DEEP SOCCER CAPTIONING WITH TRANSFORMER: DATASET, SEMANTICS-RELATED LOSSES, AND MULTI-LEVEL EVALUATION

Ahmad Hammoudeh[1,2], Bastein Vanderplaetse[2,3], Stéphane Dupont[2]
[1]ISIA Lab, [2]MAIA Artificial Intelligence Lab, [3]MARO Lab
University of Mons, Mons 7000, Belgium
at.hammoudeh@gmail.com

ABSTRACT

This work aims at generating captions for soccer videos using deep learning. In this context, this paper introduces a dataset, model, and triple-level evaluation. The dataset consists of 22k caption-clip pairs and three visual features (images, optical flow, inpainting) for 500 hours of *SoccerNet* videos. The model is divided into three parts: a transformer learns language, ConvNets learn vision, and a fusion of linguistic and visual features generates captions. The paper suggests evaluating generated captions at three levels: syntax (the commonly used evaluation metrics such as BLEU-score and CIDEr), meaning (the quality of descriptions for a domain expert), and corpus (the diversity of generated captions). The paper shows that the diversity of generated captions has improved (from 0.07 reaching 0.18) with semantics-related losses that prioritize selected words. Semantics-related losses and the utilization of more visual features (optical flow, inpainting) improved the normalized captioning score by 28%. The web page of this work : https://sites.google.com/view/soccercaptioning

***K*eywords** video captioning · transformer · soccer · deep learning · football

1 Introduction

Before the spread of televisions, sports fans followed games by listening to sports commentators on the radio [1]. Since then, commentary has been an essential part of sports broadcasting. A sports commentator tells what is happening in a game and highlights main events such as goals and substitutions. Soccer commentators are typically humans with sports knowledge. That knowledge comes in two folds: A) a low level of knowledge extracted from what is happening inside the pitch (e.g., goal, pass, and penalty), B) a higher level of knowledge that relies on external information such as the context of a game, and the history of teams /players. The higher level of knowledge cannot be extracted by just watching a match. For example, when a commentator says: "<player> scores his seventh goal in the tournament," the commentator relies on a statistic that the player scored six other goals in the tournament.

Creativity is an additional dimension of soccer commentary. Some soccer commentators use figurative language or rhyming sentences to describe important actions. For example, a commentator described a goal in the Spanish league as firing a ball past a diving goalkeeper. Another commentator, Issam Chawali, said:

> "This player is an artist. In terms of Art, he is Charlie Chaplin presenting a silent show. If I am going to talk about scientists, then this is Alexander Fleming. This player has just discovered a treat for his team like Fleming, who discovered Penicillin."

The commentator created diverse sentences that describe a single scene. Although the created statements do not provide facts about the game, they entertain listeners. Sports commentary is not just reporting a game, but it should also be entertaining, metaphorical, emotional, and reflecting the personality of the commentator. In [2, 3], Schultz elaborated on the profound factor of entertainment in the sports broadcasting industry, indicating the fact that the American Broadcasting Company (ABC) hired a comedian as an announcer for a live television broadcast of a National Football League game.

With the expansion of recorded sports data, there has been increased interest in utilizing data-oriented decisions in soccer games. In [4], Schoenfeld illustrated how utilizing data helped Liverpool, a British Soccer team, to make better decisions. For example, hiring the attacking midfielder, Moe Salah, was based on a recommendation from Liverpool's data analysis team. Liverpool hired Salah for £36M in 2017 while his estimated market value as of June 2021 is £90M [transfermarkt.co.uk]. Liverpool and DeepMind, the well-known artificial intelligence research laboratory, aim at advancing sports analytics using artificial intelligence research. In a recently published paper, Liverpool and DeepMind proposed combining computer vision, game theory, and statistical learning to form a microcosm for sports analysis research [5].

This work introduces three main contributions:

- A dataset for soccer captioning. The dataset consists of 22k video-caption pairs along with features extracted from 500 hours of SoccerNet videos.
- A model for captioning soccer actions using semantics-related losses.
- A criterion for evaluating captioning models at three levels: Syntax, semantics, and corpus (diversity).

Fig.1 shows an example of a generated cation. More examples with videos are available on the web page of this paper: https://sites.google.com/view/soccercaptioning.

Generated caption: <player> <team> unleashes a shot, but his effort is poor and floats high over the bar.
Ground truth caption: <player> <team> unleashes a shot towards the goal, but his effort is not precise at all and it flies high over the bar.

Figure 1: An example of a generated caption

2 Related work

Generic Video captioning
Earlier approaches to converting videos into sentences relied on template-based hierarchical language models [6, 7, 8]. The idea was to identify a set of categories in a video and predict semantic relations between them to generate a sentence according to a pre-defined ontology (i.e., a person does something). Later, deep learning approaches left out template-based language models. The deep learning architectures for video captioning included an encoder-decoder architecture [9]. The idea of the encoder-decoder architecture is that transforming content from a visual domain(video) into a linguistic domain (text) is done through an intermediate domain (hidden state). The encoder transforms the visual features into a hidden state, and the decoder transforms the hidden state into a linguistic representation. After the attention mechanism's breakthrough in the area of natural language processing, it was implemented in video captioning [10, 11]. Recurrent Neural Networks (RNN) were used as decoders in earlier work in video captioning to learn the sequential nature of captions. However, transformers [12, 13] were shown to outperform RNNs in terms of ease of training and learning long-range sequential dependencies. While video-captioning typically relied on visual features extracted from video-frames such as Resnet and I3D features, including more modalities like audio (VGGish features) and speech improved video captioning as shown in [14, 15].

Sports video captioning
Although dense captioning [15] enables to describe details and subparts of the content, generic video captioning models only generate sentences from a macroscopic and general perspective, with no domain-specific information and details. For example, a soccer clip would be captioned as "people play football." A hierarchically grouped recurrent architecture was proposed for more domain-specific and detailed captions in the basketball domain [16]. The architecture is a fusion of three parts: 1) a CNN model for pixel-segmentation where every pixel is assigned to one of four categories: ball, first-team, second-team, and background. 2) a model encodes the movement of individuals using optical-flow features 3) a part for modeling the relationship between players. The three parts are fused in a hierarchically recurrent structure to caption NBA basketball videos [16]. On the same side, attention mechanisms with hierarchical recurrent neural networks were used to caption volleyball videos [17]. In the context of captioning soccer games using deep

learning, no similar work has been reported to the best of the authors' knowledge. However, humanoid commentators were developed for soccer games played by robotic dogs [18]. The humanoid commentators select an utterance from a predefined library using a rule-based event identification. The event identification depends on the game history and event recognition using a SIFT-based vision analysis algorithm.

SoccerNet
SoccerNet [19] is a large-scale dataset of 500 soccer games with three types of labeled actions (goal, substitution, and penalty). A recently published version, SoccerNet-v2, has extended the types of actions to 17 [20]. SoccerNet has been used to utilize deep learning in soccer camera calibration [21], and action spotting [22, 23, 24]. Multimodality using both acoustic and visual features improves event detection [24, 25]. An Improvement of 4.2% of the mean average precision in soccer action spotting and 7.4% in action classification were reported [24]. The work presented in this paper for video captioning is based on the SoccerNet dataset.

3 Proposed method

The proposed model for captioning soccer actions is a deep generative model conditioned on multiple features.

3.1 Dataset

This work introduces a dataset of 22k soccer video-caption pairs. The videos were taken from SoccerNet dataset. The captions were crawled from the FlashScore website [flashscore.com]. Captions were formatted. The names of players, coaches,and teams were replaced by single tokens representing each name's category (player, coach, team, time). For example, 'Christiano Ronolado' was replaced by '<player>' and 'José Mourinho' was replaced by '<coach>'. Identifying individual names of actors is a higher level of knowledge that entails external information, such as mapping between players' names and their numbers/faces, that cannot be learned from the created dataset.

3.1.1 Multiple Features

Learning from videos is challenging due to the problem of high dimensionality. Videos were represented using low-dimensional features in order to overcome the heavy processing of spatio-temporal data. Extracting multiple features provides more clues about the video. Three features (see Fig.2) were extracted: images, optical flow, and image inpainting features.

(a) RGB image

(b) Optical flow features extracted using the RAFT model. The moving players are highlighted in yellow.

(c) Image inpainting. From left to right: input image with missing parts, VAE, and reconstructed image.

Figure 2: Multiple visual features

- **RGB images (img)** extracted from SoccerNet videos at a frame rate of 1 image every half a second (2 FPS). The resolution of the RGB was reduced to 32x64x3 pixels in order to speed up the computation. The intuition behind the low-resolution images is to provide clues about the relative locations of visible objects in a frame.
- **Optical flow features (flow)** were extracted using a pre-trained optical-flow transformer (RAFT: Recurrent all-pairs field transforms [26]). The features were extracted by processing pairs of successive 398x224 frames with a time difference of 0.1 seconds. The intuition behind optical-flow features is to provide clues about the displacement of objects (players, ball, referees). For more efficient computation, each of the optical flow channels (U, and V) was reduced from 398x224 to 256 using PCA (Principal Components Analysis, a dimensionality reduction technique). A PCA matrix was calculated for a single match and applied to other matches achieving a retained variance of 98%.
- **Image inpainting features (vae)**. The intuition of extracting features from an image inpainting model is to substitute some missing clues that are not available in the low-resolution images or the optical flow features. The task of image inpainting is the process of completing the missing parts of an image. A variational auto-encoder (VAE) was trained to reconstruct missing parts of an image achieving a reconstruction loss of

0.08. The missing parts are: a horizontal rectangle of 398x24 pixels on the bottom, a vertical rectangle of 224x48 pixels on the right, and a square of 40x40 pixels at the center. The encoder compresses an input image into a latent space (latent distribution) of 2x2000 parameters: 2000 parameters represent mean values μ of a multivariate Gaussian distribution, and 2000 parameters represent the standard deviations σ. The decoder reconstructs the image and completes the missing parts by sampling the latent distribution. The inpainting features were extracted from the latent space.

The architecture details of the VAE are as the following: the encoder consists of 5 ConvNets, with a dilation rate of 1, 1, 2, 4, 8 respectively, followed by global average pooling and two separate fully connected layers (one for estimating μ and one for σ). The decoder consists of 3 deconvolutional networks (stride of 2). The ConvNets in both the encoder and the decoder are two-dimensional with a kernel size of 4x4. The VAE was trained on 1 Million images selected randomly from 4 million images extracted from SoccerNet videos (2 images per second).

3.1.2 Labeling uncaptioned clips using semi-supervised learning

The size of the training dataset can be increased by assigning captions to 58k uncaptioned clips from SoccerNet. Hence we can talk about two datasets depending on the accuracy of the assigned captions: 1) a dataset of accurate caption-clip pairs that consists of 22k pairs crawled online, and 2) a dataset of estimated pairs that consists of 58k clips labeled in a semi-supervised way: an uncaptioned clip gets the caption of its closest clip in the crawled dataset (see Fig.3).

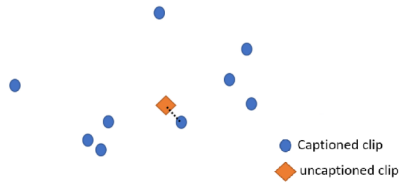
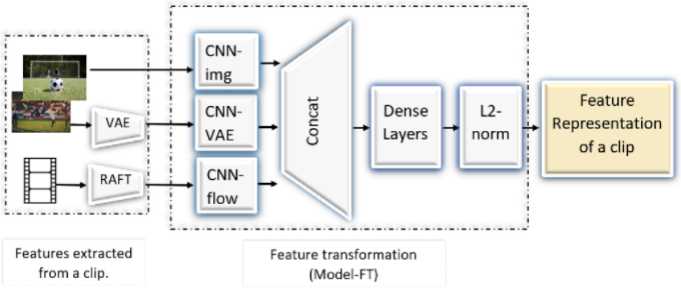
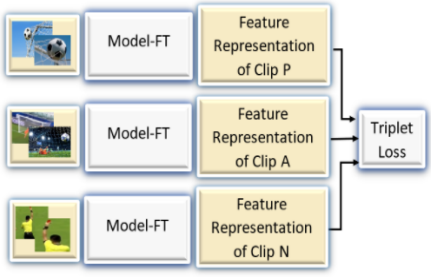


Figure 3: captioning using the nearest-neighbor approach

How was the closest clip found? The similarity between uncaptioned clips is expressed using the Euclidean distance between their feature representations. Hence the problem of measuring the similarity between two uncaptioned clips was considered as a feature representation problem. A feature transformation model (Model-FT) transforms the multiple features (images, optical flow, and image inpainting) into another representation that reflects the similarity between soccer clips. Model-FT was trained in a triplet network (see Fig.4). The triplet network takes 3 clips (two similar and one different) and learns to minimize the distance between similar clips (clip A and clip P) and maximize the distance between dissimilar clips (clip A and clip N) as in equation 1. The idea of the triplet network was introduced in [27, 28] and achieved a breakthrough in face recognition (FaceNet [28]).



(a) Left: multiple features(images, image inpainting, optical flow), center: Model FT, right: feature representation.



(b) Triplet network: A network of 2 similar clips (Anchor/A, and Positive/P) and 1 different clip (Negative/N)

Figure 4: Triplet network

$$Loss_{Triplet} = (FT(A) - FT(N))^2 - (FT(A) - FT(P))^2 + a \quad (1)$$

- $Loss_{Triplet}$ is the loss of the triplet network
- α is a constant equals to 0.2
- *FT* is Model-FT

One may wonder if soccer events can be captioned using the nearest neighbor approach (K-NN where K equals 1) in the resulting feature space. Evaluating the idea of captioning using the nearest neighbor approach in the experimental part (see 1-NN in Table 3) showed that it did not surpass the deep learning approach. Moreover, 1-NN limits the assigned captions to the available captions (no new sentences). On the other hand, caption generation using deep learning can generalize and create new captions [29].

How was the triplet network dataset created? The triplet network was trained on a dataset of 1M triplet examples were created. Each example consists of three clips from the crawled dataset. The Anchor clip (A) was sampled randomly. The positive clip (P) is the most similar captioned clip to clip (A) according to the similarity described in the following paragraph. The negative clip (N) was sampled randomly from clips with different types of actions.

How was the similarity between captioned clips measured? The similarity between two clips from the crawled dataset depends on three factors expressed in equation 2. The factors are:

- The type of actions: This factor looks at whether the two clips are from the same category or not. Clips from the same category are closer to each other than clips from 2 different categories (e.g., one is a goal and the other is a substitution). For this factor, a crisp boundary was set where captions from the same category get full score (1) and actions from different categories get 0.
- The similarity of captions: the similarity between two actions indicates the similarity between their clips. The similarity between two captions was estimated using the Bleu-1 score.
- The similarity of the significant words: More emphasis is put on selected words as some words influence the similarity more than others. For example, sharing a word like 'goal' indicates the similarity more than sharing a word like "is" or "the".

$$Similarity(clip_1, clip_2) = 0.25 \times (B1(caption_1, caption_2) + B1(SW_1, SW_2) + 2 \times I(action_1, action_2)) \quad (2)$$

- *B1*: Bleu-1 score
- SW_i: Significant words of caption i. Significant words are words that have technical meaning in the soccer-domain.
- $I(action_1, action_2)$: Identity function returns 1 if inputs are identical and 0 otherwise. The input is the type of action.

3.2 The model :Deep soccer captioning

The proposed model for captioning soccer actions is a language model conditioned on multiple features. The model predicts the next word given a video clip and a sequence of previous words. The architecture is a fusion of a transformer and ConvNets processed as explained below to produce the final caption. The main differences between the proposed model here and the transformer as in the attention paper [30] are that the encoder in this work is a set of ConvNets instead of attention, and a fully connected layer replaces the cross attention.

1. A Transformer processes linguistic features and predicts the next word given a sequence of words.The transformer (see part [A] in Fig.5) is a diminutive version of GPT that consists of just a single transformer block and two multi heads only while GPT uses a stack of such blocks [30]. For a sequence "Goal! A great ...", the transformer predicts the probability of each word in the vocabulary to be the next word. Although the transformer learns to predict the next word loosely, the final prediction is not taken from the transformer but the final layer.
2. ConvNets process the visual features described in section 3.1.1: RGB pixels, optical flow, and inpainting. The ConvNets (see part [B] in Fig.5) are identical to the ConvNets in Model-FT, followed by a fully connected layer. Model-Ft starts with three sets of ConvNets. Each set processes a particular feature: CNN-img for RGB pixels, CNN-flow for optical flow, and CNN-VAE for inpainting. The features are then concatenated and passed to the upper-level processing (part c).
3. An upper-level processing takes the outputs of part A and part B and yields the final prediction. That was represented by fully connected layers.

Tips:

- Transfer learning: The initial weights of CNN-img, CNN-flow, and CNN-vae in the captioning task were transferred from Model-FT.

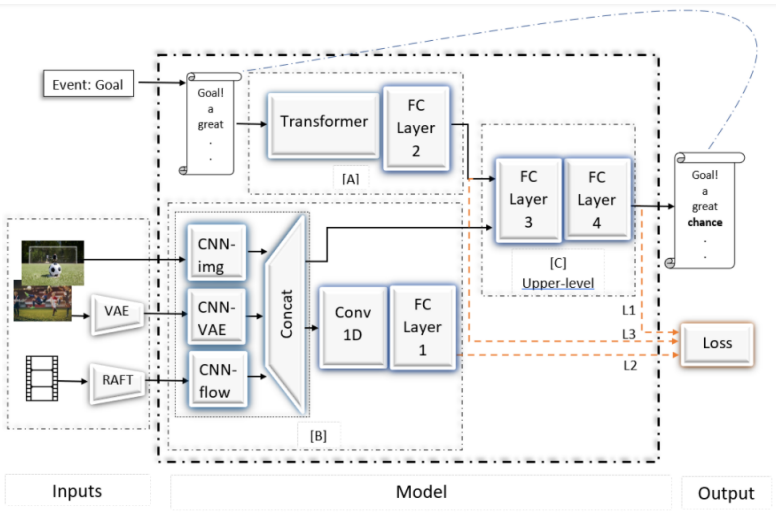


Figure 5: The proposed model

- Multi-heads: The fully connected layer of model A (FC2) is followed by two separate heads: one with a sigmoid activation function for the loss and another with Relu activation function passes the features to the next layer (FC3). Adding a sigmoid activation layer before sending the loss signal avoids the log(0) error. Feeding zero to the log cross-entropy loss results in errors as the log of zero is minus infinity. However, Relu was selected as the activation function when passing the features to FC3 because ReLU, in general, performs better with deep learning.
- The type of action was fed to the transformer as the first word of the sentence because conditioning captions on the type of action simplifies the soccer captioning task. The state-of-the-art model in SoccerNet-v2 challenge detects soccer actions with a mean average precision of 75%.

3.2.1 Loss functions and prioritizing significant words

Generated captions are learned in a supervised way by feeding three correction signals (losses) at three points of supervision. The overall loss is a weighted average of 3 losses. L1 is fed at the output layer of the captioning model, L2 is fed at the final layer of part B (vision part), L3 is fed at the final layer of part A (language part).

$$L = w1 \times L1 + w2 \times L2 + w3 \times L3 \tag{3}$$

- *Li*: Loss i
- *wi*: weight of loss i

L1 loss measures the deviation between the probability of each word in the vocabulary to be the next word *pp* and the ground truth *cp* using the categorical cross-entropy loss function CCE as in equation 4.

$$L1 = CCE(pp, cp) \tag{4}$$

L2 and L3 prioritize learning some words more than the rest by activating correction signals for the selected words only. Those words are called “Significant words” SWs. Words such as goal, free-kick, and penalty indicate specific technical meaning in the soccer game. For example, describing a goal scored from a penalty is different from a goal from a free-kick. Moreover, generating a word like ‘penalty’ mistakenly affects the next sequence of words more than a word such as ‘he’. Learning the significant words is prioritized by giving them a greater portion of the loss function.

Loss L3 at the final layer of part [A] is similar to L1 except that just significant words are corrected while other words are masked.

$$L3 = CCE(pA \times M, cp \times M) \tag{5}$$

- *pA*: the probability of each word to be the next word as predicted by the transformer (part [A])
- *M*: a mask where M[i] = 1 if i is a significant word and 0 otherwise.
- *cp*: the ground truth probability of every word in the vocabulary be the next word.

Loss L2 teaches the visual part [B] to predict the significant words in a caption given a video clip. Part [B] assigns 1 to the significant words in the caption and 0 otherwise. L2 as in equation 6 is a weighted average of the mean square error of 1s and 0s.

$$L2 = MSE(Y_{pred}, Y_{gt}) + sc \times MSE(Y_{pred} \times Y_{gt}, Y_{gt}) \tag{6}$$

- *MSE*: mean square error
- Y_{gt}: ground truth values: a vector of 1s and 0s reflecting which significant words are (1s) in the caption and which are not (0s)
- Y_{pred}: the prediction of Part [B]
- *sc*: a scaling factor. Ideally it is the ratio between the average number of 1's and 0s.

Assuming that two significant words appeared in a caption out of ten SWs, the ground truth in this case is a vector of ten digits (the number of SWs), two of them are ones. Something like Y_{gt} = 0001000010. MSE(Y_{pred} , Y_{gt}) is the mean square error for all 1s and 0s. However, the total number of significant words that are not in the caption (0s) is greater than those in the caption (1s). If a model predicts 0 always, the prediction will be correct for 80% of the digits. To overcome this, a term was added to the loss which is the mean square error of the ones only: MSE(Y_{pred} x Y_{gt}, Y_{gt}). The loss of 1s was scaled to balance the loss of the 1s and the loss of 0s. For the soccer captioning dataset, the scaling factor is 20. The semantic loss signal L2 was taken from a head added to the visual part. The head is a convolution layer and a fully connected layer added on the top of the visual part. The fully connected layer predicts the significant words in the clip and L2 penalizes wrong predictions.

3.3 Evaluation

The evaluation of generated captions included three aspects (see Fig.6): Syntax (word level), Soccer-oriented semantics (meaning level), and diversity (corpus level).

- **A)** Syntax-oriented evaluation: Captioning models are typically evaluated by comparing generated captions (hypotheses) against ground truth captions (references), word by word or chunk by chunk. That kind of evaluation is syntax-oriented as any differences at the word level count. This work used BLEU-scores (B@1 - B@4) and CIDEr from Microsoft CoCo evaluation code [31].
- **B)** Soccer-oriented evaluation (meaning level): The syntactic similarity between a generated caption and a reference caption does not always reflect the quality of a generated caption in the soccer domain. The example in table 1 shows almost two identical captions. Although 27 words out of 28 words are similar and one word only is different, that single word makes a technical difference. Such a difference does not appear clearly in syntax-oriented metrics like BLEU score or CIDEr. In order to evaluate the caption from the soccer perspective, another evaluation criterion was proposed: words with no technical meanings are removed, and just the significant words are evaluated. The quality of technical descriptions was evaluated using the precision and recall of the stemmed significant words in the caption.
- **C)** The diversity (corpus level): In the first experiments, the model learned a single caption for every type of action (1 generic caption for all fouls). Hence, we evaluate the diversity of generated captions. The diversity is measured by dividing the number of distinct generated captions N_{hypo} over the number of distinct ground truth captions N_{ref} as in equation 7.

$$Diversity = \frac{N_{hypo}}{N_{ref}} \tag{7}$$

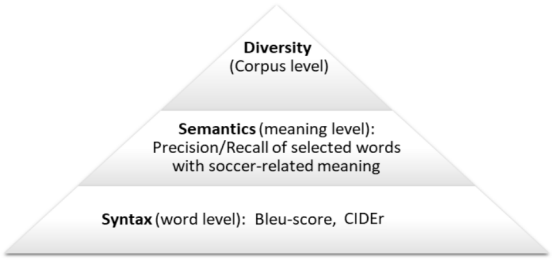


Figure 6: Three levels for evaluating soccer captions: Syntax (word level), Soccer-oriented semantics (meaning level), and diversity (corpus level)

	Reference	**Hypothesis**	**B@1**	**Precision of SW**
Caption	That was unbelievable. <PLAYER> <TEAM> changes the scoreline after getting on the end of a brilliant pass and firing a precise shot that goes inside the **right** post	That was unbelievable. <PLAYER> <TEAM> changes the scoreline after getting on the end of a brilliant pass and firing a precise shot that goes inside the **left** post	96%	-
SWs	[pass, shot, **right**, post]	[pass, shot, **left**, post]	-	75%

Table 1: A caption demonstrates how the precision of SW detects the technical differences better than generic captioning metrics (Bleu-score B@1)

The Stopping criterion depends on the captioning metrics, not the loss. The model that yields the best evaluation metrics for the caption generation task is not necessarily that with the best loss. Fig.7 shows the accuracy of predicting the next word on the validation set (green line) against the CIDEr evaluated on the validation set during training. The x-axis represents the number of epochs. The model yields the best CIDEr if the training stops slightly before the best validation loss as shown in Fig.7. The difference between the optimum point of the captioning metrics and the optimum point of loss could be attributed to the fact that the loss evaluates tokens only while the captioning metrics evaluate sentences. We train a language model to predict the next token and the loss depends only on one token. On the other hand, video captioning metrics evaluate the entire sentence (a generated sentence against a reference sentence).

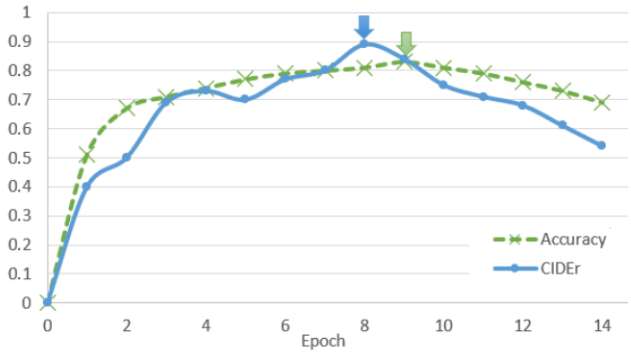


Figure 7: Caption generation metric (CIDEr) and the accuracy of predicting next word

For the purpose of using a single metric in the stopping criterion, the average of normalized BLEU-score, CIDEr, Precision and Recall was considered. In normalization, each evaluation metric was divided by a nominal value as in equation 8. The nominal values were the values scored by a baseline model.

$$NormalizedScore = \frac{\frac{B@4}{10} + \frac{CIDEr}{0.55} + \frac{Precision}{40} + \frac{Recall}{40}}{4} \tag{8}$$

4 Experimental evaluation

4.1 Data

The crawled dataset was divided randomly into 85% training, 5% validation, and 10% testing. The examples from the training dataset are used for developing the triplet network. Clips captioned using the triplet network were added to the training dataset only, but none of them was added to the validation or testing because the assigned captions' correctness is not guaranteed. The size of the vocabulary is 1400 tokens after removing words that appeared less than 4 times. Captions were tokenized after converting uppercase letters to lowercase letters. Tokens include words and a set of selected punctuations [! . ,]. The significant words and the details of how the significant words were selected are in Appendix A

4.2 Architectural details

The transformer architecture was designed to handle a predefined maximum sentence length of 40 tokens. More than 97% of the captions are of a length less than 40. Tokens and positions are embedded then fed into a transformer of 2 attention heads followed by fully connected layers. The output is a matrix of 40x1400. The row represents the position of a token, and the column represents a word in the vocabulary. The architectural details of the proposed model are shown in table 2. The model was trained with three feedback signals (L1, L2, and L3) using Adam optimizer with a learning rate of 0.00001. The initial weights of Model FT were transferred from the triplet network (no weight freezing).

Neural Networks	type	No. Layers	pooling	regularization	Input channels	Output channels
CNN-img	3D ConvNets	3	max	Dropout 0.5	30x32x64x3	lx20
CNN-flow	2D ConvNets	2	max	Dropout 0.5	30x256x2	lx20
CNN-VAE	2D ConvNets	3	max	Dropout 0.5	30x2000x2	lx20
Transformer	Attention	-	-	Dropout 0.1	30x1	lx256
FC 1	Fully connected	1	-	L2 Regularization 0.001	lx60	lx55
FC 2	Fully connected	1	-	L2 Regularization 0.001	lx256	lx1400
FC 3	Fully connected	1	-	L2 Regularization 0.001	lx1455	lx256
FC 4	Fully connected	1	-	L2 Regularization 0.001	lx256	lx1400

Table 2: Architectural details

4.3 Results and ablation study

The best model yielded a normalized score of 1.42. The best model was trained with semantics-related losses (L2 and L3) and relied on three visual features (img, flow, and vae). The initial weights of the visual part were transferred from model-FT. The role of the techniques used with the best model was investigated by evaluating the model after removing a key component. Removing the masking technique (L3 loss) resulted in the biggest degradation (normalized score of 1.34). Removing TL resulted in a slight degradation in the semantic-oriented metrics and a slight improvement in the syntax-oriented metrics yielding a normalized score of 1.41. The results are shown in Table 3, and the abbreviations are defined below.

	Syntax-oriented metrics					**Semantics-oriented metrics**		
	B@1	B@2	B@3	B@4	CIDEr	Precision	Recall	Normalized score
Best (No SL)	42.3	29.1	20.1	14.9	0.9	51.4	50.9	1.42
w/o TL	43.3	30	20.5	15.1	0.95	49	46.6	1.41
w/o VAE	40.5	27.7	18.8	13.5	0.92	51.1	50.8	1.39
with SL	40.2	26.6	16.7	9.9	0.83	54.4	53.5	1.36
w/o img	39.1	26.5	17.1	12	0.84	56	53.3	1.36
w/o flow	42.4	29.2	19.9	14.5	0.81	47.2	50.8	1.35
w/o L2	40.2	26.9	18.3	13.1	0.84	50.6	51.7	1.35
w/o L3	41.9	28.1	18.3	13.2	0.9	49.2	46.4	1.34
basic	39.8	25.8	16.2	11.8	0.65	40.6	43.3	1.11
1-NN	39.6	24.2	14.6	9.8	0.53	42.6	43.6	1.02
w/o visual part	36.2	23	14.4	9.8	0.54	39.8	40	1

Table 3: Captioning results and ablation study

The techniques used for captioning model are:

- L3: Prioritizing significant words in the language part using masked loss L3.
- L2: Prioritizing significant words in the vision part using loss L2.
- Transfer learning (TL) : Transferring the initial weights of model-FT from the triplet network.
- Semi-supervised learning (SL) : Trained on the extended dataset (both the dataset of the crawled captions and that of assigned captions).

- Multiple features: images (img), optical flow (flow), Image inpainting (VAE)
- 1-NN: captioning using the nearest neighbor approach.
- basic: trained using L1 loss on Resnet features only (None of the techniques above was used: No L2, L3, TL, SL, multiple features).

The extended dataset did not improve the results. Training the best model on the extended dataset yielded a normalized score of 1.36. That decline could be attributed to the inaccuracy of the assigned captions and the different distribution of events between the crawled dataset and the extended dataset. The 'throw-in' and 'ball out of play' events constitute 64% of the examples in the extended dataset and just 6% of the crawled dataset (see Table 5). Table 4 evaluates the generated captions per each type of action. It also shows that the idea of semantics-related losses (L3 and L2) boosted the diversity of the generated captions from 0.07 to 0.18.

Action	B@1	B@2	B@3	B@4	Precision	Recall	No. Distinct Captions (Ground Truth)	No. Distinct Captions with semantics-related losses	No. Distinct Captions without semantics-related losses
Foul	35.1	24.3	16.6	11.1	49.0	57.4	134	12	4
Shot on target	44.6	31.4	18.9	13.0	45.9	49.1	159	28	8
Shot off target	52.7	34.9	25.6	21.2	46.1	51.9	146	36	6
redcard	27.7	16.0	10.9	0.0	10.9	6.1	5	4	3
Substitution	49.6	35.7	26.5	21.9	33.6	63.3	60	5	4
direct free kick	32.0	23.8	17.7	12.7	91.9	65.5	51	9	3
corner	46.7	32.2	22.4	17.9	61.1	52.9	109	9	8
kick off	19.2	10.9	9.2	8.0	26.7	8.1	20	5	3
yellow card	41.1	24.3	15.6	10.9	50.4	41.7	72	3	1
off side	32.2	24.6	21.2	18.7	67.2	85.1	30	5	3
ball out of play	43.4	30.2	20.8	15.6	44.1	47.5	81	28	7
indirect free kick	37.5	26.6	16.9	0.0	62.6	72.1	18	6	6
goal	39.2	28.7	20.5	14.5	66.6	34.6	82	12	3
penalty	48.2	36.7	29.0	19.4	23.4	39.4	4	3	3
yellow red card	28.6	13.7	9.0	6.6	22.1	6.8	4	4	1
clearance	34.7	20.8	9.5	5.3	16.9	29.6	10	6	5
throw-in	13.2	0	0	0	0	0	1	1	1

Table 4: Results per action

5 Conclusion

This work introduced a dataset and model for soccer captioning and proposes evaluating captioning models captions at three levels: syntax (the popular evaluation metrics such as BLEU-score and CIDEr), meaning (the quality of description for a domain-expert), and corpus (the diversity of generated captions). It also shows that the diversity of generated captions improved from 0.07 to 0.18 with semantics-related losses that prioritize selected words. Semantics-related losses along with the utilization of multiple features (optical flow, inpainting) improved the normalized captioning score by 28%.

References

[1] Andrew Crisell. *An introductory history of British broadcasting*. Routledge, 2005.

[2] Brad Schultz. *Sports broadcasting*. Focal Press, 2002.

[3] Bradley Schultz. *Sports media: Reporting, producing, and planning*. Routledge, 2012.

[4] Bruce Schoenfeld. How data (and some breathtaking soccer) brought liverpool to the cusp of glory. *The New York Times*, 2019.

[5] Karl Tuyls, Shayegan Omidshafiei, Paul Muller, Zhe Wang, Jerome Connor, Daniel Hennes, Ian Graham, William Spearman, Tim Waskett, Dafydd Steel, et al. Game plan: What ai can do for football, and what football can do for ai. *Journal of Artificial Intelligence Research*, 71:41–88, 2021.

Action	Number of clips captioned by crawling	Number of uncaptioned clips
Shots on target	2326	2059
corner	3891	317
substitution	2171	160
yellowcard	1646	108
Shots off target	2528	1716
foul	3085	7022
kick-off	769	292
Ball out of play	1252	25419
goal	1295	137
Direct free-kick	795	862
offside	1267	581
Indirect free-kick	399	4817
penalty	87	10
redcard	44	2
clearance	68	3197
yellow-red card	33	4
throw-in	11	12023
Total	21667	58726

Table 5: Soccer captioning dataset - Number of examples

[6] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2712–2719, 2013.

[7] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. Translating video content to natural language descriptions. In *Proceedings of the IEEE international conference on computer vision*, pages 433–440, 2013.

[8] Jesse Thomason, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Raymond Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1218–1227, 2014.

[9] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.

[10] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015.

[11] Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. Less is more: Picking informative frames for video captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 358–373, 2018.

[12] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6578–6587, 2019.

[13] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10870–10879, 2020.

[14] Jack Hessel, Bo Pang, Zhenhai Zhu, and Radu Soricut. A case study on combining asr and visual features for generating instructional video captions. *arXiv preprint arXiv:1910.02930*, 2019.

[15] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 958–959, 2020.

[16] Huanyu Yu, Shuo Cheng, Bingbing Ni, Minsi Wang, Jian Zhang, and Xiaokang Yang. Fine-grained video captioning for sports narrative. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6006–6015, 2018.

[17] Mengshi Qi, Yunhong Wang, Annan Li, and Jiebo Luo. Sports video captioning via attentive motion representation and group relationship modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2617–2633, 2019.

[18] Manuela Veloso, Nicholas Armstrong-Crews, Sonia Chernova, Elisabeth Crawford, Colin McMillen, Maayan Roth, Douglas Vail, and Stefan Zickler. A team of humanoid game commentators. *International Journal of Humanoid Robotics*, 5(03):457–480, 2008.

[19] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1711–1721, 2018.

[20] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc 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*, pages 4508–4519, 2021.

[21] Anthony Cioppa, Adrien Deliege, Floriane Magera, Silvio Giancola, Olivier Barnich, Bernard Ghanem, and Marc Van Droogenbroeck. Camera calibration and player localization in soccernet-v2 and investigation of their representations for action spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4537–4546, 2021.

[22] Silvio Giancola and Bernard Ghanem. Temporally-aware feature pooling for action spotting in soccer broadcasts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2021.

[23] Matteo Tomei, Lorenzo Baraldi, Simone Calderara, Simone Bronzin, and Rita Cucchiara. Rms-net: Regression and masking for soccer event spotting. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7699–7706. IEEE, 2021.

[24] Bastien Vanderplaetse and Stephane 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*, pages 896–897, 2020.

[25] Mathilde Brousmiche, Stéphane Dupont, and Jean Rout. Intra and inter-modality interactions for audio-visual event detection. In *Proceedings of the 1st International Workshop on Human-centric Multimedia Analysis*, pages 5–11, 2020.

[26] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[27] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1386–1393, 2014.

[28] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015.

[29] Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. Compositional generalization in image captioning. *arXiv preprint arXiv:1909.04402*, 2019.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[31] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

A Caption skeleton: How were the significant words selected?

Significant words were selected based on the ontologies behind soccer captions. The ontologies of soccer were found by induction after analyzing soccer captions. For example, a caption that describes a goal conveys three kinds of information: A) what action preceded the goal (e.g., penalty, corner-kick, free kick), B) how the goal was kicked (by head or by foot), and C) the final destination of the ball (right, left, middle). Inducted ontologies for goals, shots, free-kicks, corner-kicks, and penalties are represented in tables 6-8 along with the related significant words.

Class	Subclass	Related significant word(s)
From	pass	pass
	rebound	rebound
	penalty	penalty
	corner kick	corner, kick
	free kick	free, kick
How	by head	head
	By foot	-
To	right	right
	left	left
	middle	middle
	back	back
	post	post
	high	high

Table 6: Captions skeleton - Goal/Shot

Class	Subclass	Related significant word(s)
Before	Intended or not	work, decide
How far	short range	short
	mid range	long
	long range	long
	destination	float, pass, cross, penalty, area
After	cleared	clear, intercept, cut, defend, opponent, fail, nothing
	goal	goal
	reached teammates	teammate

Table 7: Captions skeleton - Corner-kick/Free-kick

Class	Subclass	Related significant word(s)
After	wasted	wasted save, waste
	goal	goal
Direction	right	right
	left	left
	middle	middle
	high	high

Table 8: Captions skeleton - Penalty

The significant words are [’goal’, ’post’, ’pass’, ’net’, ’corner’, ’goalkeeper’, ’penalty’, ’bar’, ’kick’, ‘shot’, ’cross’,’freekick’, ’yellow’, ’red’, ’card’, ’area’, ’rebound’, ’free’, ’head’, ’offside’, ’throw-in’, ’box’, ’right’, ’left’,’over’, ’inside’, ’bottom’, ’back’, ’up’, ’side’, ’top’, (’loft’, ’float’), ’middle’, ’ ’outside’,’high’, ’mid-range’, ’roof’, ’out’, ’off’, ’first’, ’second’, ’half’, ’long’, ’low’, ’flag’, ’linesman’, ’short’,’defender’, ’teammate’, ’opponent’, (’work’, ’decides’), (’replace’, ’change’,’substitution’, ’substitute’), (’cut’, ’intercept’), (’foul’, ’tackle’, ’challenge’),(’nothing’, ’clear’, ’save’, ’fail’, ’waste”,’block’)]. The number of SWs is 55 after combining semantically equivalent words (between parentheses).

B The effect of the linguistic variability

An ideal model generates a caption identical to the reference. In this case, the theoretical upper limits of the evaluation metrics are: B@1: 100%, B@2: 100%, B@3: 100%, B@4: 100%, METEOR:100%, ROUGE-L: 100%, CIDEr: 10. However, to evaluate the effect of the linguistic variability on the metrics, a simple experiment was conducted. A small set of reference captions and hypothesis captions was created under the following conditions:

- The captions are from the dataset
- The reference caption and the hypothesis caption describe the same action
- The captions are from the top 4 most frequent captions for that action.
- The reference caption and the hypothesis caption are syntactically different but semantically similar.

Table 9 shows the selected captions.and the results of evaluating the effect of the linguistic variability on the evaluation metrics are in table 10. One may conclude that the model of this work performs well as it achieves metrics higher than the results of this experiment.

Action	Caption 1	Caption 2
Corner	<PLAYER> <TEAM> takes the corner but it's cleared.	This corner kick is taken by <PLAYER> <TEAM>.
Penalty	<PLAYER> <TEAM> is going to take the penalty!	<PLAYER> <TEAM> is heading towards the penalty spot to take it.
Substitution	<COACH> has decided to make a change. <PLAYER> <TEAM> replaces <PLAYER>.	Here is a change. <PLAYER> is going off and <COACH> gives the last tactical orders to <PLAYER> <TEAM>.
Goal	Goal! The ball reaches to <PLAYER> <TEAM> and he fired home from close range. The score is <STAT>.	Goal! <PLAYER> puts the ball on a plate for <PLAYER> <TEAM>, who scores with a simple close-range finish. It's <STAT>.
Shots on target	<PLAYER> <TEAM> misses a good chance to score. An inch-perfect cross into the box finds <PLAYER> <TEAM> who sends the ball well over the bar.	<PLAYER> <TEAM> gets the ball and drives a shot high over the bar.
kick off	The first half has started	The referee blows his whistle and we are underway.
	The first half has just begun.	Today's match has just started, enjoy the game!
Throw-in	It's a throw-in for <TEAM>.	The linesman signals a throw-in for <TEAM>.
Red card	<PLAYER> <TEAM> knows that the game is over for him after receiving a red card from the referee for fouling his opponent!	<PLAYER> <TEAM> receives a red card after his awful challenge. He completely lost his temper and referee <REF> sends him off the pitch.
Yellow-Red card	<PLAYER> <TEAM> should have been more careful. He commits a bad foul, receives his second yellow card and is sent-off!	<REF> blows his whistle and <PLAYER> <TEAM> is shown a second yellow card for his foul. His manager will not be pleased. He has lot of time to think about it as he walks off.
Yellow card	<REF> shows a yellow card to <PLAYER> <TEAM> for a tough tackle.	The foul by <PLAYER> <TEAM> is worthy of a card and a yellow is duly shown by <REF>.
Foul	The foul by <PLAYER> <TEAM> was seen by <REF> who did not hesitate to blow the whistle.	<PLAYER> <TEAM> commits a foul and <REF> immediately signals a free kick.
Offside	Flag goes up against <PLAYER> <TEAM> and the referee blows his whistle for offside.	The game is interrupted as <PLAYER> <TEAM> is flagged offside.
free kick	<PLAYER> <TEAM> takes a direct free kick.	<PLAYER> <TEAM> is going to send the ball in from the free kick.
Ball out of play	<PLAYER> <TEAM> wastes a good opportunity as his pass into the box is blocked by the defence. The ball goes out of play. <TEAM> are awarded a corner kick.	<PLAYER> <TEAM> is unable to feed an accurate cross into the box. The referee signals a corner kick to <TEAM>.
Clearance	<PLAYER> <TEAM> skips past his man but cannot keep the ball in play. The ball is off of the pitch and it's a goal kick for <TEAM>.	The ball is cleared after <PLAYER> <TEAM> attempted to dribble past an opposing player. The ball goes out of play and <TEAM> will have a goal kick.

Table 9: Captions used to evaluate the effect of linguistic variability on the evaluation metrics

score	B@1	B@2	B@3	B@4	METEOR	ROUGE	CIDEr
	43	27.4	16	8.6	27	35	0.99

Table 10: The effect of the linguistic variability on the evaluation metrics - Results