚΟ ΧΡΟΝΟ

Για να δοκιμάσουμε το μοντέλο ανίχνευσης νοηματικής γλώσσας σε πραγματικό χρόνο, υλοποιήσαμε ένα ολοκληρωμένο σύστημα που χρησιμοποιεί το OpenCV για την καταγραφή βίντεο, το Mediapipe για την ανίχνευση σημείων αναφοράς και ένα μοντέλο LSTM για την αναγνώριση χειρονομιών.

Η ροή από την κάμερα διαβάζεται συνεχώς καρέ προς καρέ και για κάθε καρέ η MediaPipe ανιχνεύει τα σημεία αναφοράς. Στην συνέχεια για να μπορεί ο χρήστης να βλέπει σε πραγματικό χρόνο τα αποτελέσματα των χειρονομιών που εκτελεί, φτιάχνουμε μια λογική στην οποία το πρόγραμμα εκτυπώνει τη λέξη που θεωρεί ότι είναι πιο κοντά στη χειρονομία που κάνει. Το μοντέλο εκτυπώνει από τη δεξιά πλευρά κάθε νέα πρόβλεψη που κάνει και κρατάει την προηγούμενη στη μνήμη με μέγιστο αριθμό 5 λέξεων. σε κάθε κίνηση του χρήστη το μοντέλο θα κάνει μια πρόβλεψη σύμφωνα με την ακολουθία των κινήσεων όπως φαίνεται στην παρακάτω εικόνα.



Εικόνα 4: Δείγμα εικόνας από την δοκιμή του μοντέλου σε πραγματικό χρόνο

## V. ΣΥΜΠΕΡΑΣΜΑΤΑ

Από τα παραπάνω αποτελέσματα, παρατηρούμε ότι η εκπαίδευση του μοντέλου αναγνώρισης νοηματικής γλώσσας ήταν επιτυχής, με το μοντέλο να επιτυγχάνει άριστες επιδόσεις σε όλες τις μετρικές αξιολόγησης. Η άριστη ακρίβεια, ανάκληση, ακρίβεια (precision) και F1-score υποδεικνύουν ότι το μοντέλο μπορεί να αναγνωρίσει σωστά τις κινήσεις της νοηματικής γλώσσας που εκπαιδεύτηκε. Τα γραφήματα εκπαίδευσης δείχνουν μια συνεπή βελτίωση κατά την εκπαίδευση, υποδηλώνοντας ότι το μοντέλο μαθάνει αποτελεσματικά από τα δεδομένα εκπαίδευσης χωρίς να παρουσιάζει υπερπροσαρμογή. Η συνολική απόδοση του μοντέλου, όπως καταδεικνύεται από τα αποτελέσματα και τα γραφήματα, υποστηρίζει την αποτελεσματικότητα της χρήστης LSTM δικτύων για την αναγνώριση μοτίβων σε ακολουθίες και την ακριβή ταξινόμησή τους σε πραγματικό χρόνο.

## VI. ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] W. Sandler, D. Lillo-Martin, "Sign language and linguistic universals". Cambridge University Press, 2006
- [2] Hasim Sak, Andrew Senior, Francoise Beaufays, Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. Conference Article in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2014.
- [3] Felix Gers(Thesis Number 2366-2001 EPFL). Long Short-Term Memory in Recurrent Neural Networks.
- [4] Olah, C., (2015). Understanding LSTM Networks. [Blog] colah's blog, Available at: [Accessed 10 April 2021]
- [5] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.
- [6] "Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville, MIT Press, 2016.
- [7] Bart van Merriënboer, Dzmitry Bahdanau,Vincent Dumoulin,Dmitriy Serdyuk,David Warde-Farley,Jan Chorowski,Yoshua Bengio "Blocks and Fuel: Frameworks for deep learning"
- [8] Ammar Anuar, Khairul Muzzammil Saipullah, Nurul Atiqah Ismail, Yewguan Soo "OpenCV Based Real-Time Video Processing Using Android Smartphone"
- [9] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, Matthias Grundmann "Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs "
- [10] Wei Liu , Dragomir Anguelov , Dumitru Erhan , Christian Szegedy , Scott Reed , ChengYang Fu , Alexander C. Berg "SSD: Single Shot MultiBox Detector "
- [11] Jared Ostmeyer, Lindsay Cowell "Machine Learning on Sequential Data Using a Recurrent Weighted Average"
- [12] Kumar, P., Roy, P.P., Dogra, D.P.: Independent Bayesian classifier combination based sign language recognition using facial expression. Inf. Sci. (Ny). 428, 30–48 (2018).
- [13] Vasantha Kumar Velu, Sendhil Kumar Selvaraju. "Developing a Conceptual Framework for Short Text Categorization using Hybrid CNN- LSTM based Caledonian Crow Optimization" , Expert Systems with Applications, 2022
- [14] Jun Xie, Bo Chen, Xinglong Gu, Fengmei Liang, Xinying Xu. "Self-Attention-Based BiLSTM Model for Short Text Fine-Grained Sentiment Classification", IEEE Access, 2019
- [15] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634