Μηχανική Μάθηση Σε Πολυμεσικά Δεδομένα

Εντοπισμός Highlights σε Ποδοσφαιρικούς Αγώνες

Κυριαζόπουλος Χρήστος, Σταυριανουδάκης Βασίλειος, Χιώτης Νικόλαος
MTN2207, MTN2215, MTN2221

Εισαγωγή

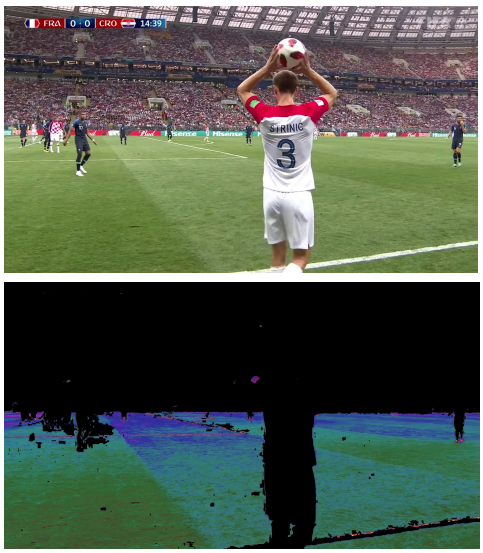
Σκοπός της παρούσας εργασίας αποτελεί ο εντοπισμός σημείων ενδιαφέροντος σε ποδοσφαιρικούς αγώνες. Το σετ δεδομένων το οποίο χρησιμοποιήθηκε, για την αξιολόγηση της επίδοσης των τεχνικής που αναπτύξαμε για την επίλυση του συγκεκριμένου προβλήματος, προέρχεται από την παρακάτω youtube playlist: https://www.youtube.com/playlist?list=PLCGIzmTE4d0iY7ryPjeltaZMJ0s_jJPYN . Αρχικά κάναμε την παραδοχή ότι κάθε φάση σε οποιονδήποτε αγώνα διαρκεί 10 δευτερόλεπτα. Έτσι, κάθε αγώνας τμηματοποιήθηκε σε segments των 10s και στην συνέχεια εξήχθησαν features που αφορούν το κάθε modality για καθένα από αυτά. Ακολουθώντας τεχνικές τύπου outlier detection(καθώς θεωρούμε πιο σπάνιο ένα κλιπ να αντιστοιχεί σε φάση) με υπολογισμό αποστάσεων στο χώρο των features για κάθε modality ξεχωριστά λάβαμε 3 ιεραρχίες(μία για κάθε modality) για τα segments. Στην τελική απόφαση των σημείων ενδιαφέροντος, χρησιμοποιήσαμε τον μέσο όρο των ιεραρχιών αυτών, ενώ παρατηρήσαμε ότι σε κάποιες περιπτώσεις ένας βεβαρημένος μέσος όρος έδινε (ποιοτικά) καλύτερα αποτελέσματα. Η λύση που προτείνουμε λοιπόν χρησιμοποιεί καθαρά unsupervised τεχνική και δεν εμπεριέχει καμία διαδικασία εκμάθησης.

1. **Εικόνα**

Η μέθοδος που ακολουθήσαμε για την περίπτωση της διάστασης της εικόνας είναι η εξής:

- Λήψη στιγμιότυπων εικόνων ανά 10 δεύτερα από το αρχείο βίντεο του αγώνα(ξεκινώντας από το 5ο δευτερόλεπτο).
- Εφόσον μετατρέψουμε σε κατάλληλη μορφή τις εικόνες χρησιμοποιούμε ένα pretrained μοντέλο (resnet18) με την βοήθεια του οποίου εξάγουμε features για κάθε εικόνα.
- Υπολογισμός των cosine αποστάσεων για κάθε feature vector από τα υπόλοιπα. Κρατάμε τα k που είναι πιο μακριά από τα υπόλοιπα οπότε και υποψήφια σημεία ενδιαφέροντος.
- Μετά τον υπολογισμό των feature vectors, εφαρμόζουμε ένα φίλτρο, που εντοπίζει σε τι ποσοστό στην εικόνα εμφανίζεται το έδαφος του γηπέδου. Έτσι, μπορούμε να αφαιρέσουμε clips που μπορεί μεν να φαίνονται διαφορετικά από τα υπόλοιπα, αλλά δεν αφορούν πλάνα εκτός του γηπέδου, όπως για παράδειγμα οι θεατές. Η υλοποίηση έγινε μέσω της βιβλιοθήκης OpenCV[1]. Τα αποτελέσματα φαίνονται στην Εικόνα 1. Με βάση αυτό, δημιουργείται ένα threshold, με το οποίο φιλτράρονται οι cosine αποστάσεις των feature vectors.

Έτσι στο τέλος αυτής της διαδικασίας ένα score προτεραιότητας με βάση την απόσταση των εικόνων που εξήχθησαν από κάθε segment του αρχικού αγώνα.



Εικόνα 1. Στιγμιότυπο πριν και μετά την χρήση φίλτρου εντοπισμού γηπέδου.

2. Ήχος

Η μέθοδος που ακολουθήσαμε για την περίπτωση της διάστασης του ήχου είναι η εξής:

- Τμηματοποίηση του ήχου σε segments των 10s.
- Υπολογισμός της μέσης έντασης των segments που προέκυψαν χρησιμοποιώντας την librosa.
- Εφαρμογή ενός threshold βάση του οποίου χαρακτηρίζουμε ως irrelevant(δηλαδή μη υποψήφια highlights) τα segments με ένταση μικρότερη από το 30% της μέσης έντασης όλων. Υποθέτουμε δηλαδή ότι στην χειρότερη περίπτωση σε ένα highlight μπορεί να υπάρχει μέση ένταση του σήματος το πολύ λίγο μικρότερη της μέσης έντασης σήματος σε όλο τον αγώνα
- Εξαγωγή features με την pyAudioAnalysis.
- Επιλογή πιο σημαντικών features διερευνώντας ποια από αυτά παρουσιάζουν την μεγαλύτερη διακύμανση μεταξύ των segments.
- Υπολογισμός των cosine αποστάσεων για κάθε feature vector από τα υπόλοιπα. Κρατάμε τα k που είναι πιο μακριά από τα υπόλοιπα οπότε και υποψήφια σημεία ενδιαφέροντος.

Έτσι στο τέλος αυτής της διαδικασίας εξάγουμε ένα score προτεραιότητας με βάση την απόσταση των ηχητικών σημάτων για κάθε segment.

3. Κείμενο

Η μέθοδος που ακολουθήσαμε για την περίπτωση της διάστασης του ήχου είναι η εξής:

- Εξαγωγή της περιγραφής καθενός segment σε μορφή κειμένου με την βοήθεια ενός pretrained μοντέλου(whisper: small).
- Εξαγωγή της συχνότητας ομιλίας για κάθε segment(αριθμός λέξεων ανά κλιπ).
- Εξαγωγή του sentiment που εκφράζεται από την περιγραφή σε κείμενο κάθε segment με την χρήση pretrained μοντέλο [2].
- Συνένωση των δύο παραπάνω σε ένα feature vector.
- Υπολογισμός των cosine αποστάσεων για κάθε feature vector (μαζί με κάποιες ακόμα παραδοχές που θα παρουσιαστούν στην συνέχεια) από τα υπόλοιπα. Κρατάμε τα k που είναι πιο μακριά από τα υπόλοιπα οπότε και υποψήφια σημεία ενδιαφέροντος.

Έτσι στο τέλος αυτής της διαδικασίας εξάγουμε ένα score προτεραιότητας με βάση την απόσταση των transcriptions για κάθε segment.

4. Εξαγωγή Σημείων Ενδιαφέροντος

Στην τελική απόφαση των segments που αποτελούν σημεία ενδιαφέροντος ενός ποδοσφαιρικού αγώνα συνδυάσαμε τα αποτελέσματα που προέκυψαν από κάθε διάσταση ξεχωριστά. Πιο συγκεκριμένα:

1) Σε πρώτη φάση έγινε απλά μια ‘ψηφοφορία’ μεταξύ των αποτελεσμάτων του κάθε modality και προέκυψε το τελικό σύνολο των k σημείων ενδιαφέροντος. Ο παραπάνω πρώτος πειραματισμός έδωσε σχετικά ικανοποιητικά αποτελέσματα. Ωστόσο, ανάμεσα στα παραγόμενα highlights υπήρχαν και ορισμένα που αντιστοιχούσαν στην περίοδο του αγώνα που υπάρχει διάλειμμα. Αυτό προκύπτει από το γεγονός ότι τοσο στο μέρος του ήχου όσο και της εικόνας τα συγκεκριμένα segments διέφεραν (ως προς τα features) σε σχέση με τα υπόλοιπα σε μεγάλο βαθμό και οι distance based μέθοδοι που χρησιμοποιούμε τα εξέλαβαν σαν outliers..
2) Η λύση που υλοποιήσαμε για την αντιμετώπιση αυτού του προβλήματος αφορά το thresholding που προαναφέρθηκε βάση της μέσης έντασης και το οποίο επιβάλλει segments με χαμηλότερη μέση ένταση να μην λαμβάνονται υπόψιν. Τα αποτελέσματα αυτής της μεθόδου είναι πιο ικανοποιητικά σε σχέση με την προηγούμενη καθώς παρατηρείται ότι τα σημεία ενδιαφέροντος που προκύπτουν δεν αφαιρούν απλά τα segments που είδαμε πως αντιστοιχούν σε διαλλείματα αλλά και άλλα τα οποία δεν περιείχαν κάποια ουσιαστική φάση του αγώνα. Στην θέση αυτών προβλεφθηκαν πιο ουσιαστικά σημεία του αγώνα, γεγονός που επιβεβαίωσε την ορθότητα της επιλογής του thresholding που κάναμε.
3) Το τρίτο πείραμα αφορά την χρησιμοποίηση μόνο του κειμένου για την εξαγωγή των σημείων ενδιαφέροντος χρησιμοποιώντας και το bias για τα κλιπ που περιέχουν σιγή. Αξιοσημείωτο είναι ότι σε αυτή την περίπτωση παρατηρήσαμε ότι για

κάποιους αγώνες, στους οποίους ο εκφωνητής είχε καθαρή ομιλία και εξάγονταν υψηλής ποιότητας κείμενο μέσω του whisper, τα σημεία ενδιαφέροντος που προέκυψαν ήταν πιο εύστοχα σε σχέση με οποιοδήποτε από τα υπόλοιπα πειράματα. Ωστόσο, σε άλλους αγώνες όπου το παραγόμενο transcription ήταν χαμηλότερης ποιότητας, τα αποτελέσματα ήταν απογοητευτικά.

4) Το τέταρτο και τελευταίο πείραμα που υλοποιήσαμε είναι η ενσωμάτωση και του κειμένου σαν επιπλέον modality για να διερευνήσουμε εάν μπορεί, η συμμετοχή του συνδυαστικά με τα υπόλοιπα, να δώσει ακόμα ποιοτικότερα σημεία ενδιαφέροντος. Έχοντας λοιπόν το κίνητρο από το προηγούμενο πείραμα, αρχικά προσθέσαμε ένα επιπλέον bias στο κομμάτι του ήχου κρατώντας αυτή την φορά μονάχα τα segments τα οποία παρουσίαζαν μέση ένταση πάνω από το 80% της μέσης έντασης όλων. Έπειτα για κάθε ένα από αυτά χρησιμοποιήσαμε τα σημεία ενδιαφέροντος που έδινε ως αποτέλεσμα το cosine distance κάθε modality και στο τελευταίο βήμα κάνουμε πάλι μία απλή ‘ψηφοφορία’. Τα κλιπ που προέκυψαν ως σημεία ενδιαφέροντος ήταν εξίσου ποιοτικά με την περίπτωση του δεύτερου πειράματος ενώ σε ορισμένες περιπτώσεις ακόμα περισσότερο.

5. Αξιολόγηση Αποτελεσμάτων

Για την αξιολόγηση των πειραμάτων μας δημιουργήσαμε ένα ερωτηματολόγιο σε μορφή google form στο οποίο ζητήθηκε από άλλα άτομα να βαθμολογήσουν τα παραγόμενα σημεία ενδιαφέροντος από 5 αγώνες σε σχέση με τα πραγματικά highlight των αγώνων που είχε βγάλει η εκάστοτε διοργάνωση. Τα αποτελέσματα που προέκυψαν από την παραπάνω διαδικασία και για δείγμα περίπου 20 ατόμων παρουσιάζονται παρακάτω και αναδεικνύουν την ποιότητα των εν λόγω σημείων που εξάγαμε. Πιο συγκεκριμένα, η μέση αξιολόγηση που δεχτήκαμε, από το 1 (καθόλου highlights στο βίντεο) εώς το 5 (όλα τα highlight περιέχονται στο βίντεο), ήταν στο 3.2 .

	Game1	Game2	Game3	Game4	Game5
Score	3	4	3.5	2.5	3

6. Συμπεράσματα - Μελλοντική Δουλειά

Στην παρούσα εργασία καταφέραμε επιτυχώς την εξαγωγή των σημείων ενδιαφέροντος ενός ποδοσφαιρικού αγώνα. Θετικά και αρνητικά σημεία των τεχνικών που χρησιμοποιήθηκαν είναι τα εξής:

1. Σε όλο την διαδικασία έγιναν δύο παραδοχές:

   α. Η διάρκεια ενός highlight είναι 10s.

   b. Σε κάθε αγώνα υπάρχουν k highlights.

   Η πρώτη αποτελεί ήπια παραδοχή και δεν επηρεάζει τόσο τα παραγόμενα αποτελέσματα. Αντιθέτως η δεύτερη αποτελεί μια ισχυρή και αυθαίρετη παραδοχή που έγινε για λόγους ευκολίας και απλότητας.
2. Οι τεχνικές που χρησιμοποιήθηκαν δεν εμπεριέχουν κανένα είδους training αλλά βασίζονται μόνο στα features που παράγονται από τα 3 modalities που αναφέραμε.
3. Δείξαμε ότι ο συνδυασμός των 3 modalities με τους τρόπους που αναφέραμε μπορεί αποτελεσματικά να βελτιώσει την ποιότητα των πειραματικών αποτελεσμάτων.

Τελικά, σαν μελλοντική δουλειά πάνω στο συγκεκριμένο πρόβλημα θα μπορούσε να είναι η επέκταση της υλοποίησης που έχουμε φτιάξει ώστε:

- να ορίζονται με πιο ουσιαστικό τρόπο τα biases που εισάγονται στην διαδικασία σε σχέση με το κάθε modality(το thresholding εισήχθει με αυθαίρετο τρόπο και εντελώς διαισθητικά).
- να ορίζεται πιο ουσιαστικά η βαρύτητα που έχουν τα 3 modalities στην τελική απόφαση του εάν ένα segment αποτελεί σημείο ενδιαφέροντος.
- να χρησιμοποιηθούν περισσότερα features για όλα τα modalities, είτε hand-crafted είτε από pretrained μοντέλα.

7. Αναφορές

[1] Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

[2] Jochen Hartmann, "Emotion English DistilRoBERTa-base".https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, 2022. Ashritha R Murthy and K M Anil Kumar 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1110 012009