Αναγνώριση δραστηριότητας μέσω ταξινόμησης Video

Georgios Batsis

June 26, 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.

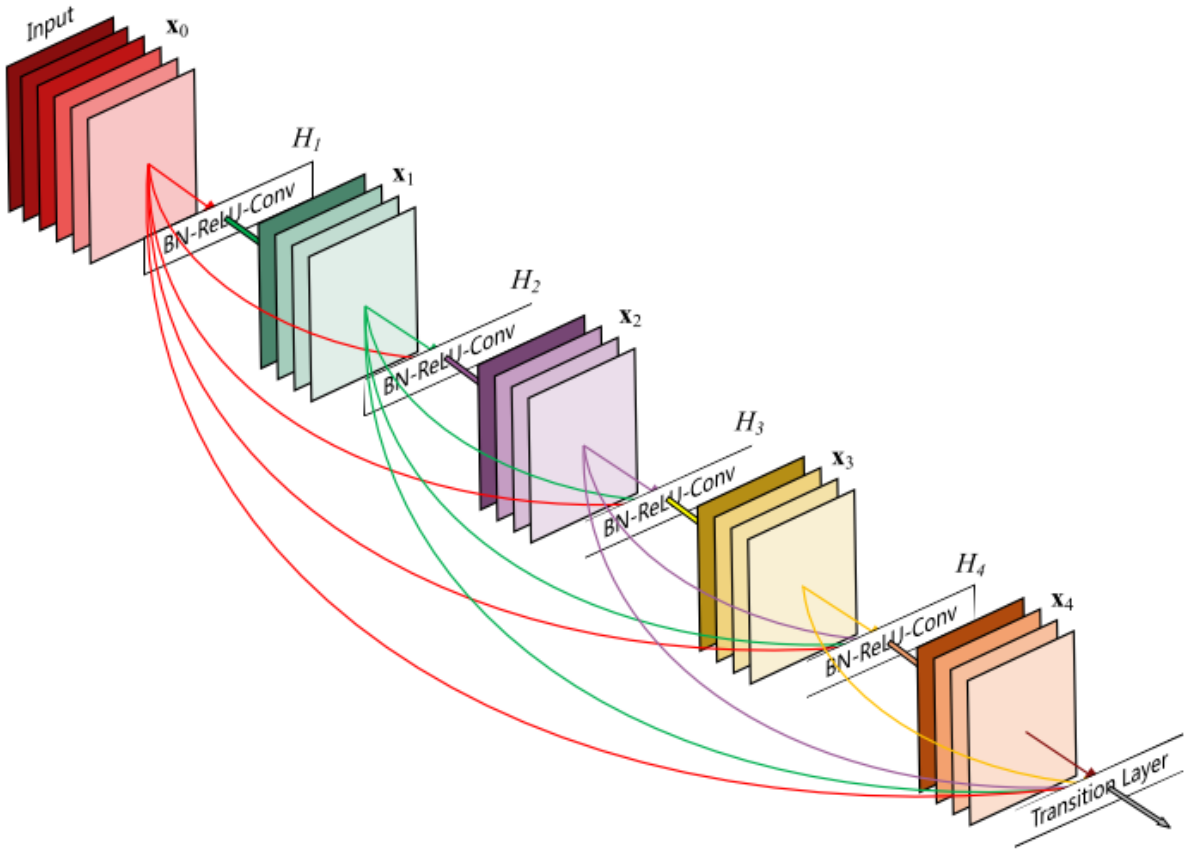

Figure 1: Παράδειγμα τοπολογίας ResNet.

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.

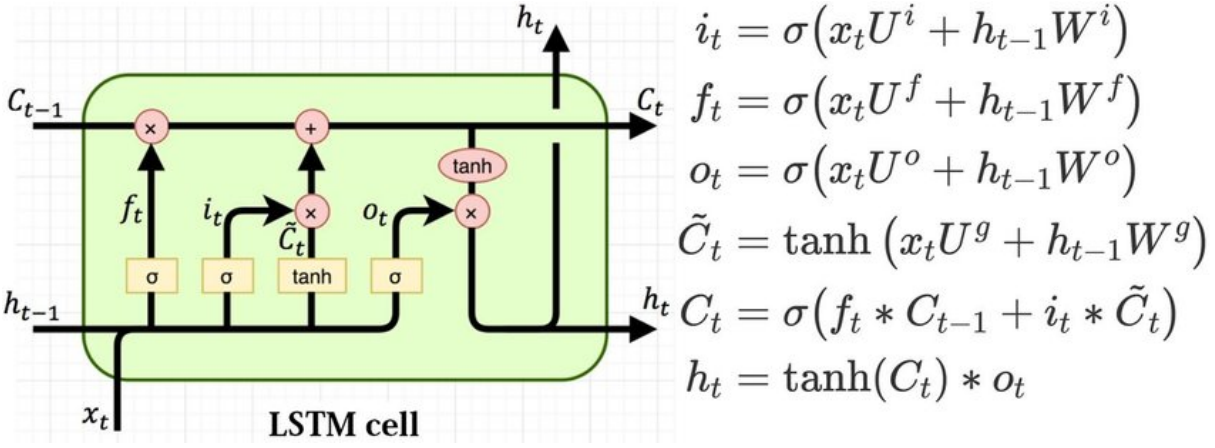

Figure 2: Δομή κελιού LSTM και οι αντίστοιχες μαθηματικές σχέσεις.

3 Προτεινόμενη Μέθοδος

3.1 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.

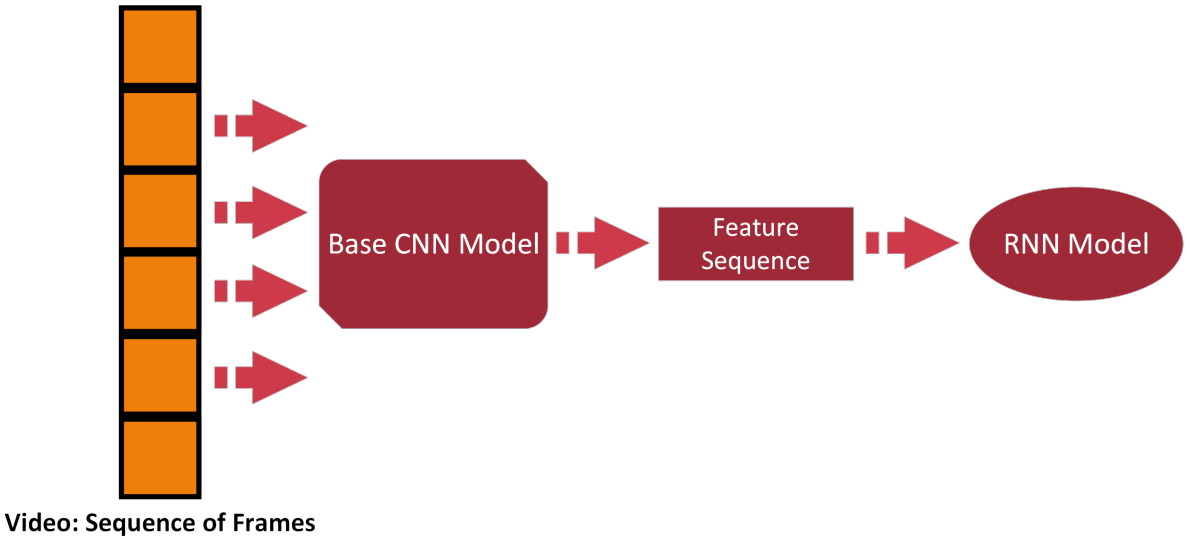

Figure 3: Μεθοδολογία ταξινόμησης βίντεο.

3.2 Επιλογή Αρχιτεκτονικής

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.