Αναγνώριση δραστηριότητας μέσω ταξινόμησης Video

Georgios Batsis

July 4, 2022

1 Εισαγωγή

Στην σημερινή εποχή παρατηρείται εκθετική αύξηση της διαθεσιμότητας των δεδομένων βίντεο στο διαδίκτυο. Οι λόγοι για τους οποίους διαμορφώθηκε η συγκεκριμένη κατάσταση είναι, μεταξύ άλλων, η δημιουργία όλο ένα και περισσότερων Video Streaming/Sharing ψηφιακών εφαρμογών (e.g. Youtube, Vimeo) στις οποίες δημοσιεύονται καθημερινά βίντεο διάρκειας εκατοντάδων ωρών καθώς και αμεσότητα και ταχύτητα sharing που προσφέρουν στους χρήστες τα social media (e.g. Facebook, Instagram, Twitter) [1, 2]. Καθίσταται αντιληπτό ότι πλέον οι χρήστες είναι αναγκασμένοι να αναζητούν τα συγκεκριμένα δεδομένα μέσα σε τεράστιες συλλογές που διατίθενται από τις ψηφιακές πλατφόρμες. Λύση στο συγκεκριμένο πρόβλημα πιστεύουν οι ερευνητές και οι μηχανικοί ότι δίνεται μέσω της αυτοματοποιημένης ανάλυσης και κατηγοριοποίησης βίντεο.

Η ταξινόμηση βίντεο βάσει μοτίβων και γεγονότων που παρατηρούνται αποτελεί πρόκληση καθώς απαιτεί αποτελεσματική ανάλυση του περιεχομένου. Ένα βίντεο αποτελείται από μια ακολουθία διαδοχικών εικόνων - frames που εναλλάσσονται με έναν συγκεκριμένο ρυθμό ο οποίος είναι ανάλογος την διάρκειάς τους, εν προκειμένω τον αριθμό των frames ανά δευτερόλεπτο (Frames per second-FPS). Συνεπώς, η ανάλυση του περιεχομένου προϋποθέτει αρχικά την εξαγωγή πληροφορίας από κάθε μεμονωμένο frame με μεθόδους που εφαρμόζονται σε εφαρμογές ανάλυσης εικόνας και εν συνεχεία την εύρεση χρονικών συσχετίσεων μεταξύ των διαδοχικών. Απαιτείται, επομένως, όχι μόνο η αναγνώριση γεγονότων αλλά και τον τρόπο με τον οποίο αυτά εξελίσσονται χρονικά. Ο συνδυασμός της χωρικής και χρονικής πληροφορίας οδηγεί στην αναγνώριση συμβάντων όπως για παράδειγμα η αναγνώριση ανθρώπινης δραστηριότητας ή η αναγνώριση της παραβατικότητας [3].

Η αυτοματοποίηση της διαδικασίας ανάλυσης βίντεο εξελίσσεται συγχρόνως με τις τεχνολογίες και μεθόδους που αφορούν την Τεχνητή Νοημοσύνη και την Μηχανική Μάθηση. Η ανάλυση κάθε μεμονωμένου frame απαιτεί την εφαρμογή μεθόδων ψηφιακής επεξεργασίας εικόνας με στόχο την εξαγωγή χαρακτηριστικών. Παραδοσιακά στην αναγνώριση εικόνας μέθοδοι όπως η SIFT, HOGs ή οι μέθοδοι χαρακτηριστικών υφής (Texture) μπορούν να χρησιμοποιηθούν στο πεδίο των βίντεο. Η σύλληψη, όμως, της χρονικής πληροφορίας είναι ένα βασικό

στοιχείο της μεθοδολογίας, αν χρησιμοποιηθούν κάποια από τις προαναφερθείς τεχνικές, απαιτεί την εφαρμογή μεθόδων όπως η στατιστική συγκέντρωση (Aggregation) ή ακόμη και την πιο σύνθετη μέθοδο των Dense Trajectories [4, 5]. Η εκτεταμένη έρευνα πάνω στις μεθόδους Βαθιάς Μηχανικής Μάθησης (Deep Learning - DL) οδήγησε στην αύξηση της απόδοσης σε διεργασίες ανάλυσης βίντεο συγκριτικά με τα handcrafted χαρακτηριστικά [6]. Τα Βαθιά Νευρωνικά Δίκτυα (Deep Neural Networks - DNN) δεν είναι ικανά μόνο να εξάγουν την καταγεγραμμένη στα βίντεο χωρική πληροφορία με μεγαλύτερη ακρίβεια, αλλά και με μικρότερο υπολογιστικό κόστος διότι μπορούν να προβλέψουν χαρακτηριστικά από πολλαπλά frames ταυτόχρονα λόγω της παράλληλης επεξεργασίας. Δεν θα ήταν υπερβολή να ισχυριστούμε ότι μοντέλα Συνελικτικών Νευρωνικών Δικτύων (Convolutional Neural Networks - CNN) έχουν προσεγγίσει τις ανθρώπινες προβλέψεις σε μεγάλους διαγωνισμούς αναγνώρισης εικόνας όπως είναι το ImageNet. Τέλος,, οι δυνατότητες των DNN εκτείνονται και στην εξαγωγή χρονικής πληροφορίας μέσω της ανάπτυξης ακολουθιακών μοντέλων όπως είναι τα Recurrent Neural Networks (RNN) τα οποία χρησιμοποιούνται σε περιπτώσεις που τα δεδομένα χαρακτηρίζονται από χρονική μεταβλητότητα.

Στην συγκεκριμένη εργασία, υλοποιείται μια υβριδική DL προσέγγιση και συγκεκριμένα ένα CNN ακολουθούμενο από RNN με στόχο την αναγνώριση ανθρώπινης δραστηριότητα σε βίντεο. Ένα βίντεο εκχωρείται στο σύστημα ως μια ακολουθία από εικόνες - frames και κάθε μία εισέρχεται σε ένα προ-εκπαιδευμένο CNN από το οποίο έχει αφαιρεθεί το τελευταίο Layer που αφορά τις κλάσεις του Dataset στο οποίο έχει εκπαιδευτεί εξ αρχής. Τα χαρακτηριστικά που προκύπτουν συνδυάζονται σε μια ενιαία ακολουθία βήματος ίσου με τον αριθμό των frames. Η συγκεκριμένη ακολουθία διέρχεται εντός ενός RNN και συγκεκριμένα ενός Long short-term memory (LSTM) υπεύθυνου για την εξαγωγή χρονικής πληροφορίας από τα βίντεο. Με αυτόν τον τρόπο δημιουργείται μια ενιαία χωρο-χρονική αναπαράσταση η οποία ταξινομείται στην αντίστοιχη κατηγορία μέσω ενός Fully Connected.

2 Μέθοδοι

2.1 Μεταφορά Μάθησης και προ-εκπαιδευμένα CNN

Η μέθοδος Tranfer Learning διάσημων αρχιτεκτονικών DNN, όπως για παράδειγμα προ-εκπαιδευμένα CNN σε μεγάλου εύρους Datasets σαν το ImageNet, έχει αποδειχτεί ότι αποτελεί αποτελεσματική προσέγγιση ενός προβλήματος για το οποίο είτε δεν υπάρχουν αρκετά δεδομένα ώστε η εκ του μηδενός εκπαίδευση να οδηγήσει σε ένα μοντέλο υψηλών επιδόσεων ή όταν είναι αναγκαία μια εναλλακτική και συγχρόνως αποδοτική μέθοδο εξαγωγής χαρακτηριστικών. Χαρακτηριστική ιδιότητα των μοντέλων που έχουν εκπαιδευτεί σε Datasets μεγάλου αριθμού κλάσεων είναι η υψηλή διακριτική ικανότητα η οποία συνεπάγεται την δυνατότητα εξαγωγής χρήσιμης πληροφορίας. Επομένως, καθίστανται ένα ισχυρό εργαλείο στην διαχείρισης δεδομένων μικρότερου μεγέθους ακόμη και αν αυτά ανήκουν σε διαφορετικό πεδίο εφαρμογών από το αντίστοιχο της αρχικής εκπαίδευσης [7].

Μια εκ των state-of-the-art τοπολογιών DNN είναι τα Residual Networks (ResNet),

μια οικογένεια CNN τα οποία ως προς την αρχιτεκτονική οργανωνονται σε Convolutional Blocks και είναι διαθέσιμα σε διάφορες μορφές ανάλογα με το βάθος τους. Βασικό δομικό στοιχείο των ResNet είναι οι μονάδες Residual που βασίζονται στην ιδέα της παράληψης (skipping) των blocks μέσω shortcut connections, δηλαδή διασύνδεση μεταξύ της εισόδου και της εξόδου ενός block. Το εσωτερικό ενός block αποτελείται από μια ακολουθία από Convolutional Layers, Activation Funcitons και Batch Normalization και προφανώς τις Residual συνδέσεις. Οι τελευταίες εμφανίζονται με δύο διαφορετικές παραλλαγές: απευθείας σύνδεση της εισόδου του block με την έξοδο ή σύνδεση αυτών μεταξύ ενός μόνο Convolutional Layer. Το συγκεκριμένο τέχνασμα επιτρέπει την κατασκευή πολύ βαθιών DNN τα οποία επιτρέπουν την αποδοτική επίλυση πολύπλοκων προβλημάτων με ταυτόχρονη βελτιστοποίηση των παραμέτρων ώστε να μην εμφανιστεί ο κίνδυνος vanishing gradients [8]. Στην Εικόνα 1 παρουσιάζεται ένα παράδειγμα της αρχιτεκτονικής ResNet.

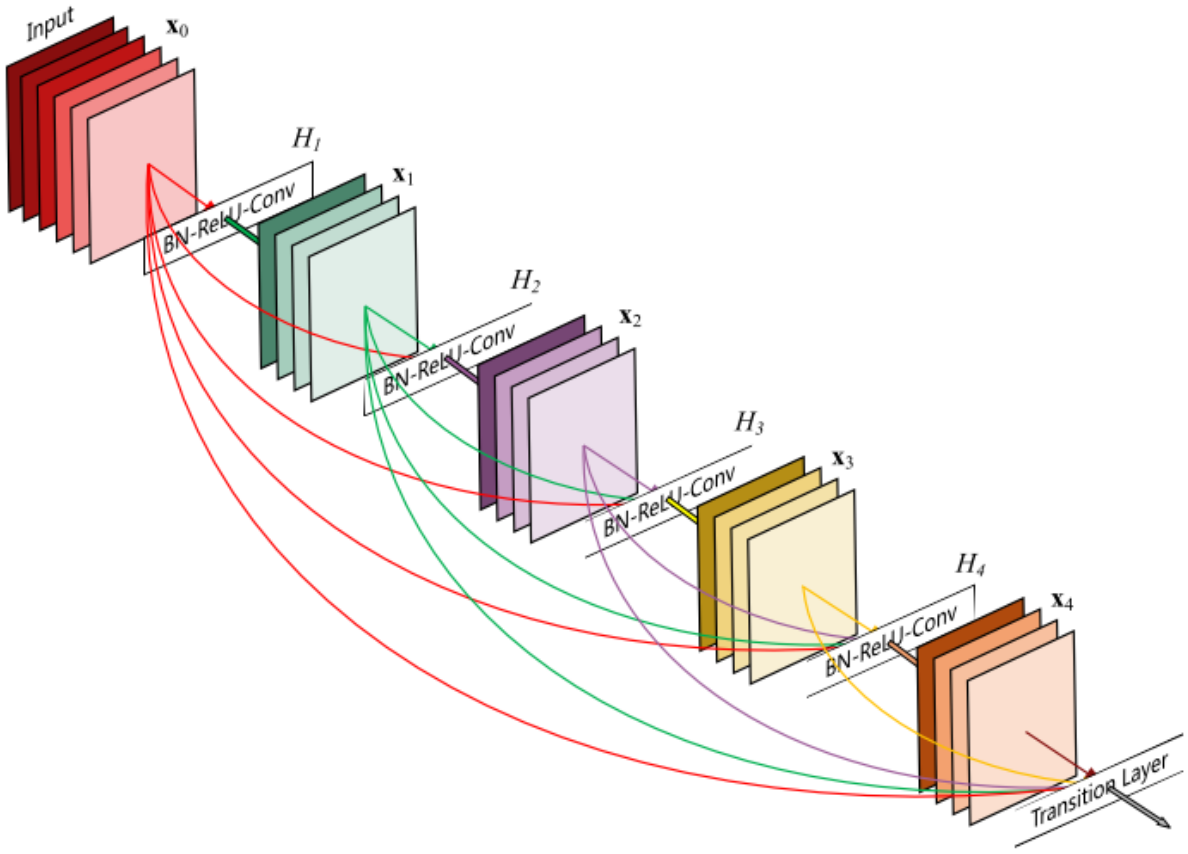

Figure 1: Παράδειγμα τοπολογίας ResNet.

2.2 Recurrent Neural Networks

Τα RNN αποτελούν μια αρχιτεκτονική DNN με βασικό χαρακτηριστικό τις ανατροφοδοτούμενες συνδέσεις μεταξύ των νευρώνων οι οποίες με την διαμορφώνουν την εσωτερική κατάσταση του δικτύου και άρα ένα είδος μνήμης. Γι αυτόν τον λόγο χρησιμοποιούνται για την επεξεργασία ακολουθιακών δεδομένων και την παραγωγή αποτελεσμάτων τα οποία μπορεί να είναι και αυτά μια νέα ακολουθία. Με αυτόν τον τρόπο, η λήψη αποφάσεων δεν βασίζεται αποκλειστικά στην τρέχουσα είσοδο, αλλά και στις προηγούμενες, αφού η έξοδος του δικτύου με βάση αυτές έχει ήδη ανατροφοδοτηθεί σε αυτό και άρα έχει ενημερωθεί η εσωτερική του κατάσταση.

Παρά το γεγονός ότι τα RNN αποτελούν μια ευέλικτη τοπολογία, εμφανίζουν πολλές φορές ζητήματα απόδοσης κατά την επεξεργασία μεγάλων ακολουθιών διότι ο κίνδυνος εμφάνισης των vanishing/exploiting gradients είναι αυξημένος. Επίλυση στο συγκεκριμένο πρόβλημα αποτελεί ένα είδος αρχιτεκτονικής RNN που αποτελούνται από πολλαπλές υπολογιστικές μονάδες υπεύθυνες για την διάδοση της πληροφορίας, τα Long Short Term Memory (LSTM). Οι υπολογιστικές μονάδες - κελιά κατέχουν τον ρόλο της μνήμης του δικτύου και όχι μόνο ελέγχουν την ροή της πληροφορίας αλλά επιλέγουν και ποια χρονικά βήματα μιας ακολουθίας είναι σημαντικά. Όσον αφορά την τελευταία ιδιότητα, η ακολουθία διέρχεται από το Forget Gate και αποφασίζεται το ποια ιστορικά βήματα αυτής είναι τα σημαντικά και ποια όχι. Ύστερα, επιλέγεται στο Input Gate το μέρος της νέας πληροφορίας που είναι σημαντικό και τέλος στο Output Gate αποφασίζεται πιο μέρος της πληροφορίας θα περάσει στην έξοδο της μονάδας. Η δομή ενός LSTM κελιού και οι αντίστοιχες μαθηματικές σχέσεις βρίσκονται στην Εικόνα 2.

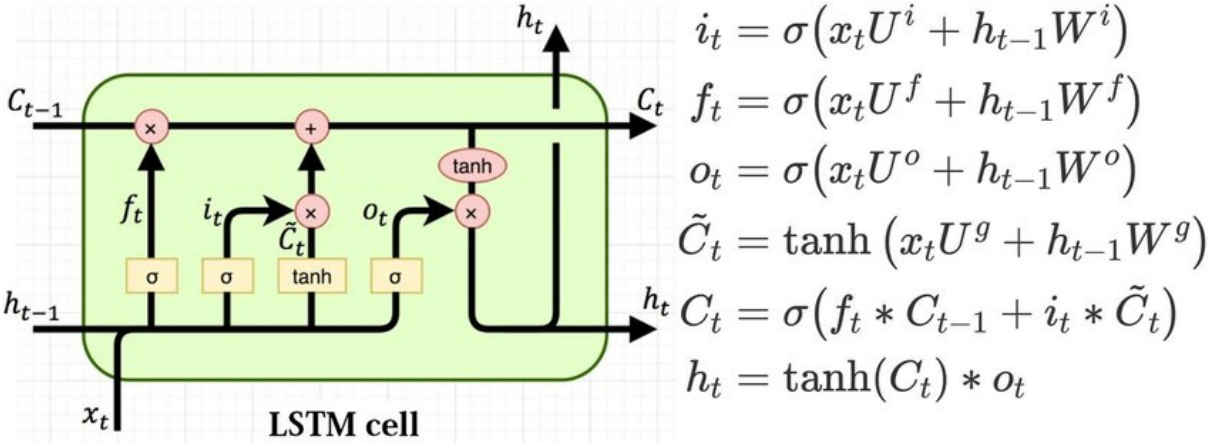

Figure 2: Δομή κελιού LSTM και οι αντίστοιχες μαθηματικές σχέσεις.

3 Προτεινόμενη Μέθοδος

3.1 Baseline Περίπτωση

Πρώτο βήμα της μεθοδολογίας αποτελεί η εξέταση της απόδοσης ενός ταξινομητή ο οποίος δεν αποτελεί μοντέλο DL. Η συγκεκριμένη προσέγγιση, όμως, απαιτεί την εξαγωγή χαρακτηριστικών σύμφωνα με τις κλασσικές μεθόδους ψηφιακής επεξεργασίας εικόνας. Στην συγκεκριμένη εργασία επιλέχθηκαν τα χαρακτηριστικά υφής (Texture), τα οποία υπολογίζονται με την βοήθεια του Gray Level Co-occurrence Matrix. Στόχος της μεθόδου είναι η μοντελοποίηση πληροφορίας σχετικά με την σχέση της φωτεινότητας ενός κεντρικού Pixel με την γειτονική του περιοχή η οποία ορίζεται μέσω ενός καθορισμένου παραθύρου. Η σχέση αποτυπώνεται με την μορφή πίνακα, κάθε στοιχεία του οποίου αποτελεί την σχετική συχνότητα με την οποία δύο pixel συγκεκριμένης τιμής εμφανίζονται σε μεταξύ μιας περιοχής και σε καθορισμένη κατεύθυνση. Τα Texture Features είναι οι στατιστικές μεταβλητές που εξάγονται από κάθε πίνακα:

1. Dissimilarity
2. Correlation
3. Contrast $= \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (i-j)^2$
4. Homogeneity $= \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \frac{P(i,j)}{1+(i-j)^2}$
5. Angular second Moment (ASM) $\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P(i,j)^2$
6. Energy $\sqrt{\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P(i,j)^2}$

Η εξαγωγή χαρακτηριστικών εφαρμόζεται για κάθε βίντεο σε επίπεδο frame και ύστερα πραγματοποιείται Aggregation ώστε τα χαρακτηριστικά κάθε frame να τροποποιηθούν σε μια ενιαία αναπαράσταση η οποία περιέχει χωρο-χρονική πληροφορία. Η συγκεκριμένη μέθοδος υλοποιείται μέσω του υπολογισμού στατιστικών στοιχείων, εν προκειμένω μέση τιμή και τυπική απόκλιση. Επομένως, για κάθε βίντεο έχουμε στην διάθεσή μας ένα διάνυσμα χαρακτηριστικών μεγέθους 12. Σε αυτό το σημείο, εκπαιδεύεται το baseline μοντέλο με την βοήθεια των παραπάνω χαρακτηριστικών. Πρόκειται για τον αλγόριθμο Support Vector Machines (SVM) με Radial basis kernel. Απαραίτητη προϋπόθεση είναι η κανονικοποίηση των χαρακτηριστικών με βάση τα δεδομένα εκπαίδευσης και εφαρμογή στο τέλος στα δεδομένα επαλήθευσης.

3.2 Συνδυασμός CNN και RNN για κατηγοριοποίηση Video

Στην παρούσα εργασία υλοποιείται ένα DNN με στόχο την ταξινόμηση ανθρώπινων δραστηριοτήτων σε βίντεο. Η βασική μεθοδολογία που ακολουθείται για τον ορισμό της αρχιτεκτονικής του δικτύου βρίσκεται στην Εικόνα 3. Το πρώτο βασικό δομικό στοιχείο του DNN είναι ένα προ-εκπαιδευμένο CNN από το οποίο έχει αφαιρεθεί το τελευταίο Layer που σχετίζεται με την ταξινόμηση στις κατηγορίες του Dataset στο οποίο έχει εκπαιδευτεί εξ αρχής. Προφανώς, από το προ-εκπαιδευμένο μοντέλο μπορεί να αφαιρεθεί και μεγαλύτερος αριθμός Layers που βρίσκονται βαθύτερα, όπως επίσης επιλέγεται και αν το μοντέλο

θα εκπαιδευτεί περαιτέρω ή αν θα παραμένει αμετάβλητο κατά την διαδικασία εκπαίδευσης (Freezed). Σε κάθε περίπτωση, το προ-εκπαιδευμένο CNN αποτελεί το βασικό εργαλείο εξαγωγής οπτικής πληροφορίας από τα διαδοχικά frames ενός βίντεο. Επιπλέον, η δυνατότητα παράλληλης επεξεργασίας επιτρέπει την πρόβλεψη χαρακτηριστικών για πολλαπλά frames ταυτόχρονα χωρίς την υλοποίηση κάποιας επαναληπτικής διαδικασίας. Συνεπώς, παράγεται για κάθε βίντεο ένα διάνυσμα χαρακτηριστικών οι διαστάσεις του οποίου ισούται με τον αριθμό των frames επί τις διαστάσεις τελευταίου Layer του προ-εκπαιδευμένου μοντέλου.

Με αυτόν τον τρόπο έχουμε διαθέσιμο για κάθε βίντεο μια ακολουθία από χαρακτηριστικά τα οποία περιέχουν την απαραίτητη χωρική πληροφορία και άρα το οπτικό περιεχόμενο των frames. Όμως τα βίντεο είναι δεδομένα τα οποία χαρακτηρίζονται από χρονική μεταβλητότητα και απαιτείται η ανάλυση της χρονικής πληροφορίας, δηλαδή το πως το οπτικό περιεχόμενο "εξελίσσεται" στο χρόνο. Η εξαγωγή της συγκεκριμένης πληροφορίας πραγματοποιείται με την τροφοδότηση της ακολουθίας χαρακτηριστικών σε ένα RNN και συγκεκριμένα LSTM. Το τελευταίο προσφέρει μια ενιαία χωρο-χρονική αναπαράσταση χαρακτηριστικών των βίντεο η οποία διέρχεται στον τελικό ταξινομητή.

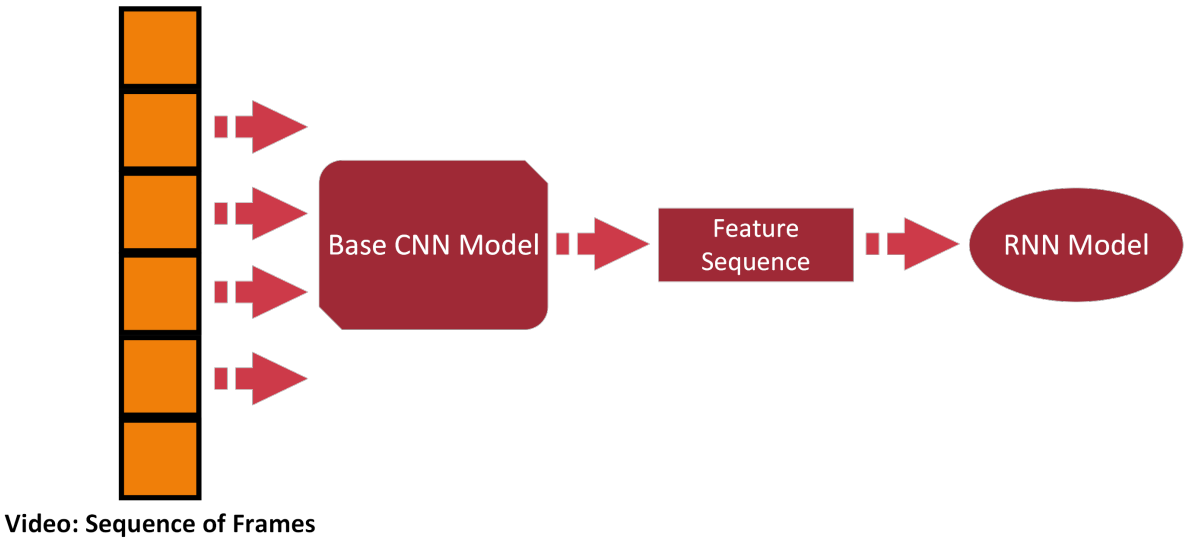

Figure 3: Μεθοδολογία ταξινόμησης βίντεο.

3.3 Επιλογή Αρχιτεκτονικής

Η επιλογή της τελικής αρχιτεκτονικής του μοντέλου πραγματοποιείται μέσω ενός συνόλου πειραμάτων, τα αποτελέσματα των οποίων παρουσιάζονται στην επόμενη ενότητα. Αρχικά, ως προ-εκπαιδευμένο CNN επιλέχθηκε το ResNet-152 το οποίο δεν μεταβάλλεται κατά την διαδικασία εκπαίδευσης. Η αρχιτεκτονική της πρώτης προσέγγισης (Approach A) παρουσιάζεται στην Εικόνα 4.

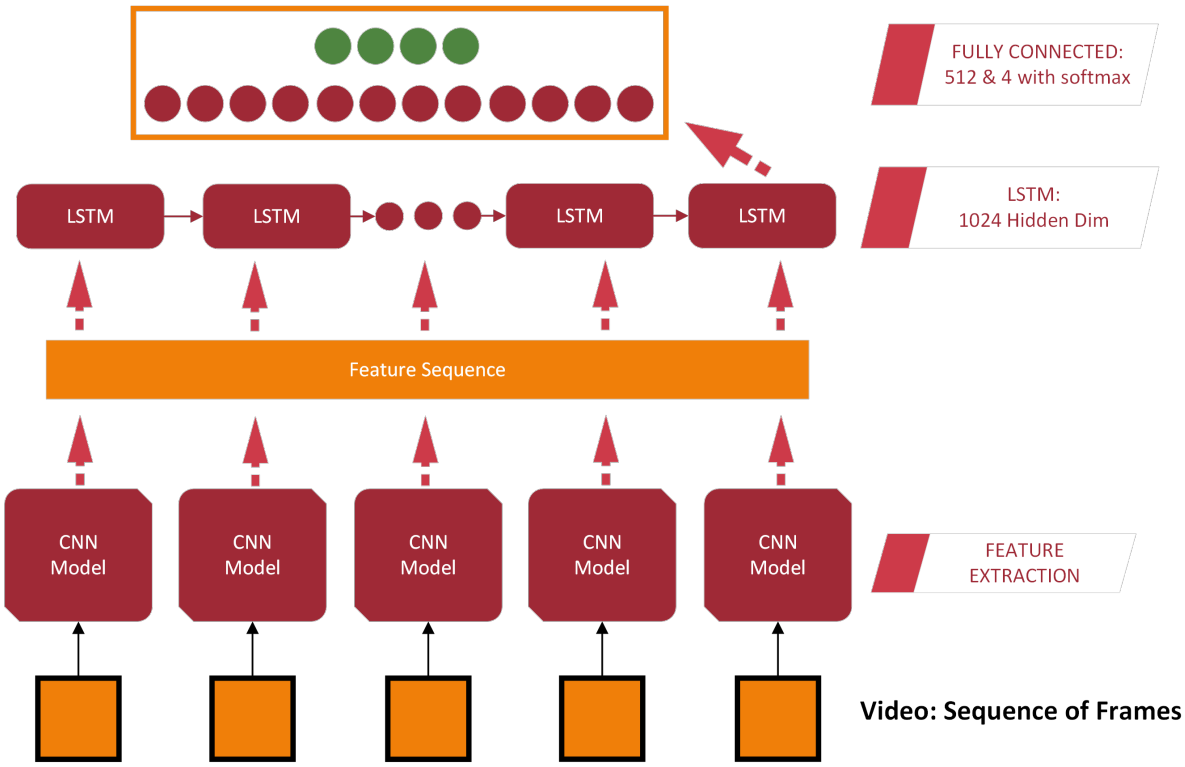

Figure 4: Αρχιτεκτονική Προσέγγισης A

Πρόκειται για την πιο απλή μορφή DNN η οποία δέχεται μια ακολουθία από frames, για κάθε ένα από αυτά εξάγονται τα 2048 χαρακτηριστικά με την βοήθεια του προ-εκπαιδευμένου μοντέλου και εισέρχονται στο στρώμα LSTM με διάσταση κρυφής κατάστασης 1024. Επειδή η διεργασία που αναπτύσσεται ανήκει στην κατηγορία των προβλημάτων ταξινόμησης, θα πρέπει η συνολική πληροφορία που έχει τροφοδοτηθεί το τελευταίο κελί του LSTM να δοθεί ως είσοδο στον τελικό ταξινομητή το οποίο είναι ένα στρώμα Fully Connected. Το τελευταίο αποτελείται από ένα κρυφό-ενδιάμεσο στρώμα 512 και το στρώμα ταξινόμησης αποτελείται από 4 Νευρώνες, όσες και οι κλάσεις προς πρόβλεψη.

Στην δεύτερη προσέγγιση (Approach B), η μετατροπή του Δικτύου αφορά την προσθήκη ενός Layer 1024 Νευρώνων αμέσως μετά το προ-εκπαιδευμένο μοντέλο. Όφελος από την συγκεκριμένη ρύθμιση, αποτελεί αφενός η μείωση της διάστασης των χαρακτηριστικών που εισέρχονται στο LSTM, αφετέρου το συγκεκριμένο στρώμα προσφέρει ένα διάνυσμα χαρακτηριστικών τροποποιημένο ως προς το Dataset το οποίο χρησιμοποιείται στην εκπαίδευση. Η δομή του δικτύου παρουσιάζεται στην Εικόνα 5.

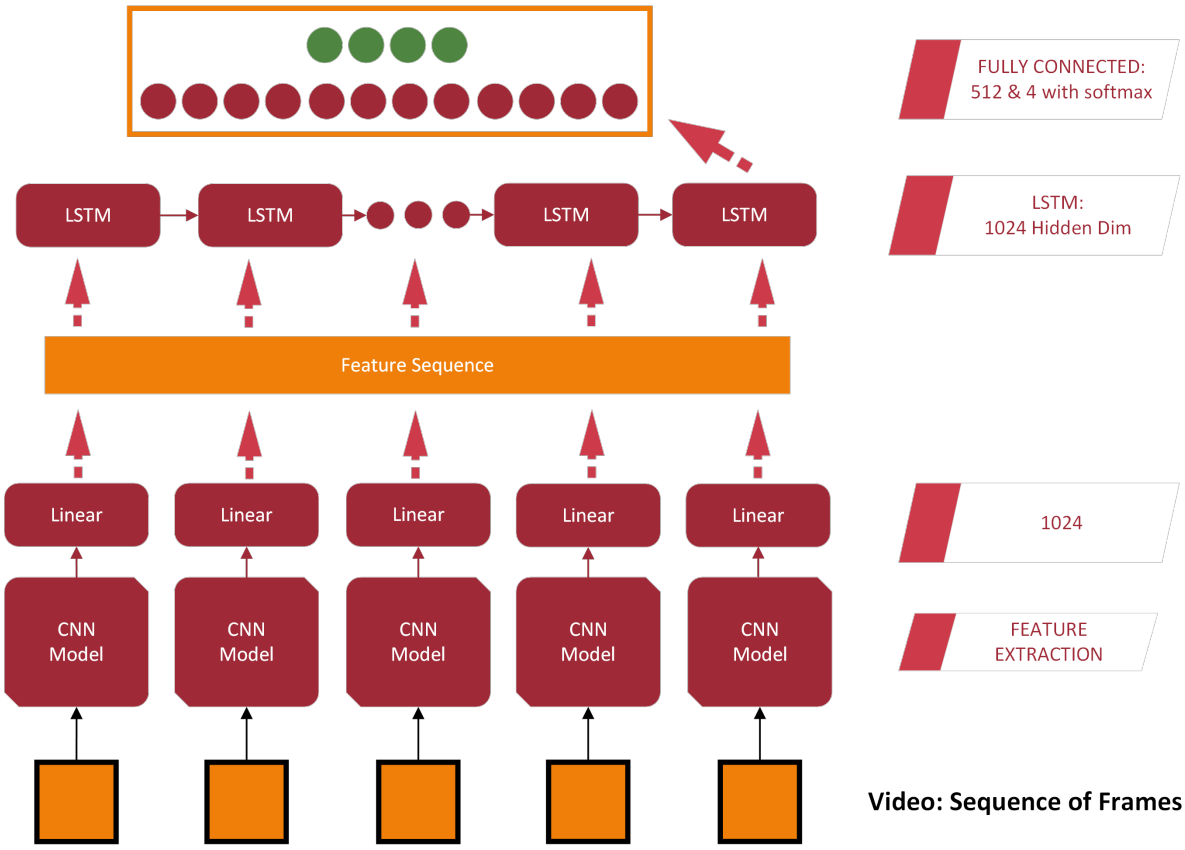

Figure 5: Αρχιτεκτονική Προσέγγισης B

Κατά την εκτέλεση πειραμάτων ώστε να βρεθεί η βέλτιστη αρχιτεκτονική ενός DL μοντέλου, μία συνηθισμένη πρακτική είναι αύξηση της πολυπλοκότητας του. Ο όρος πολυπλοκότητα του DNN αναφέρεται τόσο στο βάθος, δηλαδή των αριθμό των Layers, όσο και στο πλάτος κάθε Layer, δηλαδή από πόσους νευρώνες - μονάδες υπολογισμού αποτελείται. Η αύξηση της πολυπλοκότητας αποτελεί ένα σημαντικό βήμα κατά την ανάπτυξη της αρχιτεκτονικής διότι συμβάλλει στην αύξηση της απόδοσης του μοντέλου. Επομένως, σε αυτό το σημείο το μοντέλο της Εικόνας 5 μετατρέπεται σε μια πιο σύνθετη έκδοση με την προσθήκη ενός ακόμη επιπέδου μετά το προ-εκπαιδευμένο μοντέλο. Πλέον το ResNet ακολουθείται από ένα Fully Connected Layer 1024 - 512 διαστάσεων ανά επίπεδο. Πρόσθετα, αυξάνεται και ο αριθμός των LSTM Layer σε δύο καθώς και το Fully Connected Layer που εκτελεί την ταξινόμηση. Το τελευταίο πλέον αποτελείται από τρία επίπεδα διαστάσεων 1024 - 512 - 4 (Softmax). Η αρχιτεκτονική της τρίτης προσέγγισης παρουσιάζεται στην Εικόνα 6. Το συγκεκριμένο μοντέλο τροποποιήθηκε και ως προς την κατεύθυνση των LSTM Layers. Όπως φαίνεται στο σχήμα της Εικόνας 7, τα LSTM μετατράπηκαν σε δι-κατευθυντήρια με σκοπό την εξαγωγή πληροφορίας και από τις δύο ροές των διαδοχικών frames. Απαιτείται, όμως, η ανατροφοδότηση του Layer ταξινόμησης να πραγματοποιηθεί τόσο από το τελευταίο όσο

και από το πρώτο κελί του LSTM ώστε το διάνυσμα της κρυφής κατάστασης να αντιστοιχεί σε κάθε ροή. Αυτό συνεπάγεται τον διπλασιασμό του πλάτους του πρώτου στρώματος του Fully Connected.

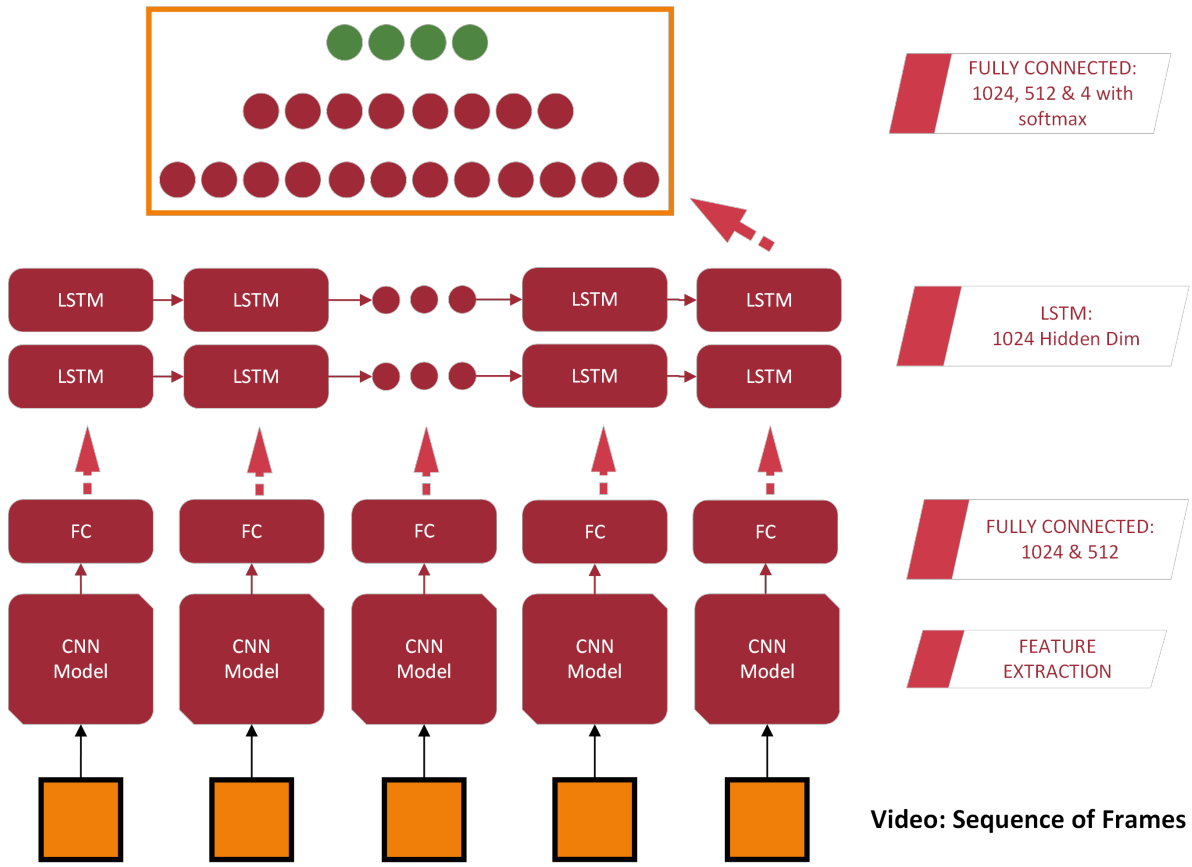

Figure 6: Αρχιτεκτονική Προσέγγισης C

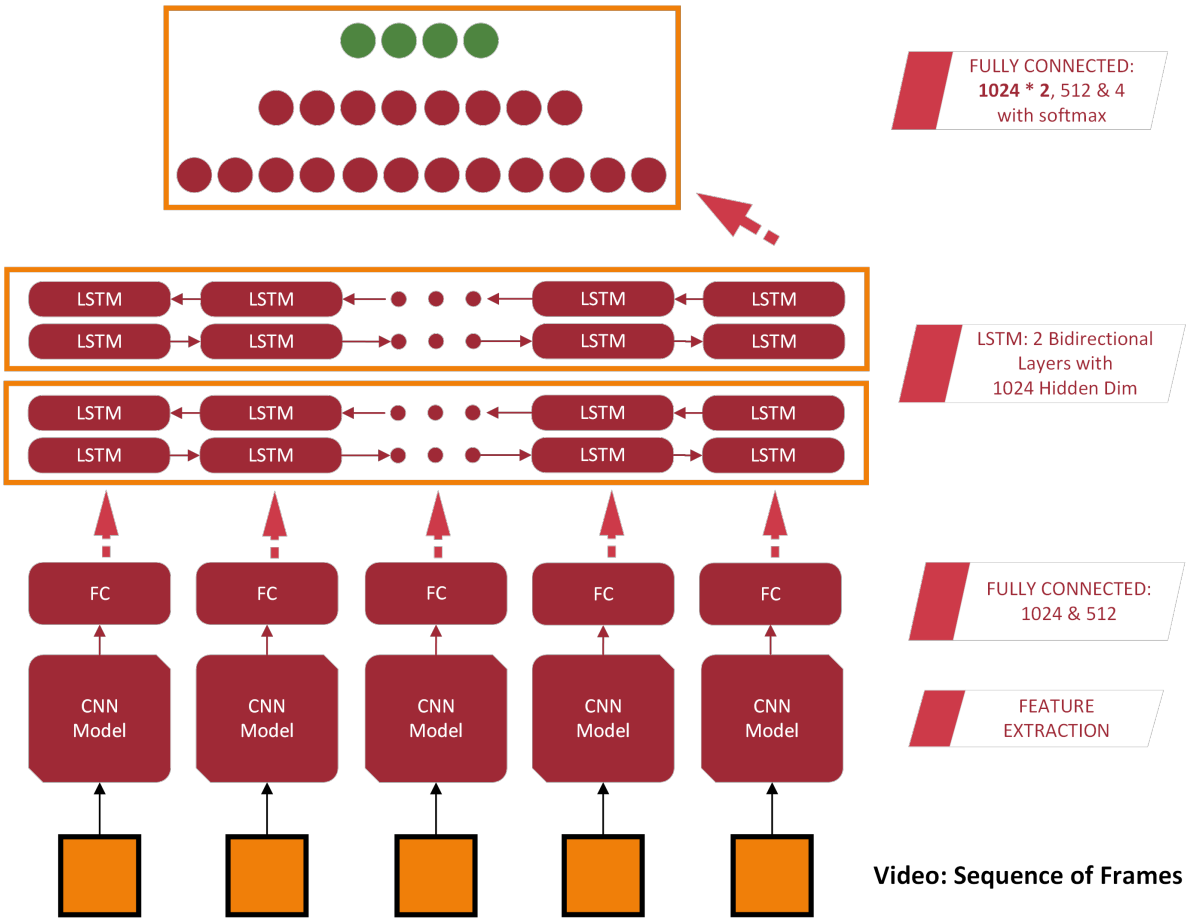

Figure 7: Αρχιτεκτονική Προσέγγισης C'

Το τελευταίο μοντέλο που αναπτύχθηκε είναι το πιο πολύπλοκο σε επίπεδο βάθους και προφανώς σε επίπεδο παραμέτρων εκπαίδευσης. Συγκεκριμένα, αποτελείται από ένα Fully Connected τριών στρωμάτων μετά το προ-εκπαιδευμένο μοντέλο, εν προκειμένω 1024-512-256. Η αναπαράσταση οπτικών χαρακτηριστικών από τα frames του βίντεο τροφοδοτείται σε τρία επίπεδα δι-κατευθυντήριων LSTM με 1024 κρυφή διάσταση. Σε αυτήν την περίπτωση, όμως, δεν μεταβάλλεται περαιτέρω το Fully Connected της ταξινόμησης. Η αρχιτεκτονική παρουσιάζεται στην Εικόνα 8.

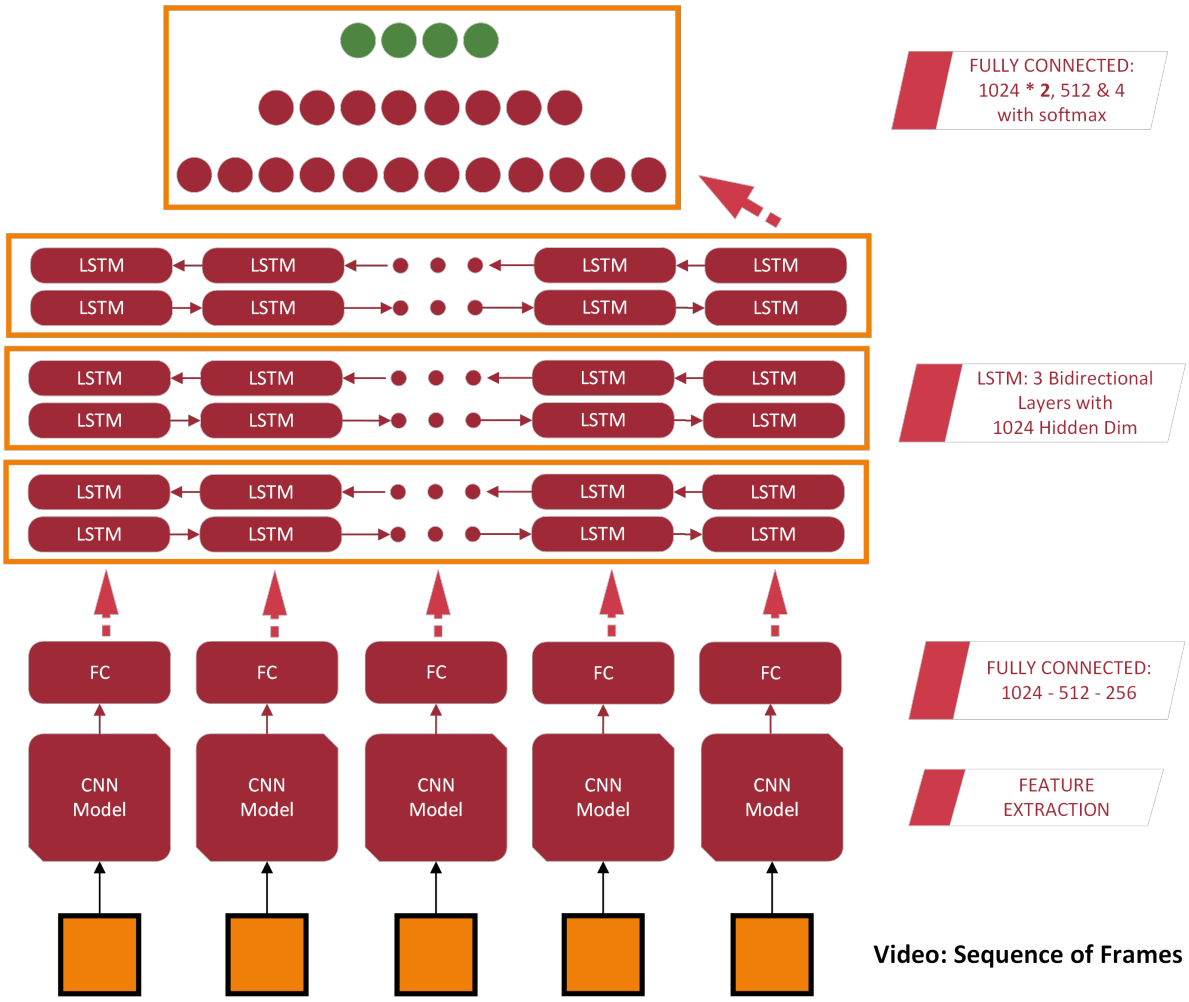

Figure 8: Αρχιτεκτονική Προσέγγισης D

4 Αποτελέσματα

4.1 Σύνολο Δεδομένων και προ-επεξεργασία

Το σύνολο δεδομένων πάνω στο οποίο τα μοντέλα εκπαιδεύτηκαν είναι το UCF 101 [9], το οποίο αποτελεί μια από τις μεγαλύτερες συλλογές δεδομένων βίντεο που απεικονίζονται ανθρώπινες δραστηριότητες. Τα διαθέσιμα βίντεο ξεπερνούν τα 13.000, η συνολική διάρκεια είναι στις 27 ώρες και αντιστοιχούν σε 101 κλάσεις. Τα βίντεο έχουν συλλεχθεί από την ψηφιακή πλατφόρμα Youtube και έχουν συγκεκριμένο αριθμό frames ανά δευτερόλεπτο, εν προκειμένω 25 FPS. Οι δημιουργεί ορίζουν και baseline μοντέλο με απόδοση 44.5 %. Στην συγκεκριμένη εργασία χρησιμοποιήθηκε ένα μικρό μέρος του UCF 101 και συγκεκριμένα 4

κατηγορίες δραστηριοτήτων:

0: Playing Guitar
1: Rock Climbing Indoor
2: Soccer Juggling
3: Band Marching

Κατά την ανάπτυξη των DL μοντέλων πραγματοποιήθηκε διαχωρισμός σε δεδομένα εκπαίδευσης και ελέγχου (Train - Test) βάσει των επίσημων οδηγιών που είναι διαθέσιμα σε μορφή txt αρχείων, ενώ ένα μικρό μέρος του Train έμεινε εκτός για λόγους επαλήθευσης (Validation). Επιπλέον, επειδή τα βίντεο δεν έχουν κοινό αριθμό frames, εξάχθηκαν από κάθε ένα μικρότερα segments των 20 frames όπως φαίνεται στην Εικόνα 9. Τέλος, σε επίπεδο frame εφαρμόστηκε αλλαγή πλάτους και ύψους σε 224 x 224 και κανονικοποίηση των τιμών των pixels.

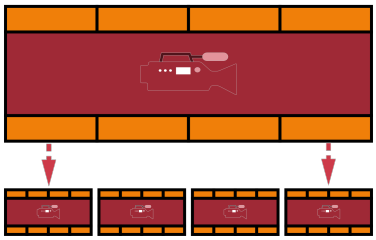

Figure 9: Εξαγωγή Segments από το αρχικό βίντεο.

4.2 Συνάρτηση απωλειών και Μετρικές απόδοσης

Η βελτιστοποίηση του μοντέλου βασίστηκε στην συνάρτηση απωλειών (Loss Function) Cross - Entropy, η οποία δέχεται την πραγματική κλάση ενός δείγματος και την πιθανότητα που προβλέφθηκε για κάθε κλάση:

$$CrossEntropoyLoss = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

Η απόδοση του μοντέλου μετράται με την βοήθεια των παρακάτω μετρικών οι οποίες βασίζονται στις ορθές και λανθασμένες προβλέψεις:

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

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

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

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

4.3 Εκπαίδευση μοντέλων και έλεγχος ορθότητας

Η ανάπτυξη των μοντέλων συμπεριλαμβανομένης της εκπαίδευσης και του ελέγχου ορθότητας πραγματοποιήθηκαν σε Python με την χρήση της βιβλιοθήκης scikit-learn και του PyTorch Framework. Η παρακολούθηση των μετρικών απόδοσης όσο διαρκούσαν οι εποχές έγινε με την βοήθεια του Tensorboard. Τα DL μοντέλα εκπαιδεύτηκαν θέτοντας ως μέγιστο αριθμό εποχών τις 100 αλλά με συνθήκη πρόωρης διακοπής εκπαίδευσης (Early Stopping) σε περίπτωση που το Loss στο Validation μέρος των δεδομένων δεν βελτιωνόταν περαιτέρω για 5 συνεχόμενες εποχές (Patience). Ως optimizer χρησιμοποιήθηκε ο Adam ενώ το μέγεθος κάθε batch ορίστηκε στα 8. Τα πειράματα έλαβαν χώρα σε τοπικό υπολογιστικό σύστημα με GPU NVIDIA RTX 2060 και στην υπηρεσία Cloud Computing Kaggle kernels - notebooks με την υποστήριξη GPU.

4.3.1 Baseline Περίπτωση

Στην περίπτωση εξαγωγής handcrafted χαρακτηριστικών και εν συνεχεία ταξινόμησης με SVM, τα αποτελέσματα που προκύπτουν δεν είναι τα επιθυμητά. Συγκεκριμένα, οι σωστές προβλέψεις αντιστοιχούν σε ποσοστό 61% του Validation μέρους των δεδομένων. Στην Εικόνα 10 εμφανίζονται όλες οι μετρικές εγκυρότητας του μοντέλου. Οι τόσο χαμηλές επιδόσεις οφείλονται αφενός στην φύση των χαρακτηριστικά υφής, αφετέρου στην μέθοδο της στατιστικής χρονικής συγκέντρωσης (Aggregation). Συγκεκριμένα, είναι δύσκολο σε ένα πολύπλοκο πρόβλημα ταξινόμησης οι δύο αυτές μέθοδοι να αποτυπώσουν την οπτική πληροφορία και χρονική συσχέτιση μεταξύ των frames.

	precision	recall	f1-score	support
0	0.41	0.32	0.36	168
1	0.64	0.70	0.67	92
2	0.55	0.67	0.60	202
3	0.97	0.84	0.90	134
accuracy			0.61	596
macro avg	0.64	0.63	0.63	596
weighted avg	0.62	0.61	0.61	596

Figure 10: Αποτελέσματα Baseline περίπτωσης.

4.3.2 Βαθιά Μοντέλα

Πρώτο βήμα για την επιλογή της αρχιτεκτονικής, αποτέλεσε η εύρεση του κατάλληλου προ-εκπαιδευμένου CNN μοντέλου που εκτελεί εξαγωγή χαρακτηριστικών από τα frames. Δοκιμάστηκαν το ResNet - 152 freezed στο τελευταίο Convolutional στρώμα και στην πραγματικότητα λήφθηκαν τα 2048 χαρακτηριστικά που εξάγονται από το Average Pooling μετά από αυτό, το VGG - 19 freezed στο προτελευταίο Layer στου ταξινομητή και το MobileNet.

Το τελευταίο δεν ειναι αρκετό ώστε το δίκτυο να ξεχωρίσει τις κλάσεις. Συνεπώς, μεταξύ του ResNet και του VGG επιλέχθηκε το πρώτο αφού στο δίκτυο της Approach A επιτυγχάνει απόδοση 87.8% έναντι 83.4% F1 στο Validation μέρος των δεδομένων.

Η πρώτη περίπτωση, όμως, δεν αποτελεί ένα αποτελεσματικό Νευρωνικό δίκτυο καθώς το Loss Function στο Validation δεν καταφέρνει να μειωθεί αισθητά σε σχέση με το αντίστοιχο στο Training μέρος. Αυτός είναι και ο λόγος που επιλέχθηκε η ανάπτυξη πιο βαθέων μοντέλων στα οποία προστίθενται και εκπαιδευόμενα στρώματα μετά το pre-trained CNN. Αναλυτικά οι επιδόσεις των μοντέλων βρίσκονται στον παρακάτω πίνακα:

Model Name	F1 Score	Loss
Approach A	87.8	2.9
Approach B	97.7	0.64
Approach C	98.4	0.44
Approach C (Bid)	97.7	0.57
Approach D	98.6	0.29

Επιπλέον, στην Εικόνα 11 φαίνεται ο τρόπος με τον οποίο διαμορφώνονται τα Loss - F1 στο Train και Validation μέρος κατά το πέρασμα των εποχών. Τα μοντέλα τα οποία επιτυγχάνουν πολύ καλές αποδόσεις με ταυτόχρονη σύγκλιση των μεγεθών απόδοσης για το Train και Validation μέρος κατά το Early Stopping, είναι τα C και D.

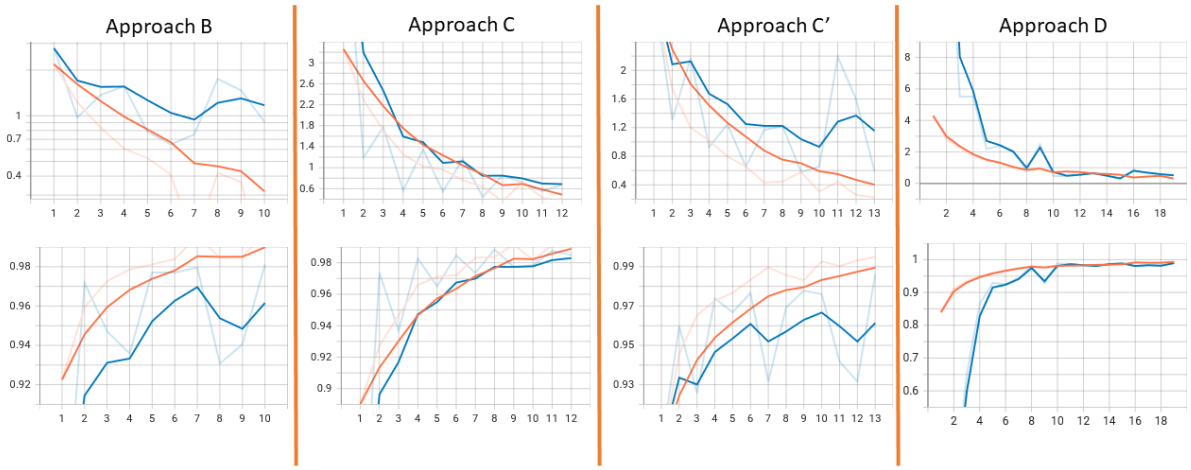

Figure 11: Αποτελέσματα Εκπαίδευσης DNN όλων των προσεγγίσεων.

Αν, όμως, στόχος είναι η επιλογή ενός μοντέλου για μελλοντικές εφαρμογές σίγουρα θα έπρεπε να είναι το C διότι επιτυγχάνει συγκρίσιμα αποτελέσματα με το D που χαρακτηρίζεται από μεγάλη πολυπλοκότητα. Γι αυτόν τον λόγο το μοντέλο Approach C χρησιμοποιήθηκε για τις προβλέψεις στο Test μέρος των δεδομένων τόσο σε επίπεδο Segments όσο και σε επίπεδο βίντεο. Οι τελευταίες πραγματοποιήθηκαν συγκεντρώνοντας για κάθε βίντεο τα posteriors που προβλέπει το μοντέλο για τα Segment αυτού και εφαρμόζοντας το Θεώρημα Ολικής Πιθανότητας. Το μοντέλο επιτυγχάνει 98% στα Segments και 99% στα Videos. Τα τελικά αποτελέσματα και οι Confusion Matrices φαίνονται στην Εικόνα 12.

Segments

	precision	recall	f1-score	support
0	0.97	1.00	0.99	510
1	1.00	0.94	0.97	865
2	0.94	0.97	0.96	621
3	0.95	1.00	0.97	347
accuracy			0.97	2343
macro avg	0.97	0.98	0.97	2343
weighted avg	0.97	0.97	0.97	2343

True\Predicted	0	1	2	4
0	512	0	0	0
1	2	812	37	14
2	13	1	603	4
3	0	1	0	346

Video

	precision	recall	f1-score	support
0	1.00	1.00	1.00	43
1	1.00	0.95	0.97	41
2	0.95	0.97	0.96	39
3	0.98	1.00	0.99	43
accuracy			0.98	166
macro avg	0.98	0.98	0.98	166
weighted avg	0.98	0.98	0.98	166

True\Predicted	0	1	2	4
0	43	0	0	0
1	0	39	2	0
2	0	0	38	1
3	0	0	0	43

Figure 12: Αποτελέσματα προβλέψεων στο Test Dataset. Αριστερά φαίνονται οι μετρικές απόδοσης και δεξιά ο Confusion Matrix. Οι προβλέψεις εκτελούνται σε επίπεδο Segments και ολόκληρων βίντεο.

5 Συμπεράσματα

Στην παρούσα εργασία, προσεγγίστηκε η αναγνώριση ανθρώπινης δραστηριότητας σε βίντεο με την χρήση Βαθιάς Μάθησης. Συγκεκριμένα, πραγματοποιήθηκε ένα πλήθος πειραμάτων με αφετηρία την υλοποίηση ενός απλού ταξινομητή σε συνδυασμό με μεθόδους εξαγωγής χαρακτηριστικών από τον τομέα της Ψηφιακής Επεξεργασίας Εικόνας. Στην συνέχεια, αναπτύχθηκαν μοντέλα Βαθιάς Μάθησης στα οποία η εξαγωγή χαρακτηριστικών βασίζεται σε προ-εκπαιδευμένο μοντέλο το οποίο είχε πετύχει σημαντικά υψηλές αποδόσεις σε διαγωνισμούς Μηχανικής Όρασης, όπως το ImageNet. Τα πειράματα που αφορούν τα Βαθιά

Νευρωνικά Δίκτυα πραγματοποιήθηκαν με σκοπό της εύρεση της τελικής αρχιτεκτονικής που επιλύει το πρόβλημα ταξινόμησης βίντεο αποδοτικότερα. Ο έλεγχος των μοντέλων σχετίστηκε με την μεθοδολογία Μεταφοράς Μάθησης, τον τρόπο με τον οποίο το προ-εκπαιδευμένο CNN ενσωματώθηκε το τελικό δίκτυο και την εύρεση της ιδανικής διάταξης του ακολουθιακού μοντέλου το οποίο εξάγει την χρονική συσχέτιση μεταξύ των οπτικών χαρακτηριστικών. Επιπλέον, εξάχθηκαν συμπεράσματα σχετικά με την απόδοση ενός μοντέλου αν αυξηθεί η πολυπλοκότητά του. Το τελικό βαθύ μοντέλο επιτυγχάνει 98% F1 Score σε επίπεδο Segments και 99% σε επίπεδο Videos που ανήκουν στο Test μέρος των δεδομένων. Τέλος, όσον αφορά τις μελλοντικές επεκτάσεις της προτεινόμενης μεθόδου, ενδιαφέρουσα προσέγγιση θα ήταν η απόπειρα ανανέωσης των βαρών του προ-εκπαιδευμένου CNN κατά την διαδικασία εκπαίδευσης αντί της διατήρησης αυτών αμετάβλητων (Freezed). Επιπλέον, η βασική αρχιτεκτονική μπορεί να χρησιμοποιηθεί και στο σύνολο του UCF101 ή κάποιας εναλλακτικής προσθέτοντας τις ακόλουθες ρυθμίσεις: χρήση Multimodal δεδομένων (πχ Βίντεο και ήχος) και εισαγωγή του μηχανισμού Attention.

References

[1] *160 amazing YouTube statistics and facts: By the numbers*, 2022, available at https://expandedramblings.com/index.php/youtube-statistics/.

[2] B. N. Subudhi, D. K. Rout, and A. Ghosh, "Big data analytics for video surveillance," *Multimedia Tools and Applications*, vol. 78, no. 18, pp. 26 129–26 162, 2019.

[3] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[4] G. LoweDavid, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 2004.

[5] H. Wang and C. Schmid, "Action recognition with improved trajectories," 12 2013, pp. 3551–3558.

[6] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[7] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *CoRR*, vol. abs/1411.1792, 2014. [Online]. Available: http://arxiv.org/abs/1411.1792

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[9] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," 2012. [Online]. Available: https://arxiv.org/abs/1212.0402