Action Spotting in Soccer Videos

Ahmed Symum Swapno
School of Data and Sciences
Brac University
Dhaka, Bangladesh
ahmed.symum.swapno@g.bracu.ac.bd

***Abstract*—This work proposes modifications to the Context-Aware Loss Function (CALF) model with the aim of enhancing the performance of action detection. The proposed changes include expanding the base convolutional network to six convolutional layers, replacing the convolutional layer in the segmentation module with two parallel convolutional layers, and switching from MaxPool to AvgPool as the pooling algorithm in the spotting module. The modified model achieves an average mAP of 43.7%, which is approximately 3% better than the original model. The classes with the greatest improvement are Direct free-kicks and Penalties. In contrast, the original model outperforms the modified model in terms of invisible actions. This work provides insights into the modifications that can be made to the CALF model to improve its performance for soccer action spotting.**

***Index Terms*—Action spotting, SoccerNet-v2, Context-Aware Loss Function (CALF), Convolutional Neural Network (CNN), Video interpretation, Temporal segmentation, Segmentation module, Pooling, Max Pool, Average Pool.**

I. Introduction

Sports video analysis has been a popular research area in recent years, with soccer being one of the most widely analyzed sports. The ability to detect actions in soccer videos, such as goals, tackles, and free-kicks, has many applications, including sports broadcasting, player performance analysis, coaching, referee assistance, fan engagement and many more. However, accurately detecting these actions in real-world soccer videos is a challenging task due to the complexity and variability of soccer games. Manually doing this task requires quite a lot of manpower and resources. An automated system to detect these action would have great implications on broadcasting and match analysis.

In recent years, the use of deep learning techniques has shown great potential in sports video analysis, including action detection in soccer videos. One such model is the Context-Aware Loss Function (CALF). The aim of this work is to try and improve the performance of the model CALF and show the comparison between the original model and the approaches taken in this work.

In the following sections, this paper provides an extensive literature review of the research field (II), a detailed description of the methodologies used (III), the acquired results (IV) along with an analytical discussion (V) and concluding remarks (VI).

II. Related Words

Action spotting was first introduced in SoccerNet [1]. This work provides a large scalable dataset of Soccer Videos annotated with specific actions. The dataset includes three classes (goal, card, substitution) and background class for no action. This paper puts forth mainly two tasks, Action Spotting and Action Classification. The authors define Action Spotting as finding an action with its temporal spot or anchor time. Action classification is classifying the spotted action into one of the three defined classes. Apart from the dataset, the authors also establish a baseline for these tasks. Their model architecture contains a feature extraction layer, a dimension reduction layer, a pooling layer and a classification layer. For feature extraction, they try out using ResNet, C3D, I3D and find that ResNet performs the best. They use PCA for dimension reduction. They investigate to find the best performing pooling method by trying out mean and max pooling, a custom CNN, SoftDBOW, NetFV, NetVLAD and NetRVLAD. Their experiment shows the VLAD based methods to be the performant. Their best model uses NetVLAD as the pooling layer. Lastly, for classification they use a sigmoid activation with Adam optimizer to minimize the multi-binary cross entropy loss. They also introduce a metric for general comparison of models, mAP(Mean Average Precision) which is calculated by taking the mean of Average Precision over all classes. Their best model reaches a 40.6% mAP.

Next up, we have CALF [2] which introduces a novel context aware loss function. Their loss function makes use of temporal context. Specifically, they classify frames as far before, just before, action, just after and far after. For their model architecture, like SoccerNet [1], they use ResNet for feature extraction. They integrate a segmentation module using a Time Shift Encoding and an action spotting module using an YOLO like loss function. The overall loss function is defined as the sum of the loss from spotting module and the weighted loss from segmentation module where the weight is a hyper-parameter. Their model establishes a new state of the art in action spotting in soccer with an mAP of 62.5%.

RMS-net [3] introduces a novel lightweight architecture for action spotting and classification that takes short video snippets as input and outputs the actions class with temporal offset. With effective handling of data imbalance combined with a masking strategy the model performance is further improved. The described model uses a 2D backbone to extract feature vectors from input frames. The features are combined using 1D convolutions and a maximum operation, followed by fully-connected layers for action classification and temporal offset regression. Using the same feature extraction technique as SoccerNet [1], the authors achieve a higher mAP of 65.5%.

AudioVid [4] proposes to integrate the audio along with the video for the task of action spotting. They argue that the audio can provide valuable insights into the actions taking place in the field. For instance, a goal leads to the crowd cheering and a foul or a red card leads to dissatisfaction. With this in mind, they implemented a multi-modal architecture to integrate both audio and video into consideration while spotting actions. The basic structure of both of their architecture follows a similar route to the one used in SoccerNet [1]. For feature extraction, they use ResNet for the video and VGG for audio stream. After the NetVLAD pooling layer, they integrate a dropout layer, a fully connected layer, a logit layer and a sigmoid activation layer. They experiment with different merge points to identify the best merge point to be before the fully connected layer. They investigate by using only audio, only video and both audio and video. Their findings show that the integration of audio leads to better performance. Their best model was evaluated to have an mAP of 56%.

SoccerNet-v2 [5] introduced a new dataset that extended on the previously mentioned SoccerNet [1]. Both of them have the same videos, but SoccerNet-v2 annotates a lot more actions while also extending the number of classes from 3 to 17. It reintroduced the task of action spotting and introduced two new tasks, camera segmentation and replay grounding. As baseline of action spotting, the authors use two of the baselines used in SoccerNet [1], Maxpool and NetVLAD, along with CALF and AudioVid. Their research found CALF to be the most perfomant among all with a mAP of 40%.

NetVLAD++ [6] proposed a novel temporally aware pooling method. Their pooling method is an extension of NetVLAD, called NetVLAD++, that takes the past and the future context into account. Their pooling module contains two layers of NetVLAD for the context before and after the action and is then aggregated. The authors follow the same architecture as in SoccerNet [1] with two key changes. For dimension reduction, this model uses a linear reduction layer instead of PCA and for the pooling layer, it uses NetVLAD++ instead of NetVLAD. These changes increase the mAP to 53.4% over all classes.

Cartas et al. [7] approaches the problem of action spotting in a new way to represent the input videos by representing the players in frame with a graph. To generate the graph from frames, they make use of player segmentation and localization, player motion vector and player classification. They use camera calibration proposed in SoccerNet-v2 [5] for localization. They implement an unsupervised player classifier for the classification of players into one of the 5 types, player of team 1, player of team 2, referee, goal keeper A and goal keeper B. For motion vector, they make use of optical flow. Using these information, the graph representation is generated. The architecture of their model for action spotting is multi-modal with graph and video input. For the graph data, they settle on DynamicEdgeConv after considering several graph neural network. For the video data, the authors follow SoccerNet [1] and use ResNet for RGB feature extraction.

Similar to AudioVid [4], the authors of this paper also integrate audio data in their model. So, in total their model consider 3 types of data, the RGB features from the video, the generated graph and the audio stream. Their experiment with different combination of using and not using these data shows that the model containing all three types of input performs the best with an mAP of 57.83%.

Zhou et al. [8] improves the feature extraction layer to improve the performance of action spotting. Instead of using ResNet for feature extraction, the authors of this paper propose using an ensemble of several pre-trained models namely TPN, GTA, VTN, irCSN, I3D slow. These were fine tuned for the specific purpose of action spotting using the SoccerNet-v2 dataset along with the some extra game videos that the authors collected. For the next layer of their architecture they consider NetVLAD++ [6] and Transformer. Their best model uses Transformer. The last layer uses a sigmoid activation and they use BCE loss function for their model. This model achieves state of the art state of the art performance with an mAP of 74.84%.

III. Research Methodology

A. Data

SoccerNet-v2, the data used for this work was provided by [5]. SoccerNet-v2 is a comprehensive dataset designed to facilitate soccer game video interpretation. It consists of 764 hours of footage from 500 soccer games, with approximately 300k manually annotated timestamps corresponding to approximately 110k actions distributed across 17 classes. SoccerNet v2 is the largest dataset in terms of event instances per class in the soccer domain, with a well-defined and consistent vocabulary concentrating on the soccer game and soccer broadcast domains. CALF [2] uses the features extracted by ResNet-152 and reduced to the size 512 using PCA. These are readily provided by SoccerNet. I will be using the same for this work.

B. CALF Overview

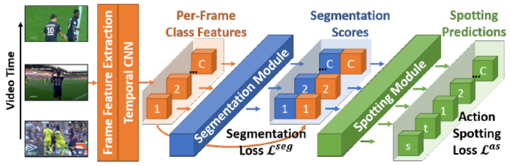

Fig. 1. CALF-Model Architecture

The model uses two different encoding for the initial annotations: a time-shift encoding (TSE) for temporal segmentation and a YOLO-like encoding for action spotting. The TSE is used to segment the temporal context surrounding each action based on their proximity to the action. The segments are then utilized by the temporal segmentation module to indicate whether an action occurred before, during, or after an action event.

On the other hand, the YOLO-like encoding is inspired by YOLO (You Only Look Once) and represents each ground-truth action of the video as an action vector composed of

2 + C values. The first value indicates the presence of the action, while the second value indicates the position of the frame annotated as the action. The remaining C values are the one-hot encoding of the action.

The model is designed with a hierarchical architecture, incorporating a Temporal CNN(base convolutional network and a pyramid-net), a segmentation module, and a spotting module, as illustrated in the figure 1. The model employs a context-aware loss function (CALF) [2], which consists of two components: a temporal segmentation loss and a spotting loss.

To train the model, the temporal segmentation loss and the spotting loss are combined by taking a weighed sum to compute the overall loss of the model. The Adam optimizer is used for the training.

The changes made to the model in this work are described in the following III-C, III-D, III-E subsections.

C. Base Convolutional Network

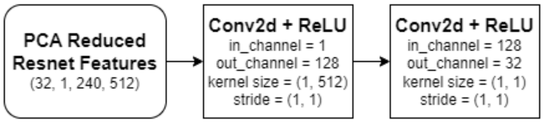

Fig. 2. CALF-Model Base Convolutional Network

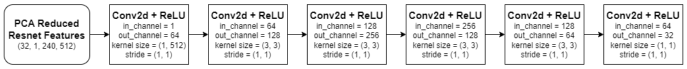

Fig. 3. Modified Base Convolutional Network

The Temporal CNN module consists of a base convolutional network and a pyramid network. As depicted in Figure 2, the original CALF model consisted of two 2D convolutional layers with kernel sizes of 1 × 512 and 1 × 1. To further enhance the performance of the model, a modification to the base convolutional network has been introduced in this work. Specifically, the base layer has been expanded to include six convolutional layers. The kernel size of the first and final layers of the modified network remain unchanged. However, the intermediate layers now utilize a larger kernel size of 3×3 Figure 3 depicts this modification by emphasizing the structure of the modified base convolutional network.

D. Segmentation Module

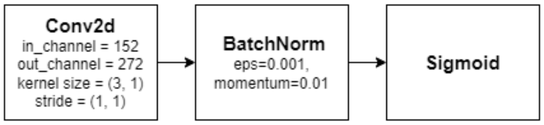

Fig. 4. CALF-Model Segmentation Module

The segmentation module is an important part of the CALF architecture. The TSE encoding in the preprocessing step is specifically done for the purposes of ensuring better learning at this module. The novel loss function proposed by CALF [2] takes into account the loss from the output of this particular module. Their segmentation module is made up of a combination of a convolutional layer and a batch normalization layer with a sigmoid activation as shown in the 4. For the purposes of improving the model, in this work, the convolution layer was replaced with two parallel convolution layers with half the output channels each. The outputs of these two layers were concatenated to feed to the batch normalization layer. This is visualized through figure 5.

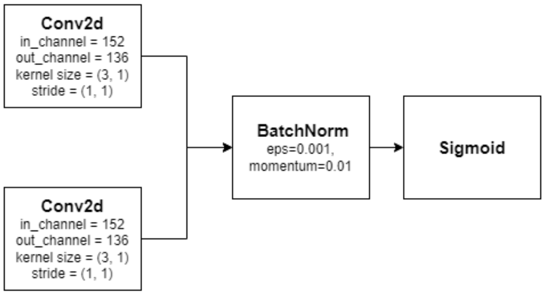

Fig. 5. Modified Segmentation Module

E. Spotting Module

Figure 6 shows, detection module used by CALF. It is comprised of a series of MaxPool and convolutional layers. The confidence branch and class branches are not shown in the figure in order to highlight the differences only. In this work, the pooling method was changed from MaxPool to AvgPool as shown in the figure 7.

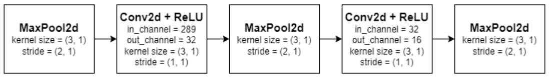

Fig. 6. CALF-Model Spotting Module

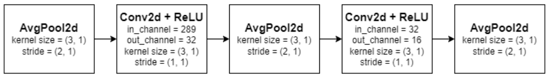

Fig. 7. Modified Spotting Module

IV. Results

The best model from the experimentation is the one with a modified base convolutional network and a segmentation module and average pooling for the spotting module. This model outperforms the original CALF model by approximately 3%. The average mAP comparison between the models is shown in the figure 8

The table I shows the comparison between the per class mAP of the different models trained for this work.

As it can be seen from the table, the proposed modification performs better than the original model on Average overall

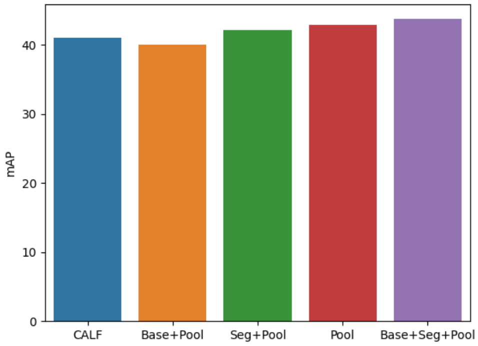

Fig. 8. Comparison of average mAP over all classes between the models. CALF: Original CALF Model, Base+Pool: Modified Base Convolutional Network and Average Pooling, Seg+Pool: Modified Segmentation Module and Average Pooling, Pool: Only Average Pooling, Base+Seg+Pool: Modified Base Convolutional Network, Modified Segmentation Module and Average Pooling.

TABLE I
COMPARISON OF PER CLASS MAP

Class	CALF	Base + Pool	Seg + Pool	Pool	Base + Seg + AvgPool
mAP	41.0	40.0	42.1	42.9	**43.7**
Visible	43.6	41.8	43.5	**45.5**	45.2
Invisible	**34.6**	29.0	31.3	29.0	32.5
Penalty	20.5	25.8	25.0	**33**.4	28.2
Kick-off	38.0	35.4	38.4	36.4	**39.8**
Goal	70.1	66.5	70.3	74.2	**75.3**
Substitution	56.0	52.3	52.7	**56.3**	54.5
Offside	24.5	13.8	**29.3**	29.0	29.0
Shots on target	27.5	29.1	28.0	23.5	**30.5**
Shots off target	28.5	30.9	29.5	30.2	**32.6**
Clearance	53.2	52.5	53.3	**55.2**	53.6
Ball out of play	64.5	66.5	65.9	65.9	**66.9**
Throw-in	59.8	61.4	60.8	59.9	**62.1**
Foul	53.4	51.6	52.4	53.8	**54.9**
Indirect free-kick	39.3	42.5	42.3	39.9	**43.9**
Direct free-kick	40.3	39.5	47.4	46.8	**49.1**
Corner	72.3	72.5	72.8	72.2	**73.8**
Yellow Card	46.9	40.0	47.0	**50.1**	49.0
Red Card	2.0	0.0	0.4	**2.8**	0.0
Yellow to Red Card	0.0	0.0	0.0	0.0	0.0

TABLE II
COMPARISON WITH OTHER BASELINE MODELS

Model	mAP (SoccerNet-v2)
MaxPool [1]	18.6
NetVLAD [1]	31.4
AudioVid [4]	39.9
Modified CALF (this)	43.7

mAP (43.7%) and average visible mAP (45.2%) but is outperformed by it in terms of invisible actions. The modified model performs better on most of the classes as well.

V. Discussion

Among the models trained for this task, the one with all the modifications performs the best overall. The highest improvements are shown in Direct free-kicks (+9% approx) and Penalty (+8% approx) classes. This model performs better than the original model in all the classes except the Red Card class. It shows the best performance of all the models trained in this work in 10 out of the 17 classes. The model shows the highest performances in Goal (75.5%) and Corner (73.8%) classes All the modified model except the one with a modified base and average pooling method performs better than the original model in overall mAP. It should be noted that, the model with only average pooling gives the average mAP of 42.9% which is only approximately 1% worse than the one with all the modifications. However, this model performs poorly on invisible actions and does not give an improvement over the original model in all the classes.

None of the models perform well on the Red Card and Yellow to Red Card class. Further exploration of the model is required in order to figure out the reasons behind this and improve the performance on these classes.

The proposed modifications also perform better than the baselines provided in SoccerNetv2 [5] as shown in the II.

Overall, the model shows an improvement over the baseline models and the original CALF model. However, the improvement is very small and further improvement require a more in depth research.

VI. Conclusion and Future Work

In conclusion, this work proposed modifications to the Context-Aware Loss Function (CALF) model for soccer action spotting, with the aim of improving its performance. The modifications included expanding the base convolutional network, replacing the convolutional layer in the segmentation module with two parallel convolutional layers, and switching from MaxPool to AvgPool as the pooling algorithm in the spotting module. The results showed that the modified model outperformed the original CALF model by approximately 3% in terms of overall mAP and on most of the classes, except for invisible actions.

The proposed modifications also performed better than the baselines provided in SoccerNetv2. However, further improvement is required to achieve more accurate detection of actions such as Red Cards and Yellow to Red Cards.

For future work, more in-depth research could be conducted to explore the reasons behind the poor performance of the models on these classes. Additionally, other techniques such as attention mechanisms or multi-modal fusion could be integrated to further improve the performance of the model. Finally, the proposed modifications could be applied to other sports and action spotting tasks to evaluate their effectiveness and generalization.

Overall, this work provides insights into the modifications that can be made to the CALF model to improve its performance for soccer action spotting and paves the way for future research in this area.

Acknowledgment

I would like to express my sincere gratitude to Meem Arafat Manab and Joyanta J. Mondal for their invaluable guidance and support throughout my research project. I am also thankful to them for providing me access to a research PC from the university, which was essential for completing my work. I am deeply grateful for their unwavering support and dedication.

References

[1] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem, "Soccernet: A scalable dataset for action spotting in soccer videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1711–1721.

[2] A. Cioppa, A. Deliege, S. Giancola, B. Ghanem, M. V. Droogenbroeck, R. Gade, and T. B. Moeslund, "A context-aware loss function for action spotting in soccer videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 126–13 136.

[3] M. Tomei, L. Baraldi, S. Calderara, S. Bronzin, and R. Cucchiara, "Rms-net: Regression and masking for soccer event spotting," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 7699–7706.

[4] B. Vanderplaetse and S. Dupont, "Improved soccer action spotting using both audio and video streams," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 896–897.

[5] A. Deliege, A. Cioppa, S. Giancola, M. J. Seikavandi, J. V. Dueholm, K. Nasrollahi, B. Ghanem, T. B. Moeslund, and M. Van Droogenbroeck, "Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4508–4519.

[6] S. Giancola and B. Ghanem, "Temporally-aware feature pooling for action spotting in soccer broadcasts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4490–4499.

[7] A. Cartas, C. Ballester, and G. Haro, "A graph-based method for soccer action spotting using unsupervised player classification," in *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, 2022, pp. 93–102.

[8] X. Zhou, L. Kang, Z. Cheng, B. He, and J. Xin, "Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection," *arXiv preprint arXiv:2106.14447*, 2021.