SEM-POS: Grammatically and Semantically Correct Video Captioning

Asmar Nadeem\textsuperscript{1}, Adrian Hilton\textsuperscript{1}, Robert Dawes\textsuperscript{2}, Graham Thomas\textsuperscript{2}, Armin Mustafa\textsuperscript{1}

Abstract

Generating grammatically and semantically correct captions in video captioning is a challenging task. The captions generated from the existing methods are either word-by-word that do not align with grammatical structure or miss key information from the input videos. To address these issues, we introduce a novel global-local fusion network, with a Global-Local Fusion Block (GLFB) that encodes and fuses features from different parts of speech (POS) components with visual-spatial features. We use novel combinations of different POS components - 'determinant + subject', 'auxiliary verb', 'verb', and 'determinant + object' for supervision of the POS blocks - Det + Subject, Aux Verb, Verb, and Det + Object respectively. The novel global-local fusion network together with POS blocks helps align the visual features with language description to generate grammatically and semantically correct captions. Extensive qualitative and quantitative experiments on benchmark MSVD and MSRVTT datasets demonstrate that the proposed approach generates more grammatically and semantically correct captions compared to the existing methods, achieving the new state-of-the-art. Ablations on the POS blocks and the GLFB demonstrate the impact of the contributions on the proposed method.

1. Introduction

Video captioning is a challenging task as it aims to describe the contents of a video in a natural language like humans do [50]. This finds application in media production [3], visual retrieval [19, 26, 43, 53, 62], visual question answering [4], etc. Existing methods for video captioning can broadly be classified into two categories: (1) learning visual representation for the whole caption in an end-to-end manner leaving the fine-grained local features unaddressed [1, 34, 36, 38], which generates captions with key information missing in the description; and (2) incorporating

\textsuperscript{1}Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, United Kingdom. \textsuperscript{2}BBC Research and Development, United Kingdom. Correspondence to: Asmar Nadeem <asmar.nadeem@surrey.ac.uk>.
of actions and objects, generating word-by-word descriptions. [41] attempts to explore the local semantics in the input video, however, it does not align the visual and language semantics, leading to limited performance. In this paper, in addition to the visual features from the POS blocks, we also input language semantics (supervised by ‘determinant + subject’, ‘auxiliary verb’, ‘verb’, and ‘determinant + object’) to the GLFB block, which aligns visual semantics with linguistics and generating grammatically correct captions. The input, output, and pipeline from the proposed network are shown in Figure 1. We also mask input spatial as well as temporal features in end-to-end training which also results in an improvement in the performance. The proposed method is inspired by the previous work POS-CG [51] and SAAT [64]. However, these methods do not fuse language and visual information and they assume the subjects as well as the objects to be physical entities and the verbs to be motion-based actions which may not be true semantically. This paper aims to answer three main questions previously unaddressed by state-of-the-art methods: 1) How to align visual and linguistic semantics to generate grammatically correct captions? (2) How to add fine-grained local features to increase semantic correctness? and (3) How to fuse local and global features together for simultaneous grammatical and semantic correctness? To answer these questions, we have proposed the following novel contributions: • A novel global-local fusion network with GLFB that encodes and fuses fine-grained local visual features from different POS blocks with global caption language features for semantically correct captions. • We propose four POS blocks: Det+Subject, Aux Verb, Verb and Det+Object with spatial and temporal features as input. Comprehensive ablation studies on the selection of POS blocks and input features show the generation of captions closer to the natural language. • An extensive evaluation of the proposed method against state-of-the-art on two benchmark datasets, MSVD [8] and MSRVT [55], on five metrics, demonstrates improved grammatical and semantic caption generation.

2. Related Work

This section reviews video captioning techniques, and visual and linguistic features fusion to support the contributions of the paper.

2.1. Video Captioning

Video representation: Video captioning [2, 50] is a challenging task in video representation. Earlier, CNN-based methods [42,47] have been applied to learn video representations which lack to model the long-range dependencies but some [16, 46] are still popular for their spatial and temporal representations of the videos. These video features, in a global fashion, combined with RNN [9, 11, 18, 32, 59] and LSTM [12,45], have been employed for the video captioning task but visual-linguistic alignment and fine-grained features are not explored. Recently, multi-modal methods [44,56] have become popular using text and audio information for video understanding tasks which result in text or audio-dependent models lacking visual information. [15] uses a cross-attention mechanism to attend to different modalities and other, visions only, methods [45,61] use attention mechanisms for extracting linguistic semantics, globally, from the visual features. The proposed method in this paper exploits the power of LSTMs, inspired by [12], for aligning and fusing visual and linguistic semantics. We also employ an attention mechanism [61] to make the linguistic components focus on the relevant visual features for video captioning.

Transformer for Captioning: Transformer [20, 48] is increasingly being used for object detection [7], classification [14], scene understanding [63] and image captioning [30] tasks. It gained its popularity from the natural language-based models like BERT [10] and roBERTa [31], a version of BERT [10], where it demonstrates significant improvement in the performance when using a huge corpus of web text data. These methods are recently adopted for vision-based tasks while using a billion images to give a state-of-the-art performance. Our method generates more grammatically correct captions than these methods (see Section 4.1). [28, 29, 40] are not fair to be compared with as [28] has used an ensemble model together with a transformer decoder and rest have a pre-training stage for their extremely large transformer-based models using ten to hundred times more computing than ours.

2.2. Masking Techniques

Previous methods [17, 25, 54] use masking on image frames to generalize the input to the model which improves the performance of the model. Pre-training tasks for language transformers [10,31] require language masking, and similarly, image frames are masked to recover the masked portion of the image. To the best of our knowledge, we are the first to mask spatial features as well as temporal features for an end-to-end video captioning task, which results in an improvement in the performance (see Section 4.2).

2.3. Visual and Linguistic Alignment and Fusion

Existing methods [33,51,60,64] generate captions with POS in the decoder block without using separate POS blocks. We are the first to explore separate POS blocks in the encoder part to align visual semantics with linguistic POS components in a supervised manner. [21] predicts the label of the POS for every word as part of the caption generation using a softmax layer, without any POS blocks. The
proposed method aligns the visual and language components at both fine-grained local and global levels by proposing to use of four POS blocks and GLFB.

### 3. Method

This section gives an overview of the proposed method, which is followed by the problem formulation in Section 3.1 and a detailed explanation for the POS blocks in the subsequent sections.
puts to the LSTM-based POS blocks using different masking mechanisms. We have introduced random spatial features masking and chunk-wise temporal features masking as Visual Features Masking (VFM) in Section 3.2 for an end-to-end video captioning task. In this work, we proposed 4 POS blocks - Verb, Det + Subject, Aux Verb, and Det + Object. The visual features from the POS blocks are given as input to the GLFB. In addition to this, word embeddings exploiting linguistic information are obtained by supervising POS blocks with respective POS components. This helps in achieving the alignment between visual and linguistic features at both the fine-grained local and global levels, unlike previous methods. The fine-grained local and global visual and linguistic features are then fed into the Caption block to generate the final caption.

3.1. Problem Formulation

Given a video-captioning task, we aim to align the visual semantics with the linguistic semantics using the POS blocks. These POS components are inherently present in any grammatically structured caption and these blocks help to exploit fine-grained local visual features of a video to align linguistic and visual semantics. These aligned semantics on the fine-grained local level are then used as input to the GLFB to fuse local and global representations for improved semantic correctness.

For supervised training of each POS block, we extract the respective POS component from the ground-truth caption $C$. Let $\mathcal{R}$ be all the reference captions for a given video $\mathcal{V}$, ground-truth $\mathcal{G}_{POS}$ for the POS blocks (see Section 3.3) are defined as:

$$\mathcal{G}_{POS} = \{\{S^d, V, A, O^d\} \in C | \forall C \in \mathcal{R}\},$$

(1)

where $S^d, V, A, O^d$ are the ground-truth labels for Det + Subject, Verb, Aux Verb and Det + Object blocks respectively. These labels when converted to word embeddings help align the visual semantics with their linguistic counterparts in a supervised setting. Unlike existing methods that align visual and linguistic semantics on a caption or a nouns-verb level only, we perform a fine-grained local alignment of visual and linguistic semantics for four POS blocks. Our overall objective is as follows:

$$\mathcal{O}_{OVERALL} = \mathcal{O}_{POS} + \mathcal{O}_{GLFB} + \mathcal{O}_{VFM},$$

(2)

where $\mathcal{O}_{POS}$, $\mathcal{O}_{GLFB}$ and $\mathcal{O}_{VFM}$ are POS, GLFB and VFM objectives, respectively.

3.2. Visual Features Masking (VFM)

Random image masking improves the results for image detection and classification tasks in the previous methods [17,54]. A random patch of the image is masked as an input to the model and the model is forced to recover the masked patch along with the overall image. In our approach, we randomly mask 30% of the spatial features and a chunk of the temporal features, approximately 15%, before the features are input to the POS blocks. The masking percentages are selected through experimentation. When a chunk is masked in the temporal features of a video, it masks the information in the temporal dimension in all the sampled frames. This helps the proposed network model to generalize well across the entire data distribution. It also forces the spatial and temporal features to interact and find the masked information and this improves the overall performance (see Section 4.2).

3.3. POS Blocks

We propose to use four POS blocks in the proposed network, two of these four blocks are a combination of two POS components i.e. ‘determinant + subject’ and ‘determinant + object’. The other two components are ‘verb’ and ‘auxiliary verb’. We performed a detailed analysis of the ground-truth captions in benchmark datasets to analyze the presence of the individual parts of speech components in each caption. Figure 3 shows the percentage of the presence of each POS component introduced in our model in the captions for MSVD and MSRVTT benchmark datasets. It is evident from the figure that the four POS blocks used in the proposed method are the largely represented POS components in the captions and also, adding blocks in the network for other less represented POS components affects the performance adversely. An extensive ablation (see Section 4.2) demonstrates the effectiveness of using these four POS blocks within the network to generate grammatically and semantically correct captions.

v_p:\]
\[v_f = \text{lstm}_b(\text{cat}(t_f, A_{nv}[n_f])), \quad (3)\]
hence, \(v_p = f(c(v_f)). A_{nv}\) is a learnable attention for \(n_f\).
For training of the Verb block, we get 'verb' from the ground-truth caption using pre-trained roBERTa \[31\] model and use the same model for ground-truth embedding \(v_g\).
The distance with the predicted embedding \(v_p\) is minimized as part of the model learning. Let \(\mathcal{L}_v\) be the Verb block loss:
\[\mathcal{L}_v = \text{Dist.}(v_p, v_g) \quad (4)\]

**Det-Subject.** In this block, we use a combination of two POS components, i.e., 'determinant + subject'. This block uses the spatial visual features \(s_f, n_f\) and \(v_f\) (see Eqn. 3) to determine the 'determinant + subject' of the caption. Similar to the Verb block, attention is used to combine the information from the inputs. \(n_f\) and \(v_f\) are used as inputs via learnable attentions \(A_{ns}\) and \(A_{vs}\) respectively. \(s_f\) are then concatenated with these attended features and are passed through an LSTM layer similar to the Verb block to get 'determinant + subject' visual features \(d_s f\). Let \(d_{sp}\) be the predicted 'determinant + subject' linguistic embedding after passing \(d_s f\) through the fully connected layer,
\[d_s f = \text{lstm}_b(\text{cat}(s_f, A_{vs}[v_f], A_{ns}[n_f])), \quad (5)\]
hence, \(d_{sp} = f(c(d_s f)). This\ approach\ creates\ a\ dynamic\ template\ of\ captions\ being\ filled\ with\ outputs\ from\ the\ POS\ blocks\ as\ these\ become\ available\ both\ during\ training\ and\ testing.\)

For labels, we get 'determinant' and 'subject' from the ground-truth caption using the pre-trained roBERTa \[31\] model. We combine these together with a space in-between and use the same model for ground-truth embedding \(d_s g\). The distance with the predicted embedding \(d_{sp}\) is minimized as a part of the model learning. Let \(\mathcal{L}_{ds}\) be the Det-Subject block loss:
\[\mathcal{L}_{ds} = \text{Dist.}(d_{sp}, d_s g) \quad (6)\]

**Aux Verb.** Aux verb block uses \(t_f,\) noun features \(n_f,\) \(v_f\) (see Eqn. 3) and \(d_s f\) (see Eqn. 5) to predict 'auxiliary verb' for the caption. It also gives an insight into the singular or plural nature of the 'subject', which is used as an input to the GLFB. \(n_f,\) \(v_f\) and \(d_s f\) are used as the inputs via learnable attentions \(A_{na}, A_{va}\) and \(A_{dso}\) respectively. \(t_f\) is then concatenated with these attended features and is passed through an LSTM layer like in previous blocks to get 'auxiliary verb' visual semantics \(a_f,\) \(a_p\) is obtained after passing \(a_f\) through a fully connected layer:
\[a_f = \text{lstm}_b(\text{cat}(t_f, A_{va}[v_f], A_{na}[n_f], A_{dso}[d_s f])), \quad (7)\]
hence, \(a_p = f(c(a_f)). For\ ground-truth\ labels,\ we\ extract\ the\ 'auxiliary\ verb'\ from\ the\ ground-truth\ caption\ and\ also,\ its\ embedding\ \(a_g\)\ using\ pre-trained\ roBERTa\ \[31\] model. Then, the distance with \(a_p\) is minimized as part of the model training. Let \(\mathcal{L}_a\) be the Aux Verb block loss:
\[\mathcal{L}_a = \text{Dist.}(a_p, a_g) \quad (8)\]

**Det+Object.** Last, in the POS blocks, the Det+Object block is a combination of the 'determinant' and 'object'. It exploits linguistically aligned visual features from the previous POS blocks along with \(s_f,\) Like previous POS blocks, learnable attentions \(A_{no}, A_{vo}, A_{ds}o\) and \(A_{ao}\) are used to exploit information in \(n_f,\) \(v_f,\) \(d_s f\) and \(a_f\) respectively. In similar manner, \(d_{op}\) is obtained from \(d_{of}\) by:
\[d_{of} = \text{lstm}_b(\text{cat}(s_f, A_{vo}[v_f], A_{no}[n_f], A_{dso}[d_s f], A_{ao}[a_f])), \quad (9)\]
hence, \(d_{op} = f(c(d_{of})). For\ training,\ we\ get\ the\ 'determinant'\ and\ 'object'\ from\ the\ ground-truth\ caption\ using\ pre-trained\ roBERTa\ \[31\] model, combine these together with a space in-between and then, use the same model for ground-truth embedding \(d_o g\). The distance with the predicted embedding \(d_{op}\) is minimized as part of the model learning. Let \(\mathcal{L}_{do}\) be the Det+Object block loss:
\[\mathcal{L}_{do} = \text{Dist.}(d_{op}, d_o g) \quad (10)\]

### 3.4. Global Local Fusion Block (GLFB)

The proposed GLFB is used to achieve fusion between the global and local features, by using the fine-grained local and linguistically aligned visual features from the POS blocks and the visual-spatial features. Visual features \(g_f\) are the output of an LSTM layer, given a set of inputs using the attention mechanism. To predict a linguistic GLFB representation \(g_p, g_f\) is passed through a fully connected layer:
\[g_f = \text{lstm}_b(\text{cat}(s_f, A_{va}[v_f], A_{na}[n_f], A_{dso}[d_s f], A_{vo}[a_f], A_{ao}[a_f])), \quad (11)\]
hence, \(g_p = f(c(g_f)). For\ training,\ this\ block,\ the\ whole\ ground-truth\ caption\ is\ first\ converted\ to\ lower-case\ letters\ and\ then,\ passed\ through\ roBERTa\ \[31\] to\ get\ ground-truth\ embedding \(g_g\). Then, the distance with the predicted embedding \(g_p\) is minimized as part of the supervised learning. Let \(\mathcal{L}_g\) be the GLFB loss, then:
\[\mathcal{L}_g = \text{Dist.}(g_p, g_g) \quad (12)\]

### 3.5. Attention Mechanism

Our approach employs an attention mechanism in all blocks to exploit the inputs as can be seen in previous sections. Here, we generically define the attention \(\alpha_g\):
\[\alpha_g = w_g \text{tanh}(W_g X + U_g Y + b_g), \quad (13)\]
where $w_g$, $W_g$, $U_g$ and $b_g$ are the training parameters. $X$ is either spatial or temporal features, depending on whatever is the input to each block, in all blocks, except the Caption block (see Section 3.6) where it is the previous word in the generated caption. $Y$ is the visual semantics generated in the POS blocks and GLFB.

### 3.6. Overall Training Objective

The overall training objective adds the POS blocks and GLFB together along with the Caption block. The Caption block uses an LSTM layer, followed by a fully connected layer, with inputs of both visual and linguistic semantics from the POS blocks and GLFB to generate a caption, word by word, using beam search, as an output. Overall learning loss $L_{all}$ is defined as:

$$
L_{all} = L_c + L_v + L_{ds} + L_a + L_{do} + L_g,
$$

(14)

where $L_c$ is the cross entropy loss for the Caption block:

$$
L_c = - \sum_{n=1}^{N} E(w_n) \log P_n
$$

(15)

$E(w_n)$ is the one-hot encoding of the word $w_n$ and $P_n$, the output of the fully connected layer, is the probability distribution over the word vocabulary. It is worth mentioning that loss for the VFM objective is an integral part of $L_{all}$ as the model has to predict the right embedding with masked input. Also, we have done some experiments by tuning the loss weights but the effect on the overall performance was found to be minimal.

The objective of the Caption block is to generate the final video caption which is both grammatically and semantically correct.

### 4. Experimental Results and Evaluations

#### Datasets.

Similar to the existing video captioning approaches [1,33,36,38,51,58,60,64], we evaluate our model on MSRVTT [55] and MSVD [8] benchmark datasets. These datasets have a total of 10,000 and 1,970 videos respectively. We use the same train, validation, and test split as the existing methods.

#### Metrics.

For evaluation with the existing methods, we use the following set of video captioning metrics: BLEU@4 (B@4) [35], METEOR (M) [5], ROUGE-L (R) [27] and CIDEr (C) [49]. B@4 is based on n-gram (n=4) matches between the predicted and ground-truth captions. In contrast, M is a recall-based metric that uses word-to-word alignment between predicted and ground-truth captions. The final score is calculated between the best-scoring ground truth and the predicted caption. R, another recall-based metric, measures the longest common set of shared words with similar order between predicted and ground-truth captions. C is studied to be robust in the condition where the semantic meaning of the caption remains intact [6]. It measures the average cosine similarity between predicted and ground-truth captions. Also, we are the first to use GPT-2 [37] pre-trained model for measuring the Grammatical correctness Score (GS) of the captions generated by our model in comparison to state-of-the-art which we implement for the GS metric, demonstrating improved performance. GS metric measures the grammatical correctness without using the ground-truth captions unlike the other four metrics (see Section 4.2 and 4.3).

**Implementation.** For text, we have used spaCy$^1$ with roBERTa [31], a version of BERT [10], to extract POS components along with nouns from the ground-truth captions. We again use pre-trained roBERTa [31] for text embedding which gives an embedding of size 768 for 'determinant + subject', 'verb', 'auxiliary verb', 'determinant + object', and the whole caption. Following the existing methods, we use InceptionResNetV2 [46] to extract the spatial features and C3D [16] to extract the temporal features. These features are projected to 512 sizes before being input into the network. We train for epochs 25 and use a learning rate of 0.00015, batch size 16, ADAM optimizer [22], 16 samples per video as well as a hidden state size of 512 for the Caption block. Our model has 76M parameters and 0.045s inference time. Apart from that, we use Yolov7 [52] for extracting object features for the noun anchor. The whole implementation is performed using one NVIDIA GeForce RTX 3090 and PyTorch.

### 4.1. Results and Comparison with Existing Methods

We do a quantitative evaluation of the benchmark datasets for several state-of-the-art methods and also, demonstrate qualitative results of the proposed method.

**Quantitative Evaluation:** Table 1 gives a quantitative comparative performance of the proposed model against the state-of-the-art methods which do not employ large-scale transformer pre-training. The proposed SEM-POS network outperforms all these methods on both MSVD [8] and MSRVTT [55] benchmarks. Improvement is around 3% and 4% on CIDEr for MSRVTT and MSVD benchmarks respectively. In terms of GS metric, our method gives an improvement of around 10% and 16% for MSRVTT and MSVD respectively.

This is due to the fact that our method explores fine-grained local features in the four POS blocks before generating a global representation, giving it a better generalization ability compared to the other approaches. Visual features aligned with linguistics in POS blocks are learned across the entire data distribution for the video captioning task, which contributes to this improved performance. POS blocks help to align the linguistic and visual semantics at

---

$^1$https://spacy.io/
the fine-grained local level and then the GLFB visibly fuses it with global representation. Both these contributions of POS blocks and GLFB help generate grammatically and semantically correct sentences. Our method also outperforms these methods [28, 29, 40] (which have a large-scale transformer-based pre-training stage) on the GS metric, improving on their scores of 655.2, 643.7, and 612.8 (MSVD) and 224.7, 212.9 and 201.6 (MSRVTT) respectively.

**Qualitative Results:** In Figure 4, we show qualitative results for our approach. It can be seen that generated captions adequately represent not only the ground-truth captions but also the fine-grained visual features. For example, in the images in the first two rows, on the right-hand side, the proposed caption is better than ground truth and provides more grammatically and semantically correct information about ‘man’, ‘riding’, ‘road’ (single) and ‘football’ (single) instead of ‘person’, ‘on’, ‘streets’ (plural) and ‘sports’ (plural) respectively, which is derived from the fine-grained local features in the POS blocks.

We also compare the qualitative results of our method with the state-of-the-art method HMN [58]. Row 1 right, row 2 left and row 3 right show that our method generates captions that are grammatically and semantically more correct than HMN [58]. The rest of the results also show that our method generates captions as well as state-of-the-art.

### 4.2. Ablation Results

**Effectiveness of POS blocks and GLFB.** This section supports our first two contributions of using POS blocks and GLFB in the network and the results are shown in Table 2. Ablation is performed on the proposed method by removing the POS blocks one by one from the proposed network and evaluating the performance. The first four rows in the table demonstrate these results. As it is evident from the table the Verb block contributes significantly to the performance of the network. Det+Subject and Det+Object blocks contribute equally to the MSRVTT [55] benchmark but Det+Object shows only a minor contribution for the MSVD [8] benchmark dataset as it has a significant number of ground-truth captions without an object.

‘w/o GLFB’ shows results without the GLFB (see Section 3.4). It demonstrates the importance of combining fine-grained local and global features for video captioning. Three rows ‘w Adverb’, ‘w Adjective’, and ‘w Conjunction’ show that the performance is affected adversely as these POS components are not widely represented in the captions. Rows ‘w Subject only’ and ‘w Object only’ are Subject and Object blocks without Determinant and these show a decrease in performance.

Row ‘w All’ show the results for adding Adverb, Adjective, and Conjunction blocks to our model. It is evident that adding blocks that are not represented enough in the POS components, decreases the performance significantly.

**Effectiveness of VFM:** We also evaluate the performance without Visual Feature Masking (VFM) which is introduced...
<table>
<thead>
<tr>
<th>Block</th>
<th>MSVD</th>
<th>MSRVTT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B@4 ↑</td>
<td>M ↑</td>
</tr>
<tr>
<td>w/o Det + Subject</td>
<td>56.4</td>
<td>36.6</td>
</tr>
<tr>
<td>w/o Verb</td>
<td>55.9</td>
<td>34.9</td>
</tr>
<tr>
<td>w/o Aux Verb</td>
<td>57.7</td>
<td>37.1</td>
</tr>
<tr>
<td>w/o Det + Object</td>
<td>57.8</td>
<td>37.1</td>
</tr>
<tr>
<td>w/o GLFB</td>
<td>56.7</td>
<td>36.0</td>
</tr>
<tr>
<td>w/o VFM</td>
<td>59.6</td>
<td>38.1</td>
</tr>
<tr>
<td>w Adverb</td>
<td>56.5</td>
<td>36.8</td>
</tr>
<tr>
<td>w Conjunction</td>
<td>56.2</td>
<td>34.5</td>
</tr>
<tr>
<td>w Subject only</td>
<td>58.4</td>
<td>37.9</td>
</tr>
<tr>
<td>w Object only</td>
<td>58.9</td>
<td>37.8</td>
</tr>
<tr>
<td>w All</td>
<td>57.9</td>
<td>35.9</td>
</tr>
<tr>
<td>SEM-POS (Ours)</td>
<td><strong>60.1</strong></td>
<td><strong>38.5</strong></td>
</tr>
</tbody>
</table>

Table 2. Ablation results for the effectiveness of POS, GLFB, and VFM on MSVD and MSRVTT benchmarks.

in Section 3.2. As seen from ‘w/o VFM’ in Table 2, masking spatial and temporal features increases the performance to a considerable extent. We use different techniques of masking for spatial and temporal features. For spatial features, we find random masking across the whole frame features to be more effective than masking particular patches in a frame. We have introduced chunk-wise masking for temporal features as random masking is counterproductive in this case. In chunk-wise masking, we mask a chunk of values that are adjacent to each other in the temporal dimension and spatially the same for each frame.

### 4.3. Grammatical Score Evaluation

In video captioning, in addition to semantic correctness, it is important to check whether the generated captions are grammatically correct. Existing methods only evaluate the semantic correctness of the generated captions, however, in this section, we evaluate the Grammatical Score (GS) of the captions generated from the proposed method against state-of-the-art methods. [39] performed a comprehensive study to use GPT-2 [37] for measuring the grammatical correctness of English sentences. We employ the same technique to measure the average GS of generated captions on both MSVD [8] and MSRVTT [55] datasets as part of our ablation study in Table 2 and we compare the GS against state-of-the-art methods in Table 1. It is based on the perplexity measurement for each caption and the lower score means that the caption has fewer grammatical errors. The results demonstrate that the proposed method with POS blocks and GLFB blocks gives the best performance and generated more descriptive and grammatically correct captions compared to the existing methods achieving the new state-of-the-art.

### 4.4. Limitations

The performance of each POS block is limited to the number of samples present in the datasets. For e.g., due to a lack of ‘auxiliary verb’ samples in the dataset, this does not contribute significantly to the final performance. Similarly, other POS, like adverbs and adjectives, are far less represented in the captions, and adding these blocks, affects the performance adversely. Also, the proposed method does not handle multiple subjects/objects like existing methods. We aim to address these limitations in our future work.

### 5. Conclusion

This paper introduces a novel global-local feature fusion network that fuses local and global, language and visual features for video captioning. Four POS and a GLFB block are introduced within the network to create a coherent representation to generate grammatically and semantically correct captions, which is proved by the performance of the proposed method on MSVD [8] and MSRVTT [55] benchmarks for five different metrics against state-of-the-art methods. The application of Visual Features Masking (VFM) within the novel network helps our model to not only achieve better performance but also generalize well across the data distribution, as seen in the ablation results. Qualitative results show the ability of our model to learn fine-grained details in the video and we believe that the insights from this work will further open up research in video-language alignment and understanding not only for semantic correctness but also for grammatical correctness.

### 6. Acknowledgement

This research was partly supported by the British Broadcasting Corporation Research and Development (BBC R&D), Engineering and Physical Sciences Research Council (EPSRC) Grant EP/V038087/1 “BBC Prosperity Partnership: Future Personalised Object-Based Media Experiences Delivered at Scale Anywhere”.

2612
References


[34] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly Modeling Embedding and Translation to Bridge Video and Language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4594–4602, 2016. 1


[39] Scribendi Inc. Comparing bert and gpt-2 as language models to score the grammatical correctness of a sentence, 2020. 8


[55] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT:A Large Video Description Dataset for Bridging Video and Language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5288–5296, 2016. 2, 6, 7, 8


