Dense Video Captioning Using Unsupervised Semantic Information

Valter Estevam[a,b,*], Rayson Laroca[b,c], Helio Pedrini[d], David Menotti[b]

[a] *Federal Institute of Paraná, Irati-PR, 84500-000, Brazil*
[b] *Federal University of Paraná, Department of Informatics, Curitiba-PR, 81531-970, Brazil*
[c] *Pontifical Catholic University of Paraná, Postgraduate Program in Informatics, Curitiba-PR, 80215-901, Brazil*
[d] *University of Campinas, Institute of Computing, Campinas-SP, 13083-852, Brazil*

vlejunior@inf.ufpr.br rayson@ppgia.pucpr.br helio@ic.unicamp.br menotti@inf.ufpr.br

Abstract

We introduce a method to learn unsupervised semantic visual information based on the premise that complex events can be decomposed into simpler events and that these simple events are shared across several complex events. We first employ a clustering method to group representations producing a visual codebook. Then, we learn a dense representation by encoding the co-occurrence probability matrix for the codebook entries. This representation leverages the performance of the dense video captioning task in a scenario with only visual features. For example, we replace the audio signal in the BMT method and produce temporal proposals with comparable performance. Furthermore, we concatenate the visual representation with our descriptor in a vanilla transformer method to achieve state-of-the-art performance in the captioning subtask compared to the methods that explore only visual features, as well as a competitive performance with multi-modal methods. Our code is available at https://github.com/valterlej/dvcusi.

Keywords: Visual Similarity, Unsupervised Learning, Co-Occurrence Estimation, Self-Attention, Bi-Modal Attention

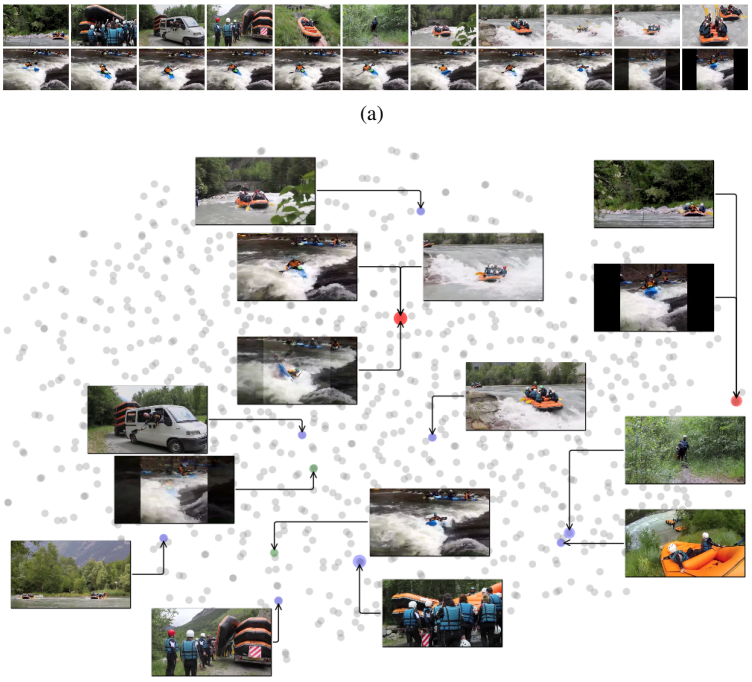
1. Introduction

In this work, we aim to perform Dense Video Captioning (DVC) [19] using only visual features. DVC is a complex task that involves localizing events and providing a suitable description for them in untrimmed videos. This task differs from Video Captioning because the events are usually not perfectly delimited, making generating accurate captions more challenging. Nowadays, DVC has been tackled using multi-modal features: visual and audio [15], visual, audio, and speech [16, 6]. Nevertheless, audio features may not always be available or indicative of the content in the video, and the same holds true for speech features. Therefore, it becomes imperative to devise approaches that rely solely on visual information. In this sense, we propose a new visual descriptor learned with an unsupervised method that can encode the co-occurrence visual similarity of short video clips (i.e., lasting a few seconds) to be used in the DVC task. Our inspiration is that humans can recognize similar video fragments and infer the later scenes from a movie they have not seen before, relying entirely on their prior knowledge and contextual information.

*Corresponding author
Email address: valter.junior@ifpr.edu.br (Valter Estevam)

This is an author-prepared version. The published version is available on *ScienceDirect* (DOI: 10.1016/j.jvcir.2024.104385).

Recently, several methods have been proposed for learning deep representations in an unsupervised manner [42, 13, 4, 14]. These methods usually combine a deep neural network (e.g., CNN or autoencoders) and a clustering method (e.g., k-means or agglomerative clustering). In the general framework, clusters are used to organize latent representations into soft labels which, in turn, are used in a supervised model that updates the encoder weights [1], improving the latent features. However, our goal is slightly different. We are interested in generating a dense representation encoding the visual relationships, where short clips are similar to each other and occur in their temporal context. These relationships are not captured by the aforementioned methods, that are optimized to produce more discriminative features.



(a)

(b)

Figure 1: Examples of visual similarities. (a) Two video fragments with about 28 seconds from YouTube (v_dBNZf90PLJ0 and v_j3QSVh_AhDc). They share some visual similar short clips. (b) A 2D t-SNE representation for the whole visual vocabulary. Some shared fragments are highlighted in red.

The intuition behind the proposed method is that long and complex events can be decomposed into short and simple events, as illustrated in Figure 1a – which shows two videos of related water sports: rafting and kayaking. We first identify similar events by splitting the videos into short clips and then extract visual features using the i3D method [5] for each short clip. A mini-batch k-means method groups representations based on their Euclidean distance producing a visual codebook, and a discrete representation is obtained by the sequence of cluster label numbers. Afterward, inspired by the GloVe method [27], we compute a co-occurrence matrix for this codebook and learn a dense representation by training a neural network to predict the pre-computed co-occurrence probability of any two

visual words, as detailed in Section 3.1.

In Figure 1b, a 2D t-SNE [23] visualization was used to project the entire visual codebook (drawn as gray dots). The clips from the first and second videos are represented with blue and green dots, respectively. Observe that the final content from the first video is much similar to the content of the second one (see the red dots) and that fragments with similar content are close to each other.

Our semantic descriptor can be employed in the DVC task, which consists of two subtasks: temporal event proposal and video captioning. In this work, we employ a popular strategy of handling these tasks independently. More specifically, we use a multi-headed bi-modal proposal module [15] for event proposal generation, and a vanilla transformer [33] for video captioning.

In summary, the contributions of this work are (i) an unsupervised descriptor that can be easily employed in dense video captioning, using either visual features alone or in combination with multiple modalities; (ii) the visual similarity proved to be efficient to generate event proposals replacing the audio signal adopted in the BMT method [15]; and (iii) our captioning results in the learned proposals scenario (i.e., the most complex setting) showcase the descriptor's ability to accurately capture the visual similarity between both seen and unseen clips, achieving state-of-the-art performance when considering only visual features and competitive performance when compared to multi-modal methods.

2. Related work

Dense video captioning was first introduced in [19]. It involves proposing a temporal event localization in untrimmed videos (i.e., event proposal generation) and providing a suitable description for the event in fluent natural language (i.e., video captioning). Presently, DVC research has primarily focused on general events [19], sport events [45], and cooking activities [46]. For this particular study, our focus lies on general events. In Section 2.1, we elaborate on the event proposal generation, while Section 2.2 is dedicated to discussing the video captioning subtask. In Section 2.3, we introduce some approaches for learning deep representations in an unsupervised manner.

2.1. Event Proposal Generation

Event proposal generation is a challenging task because events have no predefined length, ranging from short frame sequences to very long frame sequences with partial or complete overlap. The general strategy is to define a set of anchors and a deep representation that encodes the video. Each anchor receives a confidence score from binary classifiers, and the highest-scoring anchors are passed to the captioning module jointly with their associated representation.

Krishna *et al.* [19], for example, used a forward sliding window strategy, based on Deep Action Proposals (DAPs) [9], with four strides (1, 2, 4, 8), to sample video features with different time resolutions and feed them

into a Long Short-Term Memory (LSTM) unit that encodes and provides past and current contextual information. On the other hand, Wang *et al.* [37] proposed to explore not only past and current context but also the future context to predict and estimate confidence scores. They adopted a forward and a backward pass on the LSTM units and merged the confidence scores using a multiplicative strategy. They also proposed an attentive fusion approach to compute the hidden representation. In both works, there are two models, one for each task, trained with an alternate procedure where the proposal module is trained first and then the captioning module is trained while the proposals are fine-tuned.

While most works overlooked the intrinsic relationship between the linguistic description and the visual appearance of the events, taking into account only visual features obtained by the C3D model [31] pre-trained on the Sports 1-M dataset [17], Zhou *et al.* [46] leveraged the influence of the linguistic description in the proposal module with a vanilla transformer model trained in an end-to-end manner. Similarly, Iashin and Rahtu [15] proposed a Bi-Modal Transformer (BMT) model using i3D [5] and VGGish [12] features (i.e., visual and audio) to learn video representations conditioned by their linguistic description. First, the authors trained a captioning model using the ground truth events and sentences. Then, they used the encoder to feed a multi-headed event proposal module composed of 1D Convolutional Neural Networks (CNNs) with different kernel sizes.

2.2. Video Captioning

Considering the captioning task, most recent methods address this problem in two steps [35, 36, 7]. In the first step, a neural network encodes the entire video, frame by frame, into a compressed representation given by the hidden state of a Recurrent Neural Network (RNN). Then, in the second step, a decoder, usually an RNN, is fed with this representation to learn a probability distribution on a predefined vocabulary, producing a sentence, word-by-word. More recently, encoder-decoder models based on transformers [33] have been proposed [46, 16, 15], however, the best strategy for encoding video information before feeding the encoder remains an open issue. On the one hand, 2D CNN models can be fed frame by frame, producing long-range feature sequences that are difficult to process using RNN due to the well-known vanishing and exploding gradient problems [20]. LSTM and Gated Recurrent Unit (GRU) combined with soft and hard attention, or even Transformers with self-attention mechanisms, conduct the models to focus on more representative segments. These approaches boost performance but do not solve the video representation problem. On the other hand, when the entire video is fed into a 3D CNN (e.g., as in [44]), we come across the problem of information compression. All semantics are stored in a feature map with a fixed length, and converting this feature map in sentences is difficult because much relevant information can be lost or suppressed – especially on videos much longer than those used to train the 3D CNN.

This problem is more pronounced in captioning than in event proposal generation and has been circumvented by adding modalities such as audio and speech, objects, and action recognition [25, 10, 16, 15, 6]. For example, Iashin

and Rahtu [16] proposed a framework called Multi-modal Dense Video Captioning (MDVC), in which each modality is fed into a separated encoder-decoder transformer and, in the end, their hidden representations are concatenated and fed to a language generator module composed of two dense layers and one softmax layer.

Chadha *et al.* [6] proposed a method to incorporate common-sense reasoning into the MDVC method. More specifically, they adapted common-sense reasoning from images [38] to videos, thus reaching impressive results in captioning – especially for the ground truth case. Although their proposal module uses the new feature to improve the bidirectional single-stream (Bi-SST) proposal generation method [37], we demonstrate that captioning results can be largely improved by replacing the proposal generation.

2.3. Unsupervised Representation Learning

As discussed earlier, state-of-the-art DVC methods employ a combination of multiple modalities of dense representations (e.g., video, audio and speech). In our proposal, we learn a dense representation from visual features in an unsupervised manner by encoding a new semantic information on the videos given by the visual similarities of short clips (clustering) and their co-occurrences (GloVE). This dense representation would replace audio and speech modalities in state-of-the-art DVC methods.

There are a few examples of unsupervised representation learning using clustering in the literature, with remarkable differences from ours. For instance, Xie *et al.* [42] introduced an end-to-end method to learn deep embeddings for cluster analysis. In their approach, a parameterized non-linear mapping is defined to generate a lower-dimensional feature space, where a clustering objective is adopted. Their method was evaluated on image and textual datasets with a few sets of labels (4 and 10) and does not fit our goals.

Another interesting method is DeepCluster, introduced by Caron *et al.* [4]. Their approach consists in alternating between clustering of the image descriptors and updating the weights of the convolutional network by predicting the cluster assignments. Similar to Xie *et al.* [42], they also employ k-means but perform a large-scale training of convolutional architectures, incorporating clustering in the architecture and objective. Finally, Hsu *et al.* [13] also proposed a method to address the problem of effectively grouping visual representations and jointly solve the problem of clustering and representation learning.

The main difference between our proposal and these methods relies on the fact that we employ unsupervised learning to predict soft labels on *short clips* and use these soft labels to generate *visual sentences* in which a Global Vectors (GloVE) method learns a dense representation for their co-occurrences. Therefore, our features are not optimized to predict a label for the clips but to describe the relationships between the clips. As mentioned earlier, state-of-the-art DVC methods take advantage of multi-modal learning. However, it is not easy to provide more modalities for these models for three main reasons: (i) the models will be prone to overfitting due to their increased capacity; (ii) different

modalities overfit and generalize at different rates, which requires multiple optimization strategies [41]; and (iii) more preprocessing is necessary to produce the features. We provide a relevant contribution by extracting more video information using only visual features without human annotations. Our method is an improved bag-of-word approach widely used in computer vision. However, it has not yet been applied to dense video captioning to the best of our knowledge. Additionally, this type of information (i.e., co-occurrence similarity) is not easily learned by deep learning techniques, especially in an unsupervised way, justifying our choice for the combination of k-means and GloVE.

3. Methodology

In this work, we propose a dense video captioning system that leverages unsupervised semantic information and is trained in two steps. In the first step, a temporal event proposal module is responsible for generating central points, event lengths, and confidence scores, predicting whether an event is contained in that location. This proposal generator is trained by adopting the architecture and procedures from Iashin and Rahtu [15], which are described in this section. Nevertheless, we replace the audio signal with the proposed semantic descriptor. Figure 2 shows the main elements of this step: a bi-modal transformer, a proposal generator, and a language generator.

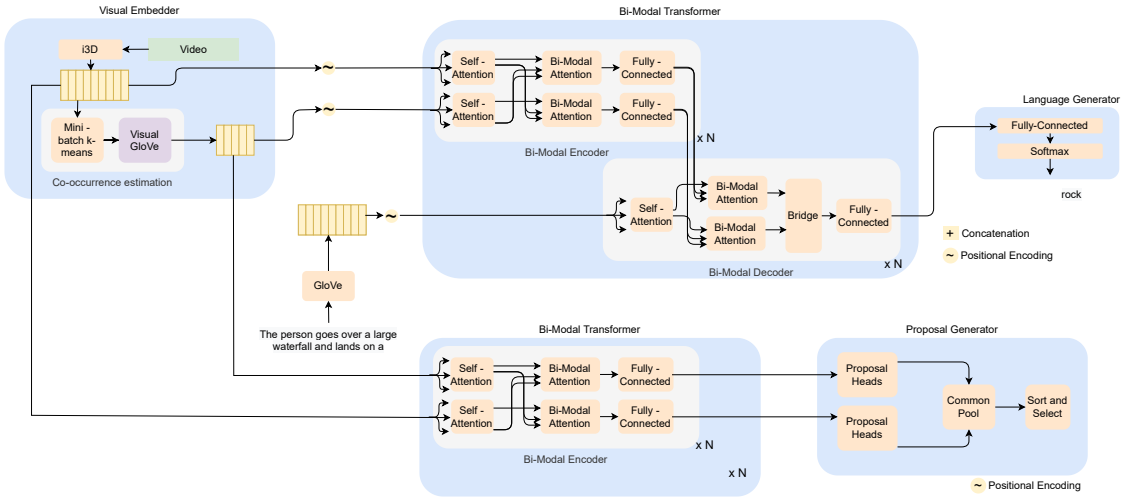
In the second step, we employ the vanilla transformer used by Iashin and Rahtu [16], replacing the Bi-SST proposal module with the Bi-Modal Transformer (BMT) proposal module due to their state-of-the-art performance in event proposal generation. The main elements in this step are a vanilla transformer and a language generator. In both of them, we employ the proposed semantic descriptor, shown in Figure 2, for the element co-occurrence estimation.

Bi-modal and vanilla transformers are composed of encoder and decoder layers. As vanilla is the base for the construction of bi-modal, we first explain how the captioning module works and then how the proposal generator works.

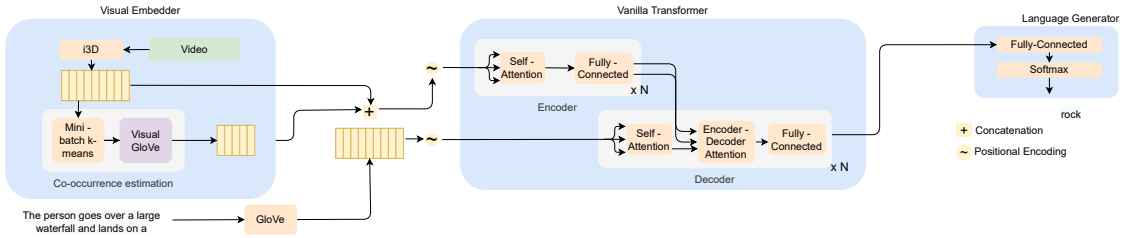
3.1. Co-Occurrence Similarity Estimation Module

Let $D_{Tr} = \{V_{Tr_1}, ..., V_{Tr|D_{Tr}|}\}$ and $D_{Te} = \{V_{Te_1}, ..., V_{Te|D_{Te}|}\}$ be the training and testing datasets, respectively, composed of videos with long duration (e.g., 1–2 min) and with more than one event per video. We first take all videos from D_{Tr} and split each one into short clips with f frames each. Then, we sample all these short clips and extract features using the i3D model [5]. As a result, a set of features $X = \{x_1, x_2, ..., x_l\}$, where $l = \lfloor n_f/f \rfloor$ and n_f is the number of frames of a given video, with $x \in \mathbb{R}^{1024}$ is produced per video. Next, a mini-batch k-means algorithm [29] is trained to minimize the Euclidean distance

$$\min \sum_{x \in X} ||Ecd(C, x) - x||^2 , \tag{1}$$



(a)



(b)

Figure 2: Overview of the proposed method. (a) describes the event proposal phase, which consists of two stages. In the first stage, a bi-modal transformer is employed in a captioning task, where visual and semantic co-occurrence-based features are used to learn the encoder parameters conditioned by language. Then, in the second stage, these encoder parameters are used to predict temporal event proposals. In the second stage (b), these proposals are used to generate captions using a vanilla transformer and a language generator trained with ground-truth events and sentences.

where $Ecd(C, x)$ stands for the nearest cluster center $c \in C$ to x and $|C|$ corresponds to our codebook size (e.g., 1,500 clusters).

Once we have trained the clustering model, a video can be processed by first splitting it into clips of f frames and then extracting the i3D features (using only the RGB stream) from these clips assigning each one of them to a cluster. These sequences of labeled clusters build a storytelling, and we can learn information about their co-occurrence properties, similarly to the dense representation from the GloVe method [27].

We compute a matrix of co-occurrence counts, denoted by Z, whose entries Z_{ij} tabulate the number of times the cluster j occurs in the context S (an arbitrary sliding window) of cluster i.

Let $Z_i = \sum_k Z_{ik}$ be the number of times any cluster appears in the context of cluster i, we define the co-occurrence probability as

$$P_{ij} = P(j|i) = \frac{Z_{ij}}{Z_i}. \tag{2}$$

Pennington *et al.* [27] showed that the vector learning should be with ratios of co-occurrence probabilities rather than with the probabilities themselves, as this choice forces a greater difference in values between clusters that occur close frequently compared to infrequent cases. This ratio can be computed considering three clusters i, j and k with (P_{ik}/P_{jk}) and the model takes the general form given by

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}, \tag{3}$$

where $w \in \mathbb{R}^{128}$ are cluster vectors and $\tilde{w} \in \mathbb{R}^{128}$ are separate context cluster vectors. Our model is a weighted least square regression trained with a cost function given by

$$J = f(Z_{ij}) \sum_{i,j=1}^{|C|} (w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log Z_{ij})^2, \tag{4}$$

where $|C|$ is the size of the vocabulary (i.e., 1,000 clusters), b_i and $\tilde{b}_j$ are bias vectors and f is a weighted function defined as

$$f(t) = \begin{cases} (t/t_{\max})^\alpha & \text{if } t < t_{\max} \\ 1 & \text{otherwise} \end{cases}, \tag{5}$$

where $t_{\max} = 100$ and $\alpha = 3/4$. More details and a complete mathematical description are provided in [27]. For our purposes, we adopt w as our semantic descriptor, represented as *Sm* in the remainder of this work.

3.2. Video Captioning Module

Given a video V, the video captioning module takes a set of n_c visual features $V_f = \{v_{f_1}, ..., v_{f_{n_c}}\}$, one per each clip, and a set of m words $Y = \{y_1, ..., y_m\}$ to estimate the conditional probability of an output sequence given an input sequence.

We encode v_{f_c}, where $1 \leq c \leq n_c$, as a concatenation of features defined as

$$v_{f_c} = [V_E(v_c), Sm(v_c)]\,, \tag{6}$$

where $V_E(\cdot)$ yields a deep representation given by an off-the-shelf neural network (e.g., i3D [5] with RGB or RGB + Optical Flow (OF) streams), $Sm(\cdot)$ produces our co-occurrence similarity representation (see Section 3.1), [ ] is a concatenation operator, and v_c is the c-th short clip for the video V.

The video features are fed to the original transformer model [33], composed of several layers (as shown in Figure 2), in which an encoder maps a sequence of visual features to a continuous representation that is used by a decoder to generate a sequence of symbols Y.

First, the visual embedding of each video is computed using Equation 6 and feeds all at once. Then, to provide information on the position of each feature we employ the same encoding method used by Vaswani *et al.* [33], a position-wise layer computes the position with sine and cosine at different frequencies as follows

$$\begin{aligned} PE_{(pos,2i)} &= \sin(pos/10000^{2i/d_{model}}), \\ PE_{(pos,2i+1)} &= \cos(pos/10000^{2i/d_{model}}), \end{aligned} \tag{7}$$

where pos is the position of the visual feature in the input sequence, $0 \leq i < d_{model}$ and d_{model} is a parameter defining the internal embedding dimension in the transformer.

In the encoder, these representations are passed through a multi-head attention layer. The attention used is the scaled dot-product and is defined in terms of queries (Q), keys (K), and values (V) as

$$Att(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \tag{8}$$

The multi-head attention layer is defined by the concatenation of several heads (1 to h) of attention applied to the input projections as

$$MHAtt(Q, K, V) = [head_1, ..., head_h]W^0\,, \tag{9}$$

where $head_i = Att(QW_i^Q, KW_i^K, VW_i^V)$ and [ ] is a concatenation operator.

Once we compute self-attention, $Q = K = V = V_f^{PE}$, which results in

$$\begin{aligned} V_f^{self-att} = [&Att(V_f^{PE}W_i^{V_f^{PE}}, V_f^{PE}W_i^{V_f^{PE}}, V_f^{PE}W_i^{V_f^{PE}}), \\ &..., Att(V_f^{PE}W_h^{V_f^{PE}}, V_f^{PE}W_h^{V_f^{PE}}, V_f^{PE}W_h^{V_f^{PE}})]. \end{aligned} \tag{10}$$

At the end of each encoder layer, a fully connected feed-forward network $FFN(\cdot)$ is applied to each position separately and identically. It consists of two linear transformations with a ReLU activation and is defined as

$$FFN(u) = \max(0, uW_1 + b_1)W_2 + b_2 \ , \tag{11}$$

resulting in V_f^{FFN} that is used in the decoder layer.

The decoder layer receives words and feeds an embedding layer $E(\cdot)$, computing a position with Equation 7 resulting in W^{PE}. Then, this representation is fed to the multi-head self-attention layer (see Equation 9), resulting in $W^{self-att}$. At this moment, the visual encoding provided by encoder layers feeds a multi-head attention layer as

$$W^{VisAtt} = MHAtt(W^{self-att}, V_f^{FFN}, V_f^{FFN}). \tag{12}$$

Finally, W^{VisAtt} feeds an $FFN(\cdot)$ and, then, a generator $G(\cdot)$ composed of a fully connected layer and a softmax layer is responsible for learning the predictions over the vocabulary distribution probability.

3.3. Event Proposal Module

The event proposal module uses the bi-modal transformer. Considering the encoder, this transformer has two differences from the vanilla encoder. It takes two streams, visual V_f and semantic Sm, separately, and it has three sub-layers in the encoder: self-attention (Equation 8), producing $V_f^{self-att}$ and $Sm^{self-att}$; bi-modal attention, i.e.,

$$V_f^{Sm-att} = MHAtt(V_f^{self}, Sm^{self}, Sm^{self}), \tag{13}$$

$$Sm^{Vis-att} = MHAtt(Sm^{self}, V_f^{self}, V_f^{self}), \tag{14}$$

and a fully connected layer $FFN(\cdot)$ for each modality attention, producing V_{Sm-att}^{FNN} and Sm_{v-att}^{FNN} used in the bi-modal attention unit on the decoder and in the multi-headed proposal generator.

In the bi-modal decoder, the differences to the vanilla decoder are the bi-modal attention and bridge layers. First,

a $W^{self-att}$ is obtained with Equation 9. Afterward, the bi-modal attention is computed as

$$W^{Sm-att} = MHAtt(W^{self-att}, Sm^{FNN}_{v-att}, Sm^{FNN}_{v-att}), \tag{15}$$

and

$$W^{V-att} = MHAtt(W^{self-att}, V^{FNN}_{Sm-att}, V^{FNN}_{Sm-att}). \tag{16}$$

The bridge is a fully connected layer on the concatenated output of bi-modal attention mechanisms, given as

$$W^{\text{FFN}} = \text{FFN}([W^{Sm-att}, W^{V-att}]). \tag{17}$$

The output of the bridge is passed through another FFN and then to the generator $G(\cdot)$. This means that the encoder parameters are learned in the captioning task, improving the visual features by conditioning them to the vocabulary.

More specifically, we focus on the Sm^{FNN}_{v-att} and V^{FNN}_{Sm-att} outputs. The proposal heads take these embeddings and make predictions for each modality individually, forming a pool of cross-modal predictions. The process begins with defining a Ψ set of anchors with a central location and a prior length. A fully connected model with three 1D convolutional layers (with kernels k_1 = arbitrary, $k_2 = k_3 = 1$) predicts the value for the length and confidence score for each anchor. Then, these predictions are grouped and sorted by their confidence levels, preserving the proportionality between the source modalities. The process of selecting a Ψ set of anchors follows the common approach of learning a k-means clustering model by grouping similar lengths using the ground-truth annotations [19, 37, 16, 6].

3.4. Training Procedure

The first stage is the training of the semantic descriptor. We split each video from the training set into clips with $f = 64$ frames and compute the i3D representation with only the RGB stream for each clip. Then, a mini-batch k-means learns a codebook with $|C| = 1500$ visual words in a procedure with 5 epochs. Once we have learned the clustering model, the semantic embedding is trained using a sliding window $S = 5$, corresponding to ≈ 10 seconds and cluster embedding vectors with 128 dimensions. The training occurs up to 1500 iterations, with early stopping applied after 100 iterations. The Adagrad optimizer [8] with learning rate $lr = 0.05$ is used.

The second stage is the training of the bi-modal encoder conditioned by the vocabulary. Thus, a captioning model is learned, using teaching forcing in which the target word is used as next input, instead of the predicted word, optimizing the *KL-divergence loss*, applying Label Smoothing [30] to make the model less confident over frequent words, and applying masking to prevent the model from attending on the next positions on the ground-truth sentences.

The model is learned up to 60 epochs with early stopping to monitor the METEOR score [3], using the Adam optimizer [18] with $\beta_1 = 0.9, \beta_2 = 0.999$, $lr = 5.10^{-5}$ and $\epsilon = 1.10^{-8}$. These procedures are also adopted in the final captioning training (i.e., using the vanilla transformer).

The bi-modal encoder is used to learn the multi-head proposal module with Mean Squared Error (MSE) for localization losses and cross-entropy for confidence losses. Then, we learn the final captioning model feeding the vanilla transformer [33] with ground-truth proposals and sentences. Thus, we predict the sentences to evaluate the performance.

All experiments were conducted on a computer equipped with an AMD Ryzen 7 2700X 3.7GHz CPU, 64 GB of RAM, and an NVIDIA Titan Xp GPU (12 GB). The experiments were performed using the Ubuntu operating system.

4. Dataset and Evaluation Metrics

All experiments were performed on the ActivityNet Captions dataset [19], which is a large-scale dataset with temporal segments annotated and described in the proportion of one sentence for each segment. ActivityNet Captions was selected because it is a challenging open-domain dataset used as a default evaluation by all reference works. The dataset contains 20,000 videos divided into training/validation/test subsets with 50/25/25% videos, respectively, and 3.65 events per video on average. As the annotations of the test set are not public, we used the validation set for testing, as in previous works [37, 6, 16, 15].

The validation set was annotated twice (val1 and val2), and we consider the average for each evaluation metric on each validation split. The captioning task was evaluated using the BLEU@1-4 [26], METEOR [3], ROUGE$_L$ [22] and CIDEr-D [34] metrics computed with the evaluation script provided by Krishna *et al.* [19], whereas event proposal was evaluated with Precision, Recall and F1-score (i.e., the harmonic mean of precision and recall).

5. Results

This section discusses our results on event proposal generation and video captioning. We provide a comparison with state-of-the-art (SOTA) methods, a qualitative analysis, and ablation studies.

5.1. Results for Event Proposals

As described in Section 3.4, we explored the captioning training to learn the parameters of the bi-modal encoder and then used this encoder to predict the proposals in the bi-modal proposal generator module. Afterward, these proposals were employed in a vanilla transformer captioning model.

Table 1 shows the BMT performance on video captioning. We highlighted as baselines the results from Iashin and Rahtu [15] with only visual features (i.e., using a vanilla transformer) and with bi-modal features (i.e., using visual and audio features denoted by BMT). Additionally, we included the performance from Iashin and Rahtu [16] with visual, audio, and speech modalities and employing Bi-SST as the event proposal module. Lastly, we investigated how BMT captioning performs with *RGB+Sm* and with *V+Sm* (i.e., *RGB* + *OptFlow* + *Sm*).

Table 1: Captioning performance comparison of the BMT and Transformer methods with different features in the same validation sets. For each metric, the top 2 results are highlighted in bold.

	GT Proposals			Learned Proposals		
	B@3	B@4	M	B@3	B@4	M
Visual [15]	3.77	1.66	10.29	2.85	1.30	7.47
BMT [15]	**4.62**	**1.99**	**10.89**	**3.84**	**1.88**	**8.44**
$\text{MDVC}_{Bi\text{-}SST}$	**4.52**	**1.98**	**11.07**	2.53	1.01	7.46
BMT_{RGB+Sm}	4.12	1.72	10.32	3.62	1.74	8.03
BMT_{V+Sm}	4.32	1.85	10.55	**3.68**	**1.81**	**8.26**

Our results with *V+Sm* presented superior performance compared to the Visual performance from [15] considering all metrics and proposals schemes (GT and Learned). Comparing BMT with BMT_{V+Sm}, we observed a slightly lower performance using *Sm* instead *A* (audio) considering all scores.

However, the proposed model was still capable of learning a high-quality encoder, as evidenced by the performance levels achieved on proposal generation (see Table 2). We reached competitive results in terms of F1-score and Precision compared to the original BMT in the *V+Sm* scenario. This slight difference in F1-score supports the adoption of only visual features for event proposal generation due to the fewer preprocessing requirements than BMT. Considering the performance on *RGB+Sm* configuration (i.e., even less preprocessing), we outperformed the popular Bi-SST method while achieving competitive performance with BMT. Masked transformer [46], which is a method that explores only visual features and linguistic information to learn temporal proposals, is outperformed by our approach in 11.8% [i.e., 59.60/53.31] in terms of F1-score.

Table 2: Comparison with state-of-the-art proposal generation. Results are reported on the validation sets using Precision, Recall and F1-score and are taken for 100 proposals per video ratio. For each metric, the top 2 results are highlighted in bold.

	FD	Prec.	Rec.	F1
MFT [43]	✓	51.41	24.31	33.01
BiSST [37]	✓	44.80	57.60	50.40
Masked Transf. [46]	✓	38.57	**86.33**	53.31
SDVC [24]	✓	**57.57**	55.58	56.56
BMT_{V+A} [15]	✗	48.23	**80.31**	**60.27**
PDVC [40]	✗	**58.07**	55.42	56.71
Ours_{RGB+Sm}	✗	47.27	78.71	59.07
Ours_{V+Sm}	✗	48.11	78.31	**59.60**

5.2. *Results for the Video Captioning Stage*

In Table 3, we show a comparison between our results using the Vanilla Transformer, fed by the concatenation of visual and semantic descriptors, and the results obtained by SOTA methods. As can be seen, there are methods based only on visual features and methods based on multi-modal features (see column *VF*). As the videos from ActivityNet captions must be downloaded from YouTube, several videos have become unavailable since the original dataset was published. Hence, we used 91% of the dataset (this information is presented in column FD, where a "✓" means that 100% of the videos were available at the time of the experiments). As we have a reduced set of videos for evaluation, the validation sets were filtered to contain only the videos downloaded. As demonstrated in [15], this procedure enables a fair comparison because the *SOTA methods reached almost unchanged results* when evaluated using these filtered validation sets. However, not considering this procedure is unfair, because the model is forced to propose events and generate captions for unseen videos, reducing performance. Finally, some works adopted a direct optimization of the METEOR score with reinforcement learning techniques (see column RL). We also listed the performance without these techniques since, as shown in Table 3 for DVC [21], these techniques boosted the METEOR score without a proportional boost in BLEU, which may not corresponds to an actual improvement in the captioning quality.

Table 3: Comparison with other methods on ActivityNet Captions (validation set). VF = Visual features only; RL = Reinforcement Learning – reward maximization (METEOR); FD = Full dataset was available. The top 2 results are highlighted in bold.

	VF	RL	FD	GT Proposals			Learned Proposals		
				B@3	B@4	M	B@3	B@4	M
DVC [21]	✓	✓	✓	4.55	1.62	10.33	2.27	0.73	6.93
SDVC [24]	✓	✓	✓	4.41	1.28	13.07	2.94	0.93	**8.82**
GVL [39]	✓	✓	✗	–	–	–	–	1.11	10.03
Dense Cap [19]	✓	✗	✓	4.09	1.60	8.88	1.90	0.71	5.69
DVC [21]	✓	✗	✓	4.51	1.71	9.31	2.05	0.74	6.14
Masked Transf. [46]	✓	✗	✓	5.76	2.71	11.16	2.91	1.44	6.91
Bi-SST [37]	✓	✗	✓	–	–	10.89	2.27	1.13	6.10
SDVC [24]	✓	✗	✓	–	–	–	–	–	6.92
MMWS [28]	✗	✗	✗	3.04	1.46	7.23	1.85	0.90	4.93
BMT [15]	✗	✗	✗	4.63	1.99	10.90	3.84	1.88	8.44
iPerceive [6]	✗	✗	✗	**6.13**	**2.98**	**12.27**	2.93	1.29	7.87
MDVC [16]	✗	✗	✗	**5.83**	**2.86**	**11.72**	2.60	1.07	7.31
TSP [2]	✗	✗	✗	–	–	–	4.16	2.02	**8.75**
PDVC [40]	✗	✗	✗	–	3.12	11.26	–	1.96	8.08
GVL [39]	✓	✗	✗	–	–	–	–	2.18	8.50
Ours$_{RGB+Sm}$	✓	✗	✗	5.40	2.55	11.06	**4.37**	**2.42**	8.52
Ours$_{V+Sm}$	✓	✗	✗	5.54	2.64	11.23	**4.57**	**2.55**	8.65

Considering only the single modality scenario, without *RL*, our model outperforms all other methods in learned proposals and has a slightly lower performance on BLEU@3-4 than the Masked Transformer for GT proposals. Compared to the multi-modal methods, our performance on ground truth is lower than the MDVC and iPerceive

methods. However, we remark that the performance on GT proposals is an indicator of how good the captions are when the event is perfectly delimited. As can be seen in Table 2, this ideal scenario is far from being the case, and the most significant performance to consider is in the learned proposals scenario, where our results are particularly noteworthy.

Finally, we highlight the results of the TSP method [2] compared to ours. This method includes an improved visual descriptor for temporal event localization that combines local features optimized by accuracy on trimmed action classification (TAC) and global features given by pooling local predictions. The authors employed the R(2+1)D architecture [32] fine-tuned on the ActivityNet v1.3 dataset [11]. They adopted the BMT model for captioning, and a critical procedure for the success of video captioning was the fine-tuning on ActivityNet with trimmed action annotations (METEOR of 8.75 with fine-tuning and 8.42 without [2]). Lastly, their model also considers the audio signal. Thus, it is noteworthy that our model reaches a comparable performance on METEOR without audio and action annotations.

5.3. Qualitative Analysis

In Figure 3, we show a qualitative analysis of dense captioning considering our semantic descriptor in combination with audio and speech. The model we propose is denoted as configuration *V+Sm*. In Section 5.4.2, we provide the quantitative results of these combinations. We selected four videos in different contexts to evaluate the ability to describe general events. The selected videos are svWiQtzgtOc, P4PQ5tC3gX8, BhAQhPasmhU, and 045Tkq12H_c.

Upon evaluating the examples in Figure 3a, it can be observed that Event 1 posed a challenge for models *V+Sm* and *V+A+Sm*, as they confused street dance with a gymnastics routine. Indeed, the movement is very similar to the one performed on the pommel horse. In Event 2, only *V+A+Sm* and *V+A+S+Sm* managed to correctly identify it as a dance. *V+Sm* described the move exactly, but even so, it would be a description with low scores B@3-4 or M. When analyzing the sentences in Figure 3b, for Event 1, none of the models were able to describe the various accordions throughout the video, although all reported that there is an accordion. This highlights the challenges in recognizing information with long-range dependencies. As for Figure 3b, Event 2, the models failed to predict the person standing.

In Figure 3c, Event 1, only V+S+Sm correctly described that the object being thrown was a discus, while all models correctly described the spinning and throwing motion. In Event 2, no model described the woman in blue. This indicates that the models tend to prioritize the foreground objects. Nevertheless, a less observant person would likely overlook the woman as well. Similarly, in Event 3, none of the models described the boy. Finally, in Figure 3d, the models correctly identify what is happening, but they generated sentences with syntactic problems, notably the repetition of fragments. In general, there was no significant improvement in sentence quality when incorporating additional modalities such as *A* or *S*, reinforcing the potential of our semantic descriptor combined with the *V* modality.

GT
A man is seen performing various flips and tricks in the middle of a public square with dozens of people watching on the side.
V+Sm
A man is seen performing a gymnastics routine on a stage and performing martial arts moves.
V+A+Sm
A man is seen performing a routine on a stage followed by a man performing a routine.
V+S+Sm
A man is seen performing a routine on the floor and performing various tricks and tricks.
V+A+S+Sm
A man is seen performing various tricks on a stage and begins performing a routine.

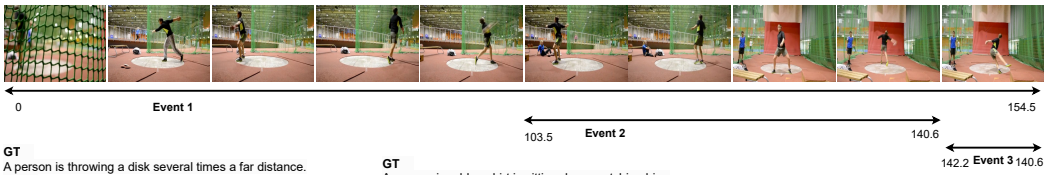
GT
The man continues to do tricks in various areas while the people around him watch and applaud.
V+Sm
The man does a handstand and spins around and does a spin routine.
V+A+Sm
The man continues to dance around while the camera captures his movements.
V+S+Sm
The man then begins performing various martial arts moves and ends with him jumping and doing tricks.
V+A+S+Sm
The man continues to dance and ends with him jumping and flipping around.

(a)

GT
The man place a small hand held accordion.
V+Sm
A person is seen sitting in a chair holding an accordion.
V+A+Sm
A man is seen sitting in front of a piano and playing an accordion.
V+S+Sm
A man is seen sitting in a chair playing an accordion.
V+A+S+Sm
A man is seen sitting in a chair playing an accordion.

GT
A man plays a song on various accordions in an office.
V+Sm
A man is seen sitting in a chair playing an accordion.
V+A+Sm
A man is seen sitting in front of a camera and leads into him playing an accordion.
V+S+Sm
A man is seen playing an accordion while moving his hands up and down the accordion.
V+A+S+Sm
A man is seen sitting in front of a piano and playing an accordion.

(b)



GT
A person is throwing a disk several times a far distance.
V+Sm
The man spins and throws the hammer in circles and throws the discus.
V+A+Sm
The man spins around and throws the ball.
V+S+Sm
The man spins and throws the hammer throw in the air and throws the discus.
V+A+S+Sm
The man then throws the object around the circle and throws the object off into the distance.

GT
A woman in a blue shirt is sitting down watching him.
V+Sm
The man spins the hammer throw in circles and throws the discus.
V+A+Sm
The man spins around and throws the discus.
V+S+Sm
A man is seen standing in a circle and throwing a discuss off into the distance.
V+A+S+Sm
A man in a white shirt throws a disc in a circle.

GT
A boy in a blue shirt is standing next to him.
V+Sm
The man spins the hammer throw in circles and throws the discus.
V+A+Sm
The man spins and throws the discuss.
V+S+Sm
The man spins and throws the ball.
V+A+S+Sm
The man spins around and throws the ball.

(c)

GT
A boy with a microphone is interviewing a man in sunglasses.
V+Sm
A man is seen speaking to the camera and leads into a man speaking to the camera.
V+A+Sm
A man is seen speaking to the camera while standing on a rock wall.
V+S+Sm
A man is seen speaking to the camera and leads into a man speaking to the camera.
V+A+S+Sm
A man is seen speaking to the camera and leads into him climbing up a rock wall.

GT
They talk back and forth in front of a large boulder.
V+Sm
The man continues climbing on the rock wall and ends with the camera panning around the area.
V+A+Sm
A man is seen climbing up a rock wall and climbing up a rock wall.
V+S+Sm
The man continues climbing and the rock and the camera pans around the area.
V+A+S+Sm
A man is seen standing on a rock and climbing wall and leads into several shots of people riding on the side.

GT
Men are then shown rappelling down the side of the cliff as they talk.
V+Sm
A man is seen speaking to the camera and leads into a man climbing a rock wall.
V+A+Sm
A man is seen speaking to the camera and leads into a man climbing up a rock wall.
V+S+Sm
The men continue to ride around the area around the area and end by stopping to speak to the camera.
V+A+S+Sm
A man is seen speaking to the camera and leads into a man climbing up a rock wall.

(d)

Figure 3: Qualitative comparison on the results including our semantic descriptor in the MDVC method. We show results for event proposals captioning considering the highest Intersection over Union for each ground truth event proposal in the following videos (a) svWiQtzgtOc, (b) P4PQ5tC3gX8, (c) BhAQhPasmhU, and (d) 045Tkq12H_c.

5.4. Ablation Studies

In this section, we conduct ablation studies to assess two key aspects: (i) the performance of the semantic descriptor under different configurations, involving changes to the vocabulary and context window, and (ii) the impact of our proposed descriptor on the MDVC method, i.e., combining it with multiple modalities.

5.4.1. Semantic Descriptor Evaluation

We evaluated the semantic descriptor for its performance on the BMT model replacing the audio signal and considering ground truth and captioning of event proposals, Figure 4 (a) and (b), and for performance of event proposals (c). We changed the vocabulary size $|C|$ (100, 500, 1000, 1500, and 2000) and the context window S (10s, 30s, and 60s).

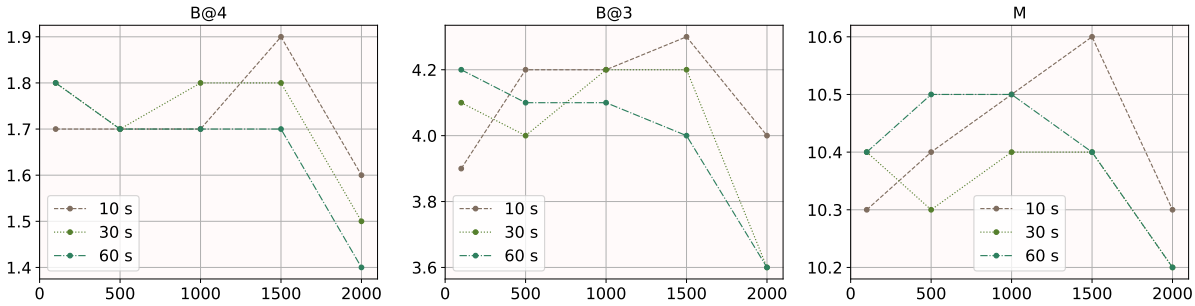
Upon evaluating the results presented in Figure 4, it is possible to note that small vocabularies (i.e., those with 100 and 500 visual words) produce inferior results than larger vocabulary sizes for all metrics (B@3, B@4, and M). The best results were achieved using vocabularies with 1,000 or 1,500 visual words. Surprisingly, the vocabulary comprising 2,000 visual words suffers a significant drop in performance for all context window sizes. We can also observe that the context window size significantly impacts the performance. Long windows (i.e., 60s) usually produce worse results than smaller windows with 30s or 10s. The optimal results were achieved using 10s.

The same conclusions are reached by analyzing the results of event generation performance, as shown in Figure 4 (c). The best results are observed when using either a 30s window with 1,000 visual words or a 10s window with 1,500 visual words. Thus, in the experiments conducted in this paper, the configuration with 1,500 visual words and 10s of context window was selected due to its superior performance across both tasks.

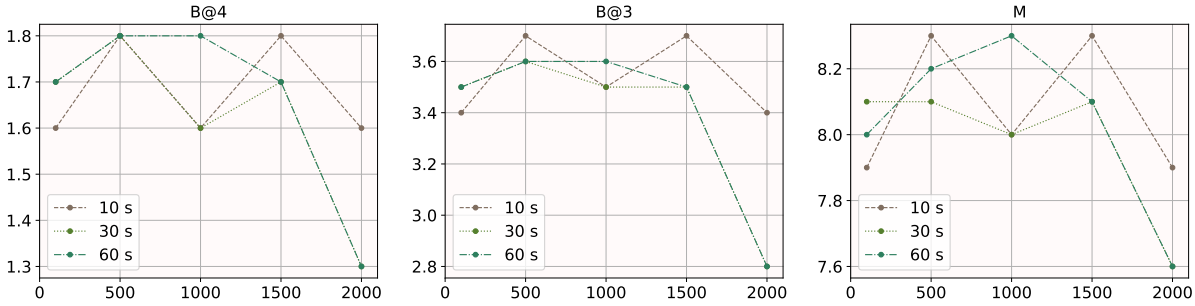
5.4.2. Impact of co-occurrence similarity estimation on the MDVC method

Motivated by the event proposal performance, we adopt the same features and validation sets from BMT in both our method and MDVC baseline. This enables a fair comparison with MDVC and iPerceive [6] SOTA methods, as there are a few differences between the filtered validation sets used to evaluate BMT and MDVC. There are also differences in the number of frames used to extract visual features with the i3D method (24 frames [16] $\times$ 64 frames in our experiments).

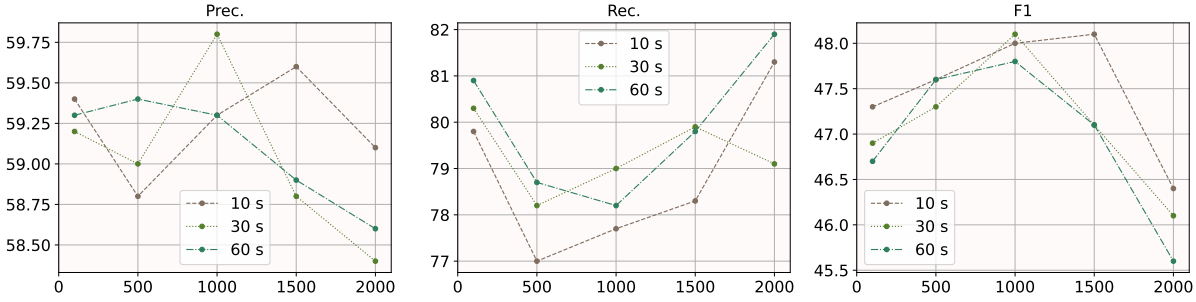
Table 4 shows the results of MDVC with the same i3D features as BMT and with our temporal proposals using *V (#1)*, *V+A (#2)*, *V+S (#3)* and *V+A+S (#4)*, where *V* = i3D output for *RGB* and OF streams, *A* = audio, *S* = speech, and *Sm* = co-occurrence similarity. We evaluate the same configurations including the *Sm* descriptor (*#6*, *#7*, *#8*, and *#9*), and an additional scenario in which only RGB stream is employed *(RGB+Sm) (#5)*. In the most challenging scenario, learned proposals, *V+A+Sm* and *V+Sm* outperformed any configuration without *Sm* considering M and B@3-4. Notably, our performance with *RGB+Sm* in learned proposals is competitive with the multi-modal approach.



(a) Captioning performance on ground truth proposals



(b) Captioning performance on event proposals



(c) Performance of event proposals

Figure 4: Captioning performance levels for ground truth and event proposals, (a) and (b), and for proposal generation (c) considering the BMT model replacing the audio signal with several semantic descriptors generated with different vocabularies (100, 500, 1000, 1500, 2000) and context windows (10s, 30s, 60s).

Table 4: Results on the ActivityNet Captions dataset [19] adopting the MDVC method and the same validation sets used in iPerceive [6]. V = i3D output for *RGB* and Optical Flow (*OF*) streams; A = audio; S = speech; Sm = co-occurrence similarity; B = BLEU@N; M = METEOR; R = Rouge$_l$; and C = CIDEr-D. The top 2 results for each metric are highlighted in bold.

#	V		A	S	Sm	GT Proposals					Learned Proposals				
	RGB	OF				B@3	B@4	M	R	C	B@3	B@4	M	R	C
1	✓	✓				5.40	2.67	11.18	22.90	44.49	4.40	2.46	8.58	13.36	**13.03**
2	✓	✓	✓			**5.67**	**2.75**	**11.37**	**23.69**	**46.19**	4.49	2.50	8.62	13.49	**13.48**
3	✓	✓		✓		**5.78**	**2.95**	10.87	22.87	41.40	4.21	2.33	8.43	13.34	11.79
4	✓	✓	✓	✓		5.61	2.69	**11.49**	**23.82**	**46.29**	4.41	2.31	8.50	13.47	13.09
5	✓				✓	5.40	2.55	11.06	23.01	42.53	4.37	2.42	8.52	13.40	12.14
6	✓	✓			✓	5.54	2.64	11.23	23.34	45.76	**4.57**	**2.55**	**8.65**	**13.62**	12.82
7	✓	✓	✓		✓	5.61	2.71	11.25	23.61	45.09	**4.60**	**2.58**	**8.69**	**13.87**	12.99
8	✓	✓		✓	✓	5.18	2.52	10.95	22.77	43.91	3.89	2.07	8.13	12.78	11.00
9	✓	✓	✓	✓	✓	5.53	2.62	11.17	23.43	45.93	4.10	2.20	8.30	13.07	11.64

It is evident that audio and speech exerted a more substantial influence on the ground truth proposals than on learned proposals results. In the case of ground truth proposals, V+Sm performs better than V (M and B@3-4). However, in all other scenarios, a slightly inferior performance is observed, possibly due to the difficulty of multi-modal training concerning the different times for optimization of each modality during the training process.

6. Conclusions and Future Work

In this work, we presented a method to enrich visual features for dense video captioning that learns visual similarities between clips from different videos and extracts information on their co-occurrence probabilities. Our conclusions are: (i) co-occurrence similarities combined with deep features can provide more meaningful semantic information for dense video captioning than only deep features from a single modality; (ii) our semantic features processed with an encoder-decoder scheme based on transformers outperformed single modality methods while achieving competitive results with multi-modal state-of-the-art methods; and (iii) we reached impressive results adopting only the RGB stream when compared to results using RGB, optical flow, and audio information.

As directions for future work, deep clustering methods could replace the mini-batch k-means. As our method is unsupervised, multiple large-scale visual datasets could be combined without the need for linguistic descriptions or human annotations. These datasets could be used to learn more accurate/detailed codebooks using co-occurrences or BERT-based models. Additionally, conducting experiments on diverse domains, such as cooking activities [46] or sport events [45], could further enrich the scope of the investigation.

Acknowledgments

This work was supported by the Federal Institute of Paraná, Federal University of Paraná, and by grants from the National Council for Scientific and Technological Development (CNPq) (grant numbers 308879/2020-1,

304836/2022-2, and 315409/2023-1). The Titan Xp GPU used for this research was donated by NVIDIA.

References

[1] Aljalbout, E., Golkov, V., Siddiqui, Y., Cremers, D., 2018. Clustering with deep learning: Taxonomy and new methods. arXiv preprint arXiv:1801.07648, 1–12.

[2] Alwassel, H., Giancola, S., Ghanem, B., 2021. TSP: Temporally-sensitive pretraining of video encoders for localization tasks, in: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3166–3176.

[3] Banerjee, S., Lavie, A., 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72.

[4] Caron, M., Bojanowski, P., Joulin, A., Douze, M., 2018. Deep clustering for unsupervised learning of visual features, in: European Conference on Computer Vision (ECCV), pp. 1–18.

[5] Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733.

[6] Chadha, A., Arora, G., Kaloty, N., 2021. iPerceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. IEEE Winter Conference on Applications of Computer Vision (WACV) , 1–13.

[7] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2625–2634.

[8] Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research 12, 2121–2159.

[9] Escorcia, V., Heilbron, F.C., Niebles, J.C., Ghanem, B., 2016. Daps: Deep action proposals for action understanding, in: European Conference on Computer Vision (ECCV), pp. 768–784.

[10] Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., Deng, L., 2017. Semantic compositional networks for visual captioning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1141–1150.

[11] Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C., 2015. ActivityNet: A large-scale video benchmark for human activity understanding, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 961–970.

[12] Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., 2017. CNN architectures for large-scale audio classification, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131–135.

[13] Hsu, C.C., Lin, C.W., 2018. CNN-based joint clustering and representation learning with feature drift compensation for large-scale image data. IEEE Transactions on Multimedia 20, 421–429.

[14] Huang, J., Gong, S., Zhu, X., 2020. Deep semantic clustering by partition confidence maximisation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8846–8855.

[15] Iashin, V., Rahtu, E., 2020. A better use of audio-visual cues: Dense video captioning with bi-modal transformer, in: British Machine Vision Conference (BMVC), pp. 1–16.

[16] Iashin, V., Rahtu, E., 2020. Multi-modal dense video captioning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4117–4126.

[17] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1725–1732.

[18] Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization, in: Bengio, Y., LeCun, Y. (Eds.), International Conference on Learning Representations (ICRL), pp. 1–15.

[19] Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C., 2017. Dense-captioning events in videos, in: International Conference on Computer Vision (ICCV), pp. 706–715.

[20] Li, S., Li, W., Cook, C., Zhu, C., Gao, Y., 2018. Independently recurrent neural network (IndRNN): Building a longer and deeper RNN, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5457–5466.

[21] Li, Y., Yao, T., Pan, Y., Chao, H., Mei, T., 2018. Jointly localizing and describing events for dense video captioning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7492–7500.

[22] Lin, C., 2004. ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain. pp. 74–81.

[23] van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2579–2605.

[24] Mun, J., Yang, L., Ren, Z., Xu, N., Han, B., 2019. Streamlined dense video captioning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6588–6597.

[25] Pan, Y., Yao, T., Li, H., Mei, T., 2017. Video captioning with transferred semantic attributes, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 984–992.

[26] Papineni, K., Roukos, S., Ward, T., Zhu, W., 2002. BLEU: a method for automatic evaluation of machine translation, in: Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311–318.

[27] Pennington, J., Socher, R., Manning, C.D., 2014. GloVe: Global vectors for word representation, in: Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

[28] Rahman, T., Xu, B., Sigal, L., 2019. Watch, listen and tell: Multi-modal weakly supervised dense event captioning, in: IEEE International Conference on Computer Vision (ICCV), pp. 8907–8916.

[29] Sculley, D., 2010. Web-scale *k*-means clustering, in: International Conference on World Wide Web, p. 1177–1178.

[30] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826.

[31] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3D convolutional networks, in: IEEE International Conference on Computer Vision (ICCV), pp. 4489–4497.

[32] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M., 2018. A closer look at spatiotemporal convolutions for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6450–6459.

[33] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: International Conference on Neural Information Processing, pp. 6000–6010.

[34] Vedantam, R., Zitnick, C.L., Parikh, D., 2015. CIDEr: consensus-based image description evaluation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4566–4575.

[35] Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K., 2015. Sequence to sequence – video to text, in: IEEE International Conference on Computer Vision (ICCV), pp. 4534–4542.

[36] Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K., 2015. Translating videos to natural language using deep recurrent neural networks, in: Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015), pp. 1494–1504.

[37] Wang, J., Jiang, W., Liu, W., Xu, Y., 2018. Bidirectional attentive fusion with context gating for dense video captioning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7190–7198.

[38] Wang, T., Huang, J., Zhang, H., Sun, Q., 2020a. Visual commonsense R-CNN, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10757–10767.

[39] Wang, T., Zhang, J., Zheng, F., Jiang, W., Cheng, R., Luo, P., 2023. Learning grounded vision-language representation for versatile understanding in untrimmed videos. ArXiv abs/2303.06378.

[40] Wang, T., Zhang, R., Lu, Z., Zheng, F., Cheng, R., Luo, P., 2021. End-to-end dense video captioning with parallel decoding, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6827–6837.

[41] Wang, W., Tran, D., Feiszli, M., 2020b. What makes training multi-modal classification networks hard?, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12692–12702.

[42] Xie, J., Girshick, R., Farhadi, A., 2016. Unsupervised deep embedding for clustering analysis, in: International Conference on Machine Learning (ICML), p. 478–487.

[43] Xiong, Y., Dai, B., Lin, D., 2018. Move forward and tell: A progressive generator of video descriptions, in: European Conference on Computer Vision (ECCV), pp. 468–483.

[44] Xu, H., Li, B., Ramanishka, V., Sigal, L., Saenko, K., 2019. Joint event detection and description in continuous video streams, in: IEEE Winter Applications of Computer Vision Workshops (WACV), pp. 25–26.

[45] Yu, H., Cheng, S., Ni, B., Wang, M., Zhang, J., Yang, X., 2018. Fine-grained video captioning for sports narrative, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6006–6015.

[46] Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C., 2018. End-to-end dense video captioning with masked transformer, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8739–8748.