Contextualise, Attend, Modulate and Tell: Visual Storytelling

Zainy M. Malakan[1,2][a], Nayyer Aafaq[1][b], Ghulam Mubashar Hassan[1][c] and Ajmal Mian[1][d]

[1] *Department of Computer Science and Software Engineering, The University of Western Australia, Australia*

[2] *Department of Information Science, Faculty of Computer Science and Information System, Umm Al-Qura University, Saudi Arabia*

Keywords: Storytelling, Image Captioning, Visual Description.

Abstract: Automatic natural language description of visual content is an emerging and fast-growing topic that has attracted extensive research attention recently. However, different from typical 'image captioning' or 'video captioning', coherent story generation from a sequence of images is a relatively less studied problem. Story generation poses the challenges of diverse language style, context modeling, coherence and latent concepts that are not even visible in the visual content. Contemporary methods fall short of modeling the context and visual variance, and generate stories devoid of language coherence among multiple sentences. To this end, we propose a novel framework Contextualize, Attend, Modulate and Tell (CAMT) that models the temporal relationship among the image sequence in forward as well as backward direction. The contextual information and the regional image features are then projected into a joint space and then subjected to an attention mechanism that captures the spatio-temporal relationships among the images. Before feeding the attentive representations of the input images into a language model, gated modulation between the attentive representation and the input word embeddings is performed to capture the interaction between the inputs and their context. To the best of our knowledge, this is the first method that exploits such a modulation technique for story generation. We evaluate our model on the Visual Storytelling Dataset (VIST) employing both automatic and human evaluation measures and demonstrate that our CAMT model achieves better performance than existing baselines.

1 INTRODUCTION

Describing a story from a sequence of images, *a.k.a. visual storytelling* (Huang et al., 2016), is a trivial task for humans but a challenging one for machines as it requires understanding of each image in isolation as well as in the wider context of the image sequence. Furthermore, the story must be described in a coherent and grammatically correct natural language. Closely related research areas to visual storytelling are *image captioning* (Pan et al., 2020; Feng et al., 2019; Yang et al., 2019; Donahue et al., 2015) and *video captioning* (Zhang and Peng, 2020; Aafaq et al., 2019a; Yan et al., 2019; Liu et al., 2020; Yao et al., 2015; Gan et al., 2017; Pan et al., 2017; Aafaq et al., 2019b). Figure 1 demonstrates the difference between descriptions of images in isolation and stories for images in a sequence. Similarly we compare the example of two captions: "*stand in front of a flag*" and "*time to honor the flag*". The first caption captures the image content which is literal and concrete.

1- Three men are wearing black and walking on the street.
2- Group of people are standing in front of a flag.
3- A man with a black shirt is playing the bagpipes
4- tow men are standing in front of a table.
5- soldier with a black hat stand in front of a flag

Story 1: the soldiers were ready to go . they were all seated and waiting for their turn . they were greeted by a man . then he spoke with his speech. Then it was time to honor the flag

Story 2: the soldiers were ready to begin their ceremony . they were all very proud of their accomplishments. they were so happy that day .everyone was excited for the ceremony .it was a great time.

Figure 1: An example highlighting the differences between image captioning and storytelling. First block (top) depicts that each image is captioned with one sentence in isolation. The second block (bottom) indicates the narrative description for the same images stream.

[a] https://orcid.org/0000-0002-6980-0992
[b] https://orcid.org/0000-0003-2763-2094
[c] https://orcid.org/0000-0002-6636-8807
[d] https://orcid.org/0000-0002-5206-3842

Malakan, Z., Aafaq, N., Hassan, G. and Mian, A.
Contextualise, Attend, Modulate and Tell: Visual Storytelling.
DOI: 10.5220/0010314301960205
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, pages 196-205
ISBN: 978-989-758-488-6

However, the second caption requires further inference about what it means to honor a flag.

In contrast to conventional image captioning where a static input image is presented devoid of context, visual storytelling is more challenging as it must capture the contextual relationship among the events as depicted by the sequence of images. Moreover, in comparison to video captioning, visual storytelling poses challenges due to large visual variation at the input (Wang et al., 2018a). On the language side, it requires narrative generation rather than literal description. To achieve the narrative, it requires to enhance long term image consistency among multiple sentences. The task becomes even more challenging as some stories involve description of human sub-cognitive concepts that do not appear in the visual content.

To address the aforementioned challenges, we propose a novel framework that starts by modeling the visual variance as a temporal relationship among the stream of images. We capture the temporal relationship in two directions *i.e*, *forward* as well as *backward* using Bi-LSTM. To maintain the image specific relevance and context of images, we project the image information and the context vectors from Bi-LSTM into a joint latent space. The spatio-temporal attention mechanism learns to relate the corresponding information by focusing on image features (spatial-attention) and context vectors (temporal attention). The attentive representation (*i.e*, encoder output) is then modulated with the input word embeddings before it is fed into the language model. We used the *Mogrifier*-LSTM (Melis et al., 2020) architecture to achieve this task. During the gated modulation in the first layer, the first gating step scales the input embedding, depending on the actual context, resulting in a contextualized representation of the input. In addition, the input token itself is a contextual representation in which it occurs. This results in the LSTM input highly context dependent and enabling the language model to generate more relevant and contextual descriptions of the images. Furthermore, intra-sentence coherence is improved by feeding the last hidden state of the first sentence generator to the next sentence generator.

Our contributions are summarised as follows:

- We propose a novel framework Contextualise, Attend, Modulate and Tell (CAMT) to generate a story from a sequence of input images. We first model the bidirectional context vectors capturing the temporal relationship among the input images from the same stream and then project the regional image features and the contextual vectors to a joint latent space and employ the spatio-temporal attention.
- We perform gated modulation on attentive representation and the input token before feeding it into the language model *i.e*, LSTM to model the interaction between the two inputs. The modulated transition function results in a highly context-dependent input to the LSTM.
- To demonstrate the effectiveness of proposed technique for visual storytelling, we perform experiments on the popular Visual Storytelling Dataset (VIST) with both automatic and human evaluation measures. The superior performance of our model over the current state-of-the-art in both automatic and human evaluations shows the efficacy of our technique.

2 RELATED WORK

Below we discuss the literature related to image captioning and video captioning which are closely associated to visual storytelling, followed by literature review of visual storytelling.

2.1 Image Captioning

Image captioning can be categorised as a single frame (*i.e*, image) described by a single sentence. The methods can further be sub-categorised into rule based methods (do Carmo Nogueira et al., 2020; Mogadala et al., 2020) and deep learning based methods (Phukan and Panda, 2020; Huang et al., 2019a). The rule based methods employ the classical approach of detecting pre-defined and limited number of subjects, actions and scenes in an image and describe them in natural language using template based techniques. Due to advancements in deep learning and introduction of larger datasets (Krizhevsky et al., 2012), most recent methods normally rely on deep learning and advanced techniques such as attention (Wang et al., 2020), reinforcement learning (Shen et al., 2020), semantic attributes integration (Li et al., 2019a) and, subjects and objects modeling (Ding et al., 2019). However, all these techniques do not perform well in generating narrative for a sequence of images.

2.2 Video Captioning

Video captioning can be treated as multi-frame (*i.e*, video) described by a single sentence. This involves capturing the variable length sequence of video frames and mapping them to variable length of words in a sentence. Like image captioning, classical video

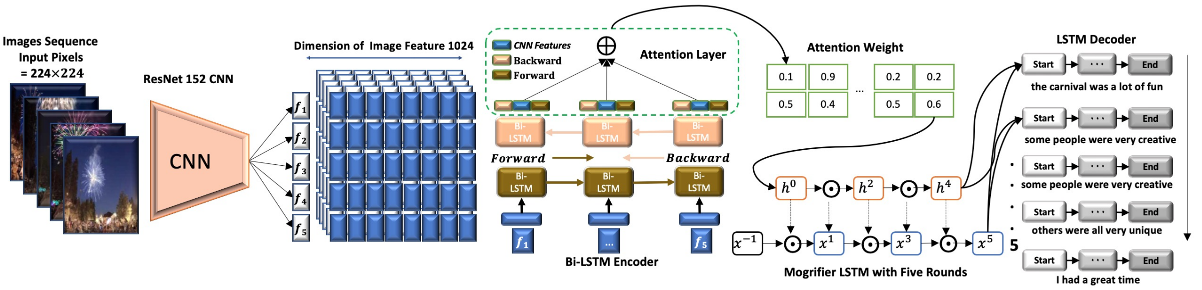


Figure 2: Our model is based on encoder-decoder architecture. The encoder part consists of pre-trained ResNet 152 to extract the visual features from each image. The feature vectors are then fed to a bidirectional Long Short-term Memory (LSTM) network sequentially, which allows the context of all images to be reflected in the entire story. The decoder is a Mogrifier LSTM comprising five rounds which improve the generated story.

captioning methods (Kojima et al., 2002; Das et al., 2013; Krishnamoorthy et al., 2013; Nayyer et al., 2019) follow template based approaches. Most recent methods rely on the neural networks based framework with encoder-decoder architecture being most widely employed. The encoder *i.e*, 2D/3D-CNN, first encodes the video frames and the decoder *i.e*, Recurrent Neural Network (RNN), decodes the visual information into natural language sentences. More recent video captioning methods employ *e.g* reinforcement learning (Wang et al., 2018d), objects and actions modeling (Pan et al., 2020; Zheng et al., 2020; Zhang et al., 2020), Fourier transform (Aafaq et al., 2019a), attention mechanism (Yan et al., 2019; Yao et al., 2015), semantic attribute learning (Gan et al., 2017; Pan et al., 2017), multimodal memory (Wang et al., 2018b; Pei et al., 2019) and audio integration (Hao et al., 2018; Xu et al., 2017) for improved performance. Few video captioning methods attempt to describe the video frames into multiple sentences (Yu et al., 2016; Xiong et al., 2018). However, describing a short video by a single or rarely by multiple sentences, the challenges of storytelling are multi-fold, as discussed in Section 1. Hence, most of the aforementioned video captioning techniques are ineffective for the task of storytelling in their original form.

2.3 Visual Storytelling

Storytelling is one of the oldest activities of mankind and has attracted extensive research recently due to improvement in computation and machine learning. Visual storytelling provides a strategy to understand activities within stream of images and summarise these images in one paragraph (Wiessner, 2014). It started with ranking and retrieval approach which is used to retrieve a paragraph rather than a sentence from multiple images rather than one image (Kim et al., 2015). Later, Coherent Recurrent Convolutional Network (CRCN) was introduced which enhanced the smooth flow of various sentences in a photostream (Park et al., 2017). This deep learning network encompasses entity-based local coherence model, bidirectional Long Short Term Memory (LSTM) networks, and convolutional neural networks. A multiple reward functions approach during training is used to understand the essential reward role from human demonstrations (Wang et al., 2018c). LSTM based model and transformer-decoder was introduced to improve coherence in stories (Hsu et al., 2019). Encoder-decoder based deep learning models are deployed to improve image features and generating sequence of sentences (Al Nahian et al., 2019; Huang et al., 2019b). Recurrent neural Network (RNNs) are introduced to improve the modulation which enhanced the relevance, coherence and expressiveness of the generated story (Hu et al., 2020). Overall, the listed studies are mostly computationally expensive due to their complexity level. In this paper, our focus is to propose a computationally efficient method that can generate human-like story.

3 METHODOLOGY

Figure 2 shows the architecture of the proposed model. It comprises of three modules: *Bi-Encoder*, *Attention* and *Decoder*. Our *Bi-Encoder* module first encodes the image features and then it also encodes the temporal context of the l images. The *Attention* module captures the relationship at two levels: image and image-sequence. Finally, the *Decoder* module generates sequence of sentences by exploiting the mogrifier LSTM architecture. Details of the three modules are discussed below.

3.1 Bi-Encoder

Given a stream of l images *i.e*, $I = (I_1, I_2, ..., I_l)$, we extract their high level feature vectors by employing a pre-trained convolutional neural network ResNet152 (He et al., 2016). We denote the extracted features by $f = [f_1, f_2, ..., f_l]$. To further capture the context information, we employ a sequence-encoder. A naive method is to simply concatenate the image features with the same sequence length, however, this loses the temporal relationship amongst the images. Alternatively, a recurrent neural network can be employed to capture the sequential information that aggregates the temporal information over time. To capture this relationship, we employ a Bi-LSTM, that summarises the sequential information of l images in both directions *i.e*, *forward* as well as *backward*. At each time step 't', our sequence encoder takes an input of image feature vector f_i where $i \in \{1,2,3,4,5\}$. At last time step *i.e*, $t = 5$, the sequence encoder has encoded the whole stream of images and provides the contextual information through the last hidden state denoted as $h_{se} = [\overrightarrow{h_{se}}; \overleftarrow{h_{se}}]$.

3.2 Attention Module

The task of visual storytelling from a sequence of images involves describing the image(s) content in the context of the other images. Here, the description of images in isolation is undesired. To capture the image features in the context of all the images, we incorporate an attention mechanism on image features and the output of the sequence encoder that includes the overall context of visual stream. Formally,

$$\varphi_i = W_a^T \cdot tanh(W_f[f_i, h_{se_i}] + b) \tag{3.1}$$

where f_i is the feature vector of i^{th} image with hidden state h_{se_i} of sequence encoder after i^{th} image has been fed.

$$\alpha_i = \frac{exp(\varphi_i)}{\sum_{k=1}^{l} exp(\varphi_k)} \tag{3.2}$$

where k is the length of the visual stream. Finally, the attentive representation becomes;

$$\zeta_i = \sum_{i=1}^{k} \alpha_i \cdot [f_i, h_{se_i}] \tag{3.3}$$

The final representations serve as the decoder inputs which attends both image specific (low level) and stream specific (high level) information.

3.3 Decoder

Recurrent networks, while successful, still lack in generalisation while modeling the sequential inputs especially for the problems where coherence and relevance are vital. In visual storytelling, the problem depends on how model inputs interact with the context in which they occur. To address these problems, we exploit the modulating mechanism of the *Mogrifier* LSTM (Melis et al., 2020) and employ as decoder in our framework. We first describe how the standard LSTM (Hochreiter and Schmidhuber, 1997) generates the current hidden state *i.e*, $h^{\langle t \rangle}$, given the previous hidden state h_{prev}, and updates its memory state $c^{\langle t \rangle}$. For that matter, LSTM uses input gates Γ_i, forget gates Γ_f, and output gates Γ_o that are computed as follows:

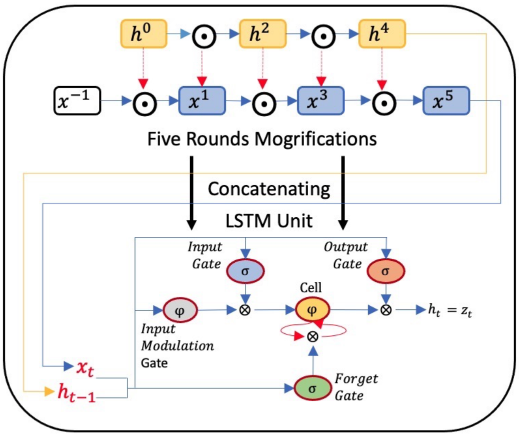


Figure 3: Illustrates how Mogrifier modulation is concatenating with LSTM unit. Before the input to LSTM, two inputs x and h_{prev} modulate each other alternatively. After five mutual gating rounds, the highest indexed updated x and h_{prev} are then fed into normal LSTM unit.

$$\Gamma_f^{\langle t \rangle} = \sigma(W_f[h_{prev}, x_t] + b_f), \tag{3.4}$$

$$\Gamma_i^{\langle t \rangle} = \sigma(W_i[h_{prev}, x_t] + b_i), \tag{3.5}$$

$$\tilde{c}^{\langle t \rangle} = \tanh(W_c[h_{prev}, x_t] + b_c), \tag{3.6}$$

$$c^{\langle t \rangle} = \Gamma_f^{\langle t \rangle} \odot c^{\langle t-1 \rangle} + \Gamma_i^{\langle t \rangle} \odot \tilde{c}^{\langle t \rangle}, \tag{3.7}$$

$$\Gamma_o^{\langle t \rangle} = \sigma(W_o[h_{prev}, x_t] + b_o), \tag{3.8}$$

$$h^{\langle t \rangle} = \Gamma_o^{\langle t \rangle} \odot tanh(c^{\langle t \rangle}) \tag{3.9}$$

where x is the input word embedding vector at time step 't' (we suppress t at all places for readability), W_* represents the transformation matrix to be learned in all cases, b_* represent the biases, σ is the logistic sigmoid function, and $\odot$ represent the hadamard product of the vectors. In our decoder network, the LSTM hidden state h is initialized by attentive vector ζ_i from the encoder output.

LSTM has proven to be a good solution for the vanishing gradient problem. However, its input gate Γ_i also scales the *rows* of the weight matrices W_c (ignore the non-linearity in c). To this end, in the *Mogrifier* LSTM instead, the *columns* of all its weight matrices W_* are scaled by gated modulation. Before the input to the LSTM, two inputs x and h_{prev} modulate each other alternatively. Formally, x is gated conditioned on the output of the previous step h_{prev}. Likewise, the gated input is utilised in a similar fashion to gate previous time step output. After a few mutual gating rounds, the highest indexed updated x and h_{prev} are then fed into LSTM as presented in Figure 3. Thus, it can be expressed as: *Mogrify* $(x, c_{prev}, h_{prev}) = LSTM(x^{\uparrow}, c_{prev}, h^{\uparrow}_{prev})$ where $x^{\uparrow}$ and $h^{\uparrow}_{prev}$ are the modulated inputs being the highest indexed x^i and h^i_{prev} respectively. Formally,

$$x^i = 2\sigma(W^i_{xh} h^{i-1}_{prev}) \odot x^{i-2}, \quad \text{for odd i} \in [1, 2, ..., r] \tag{3.10}$$

$$h^i_{prev} = 2\sigma(W^i_{hx} x^{i-1}) \odot h^{i-2}_{prev}, \text{for even i} \in [1, 2, ..., r] \tag{3.11}$$

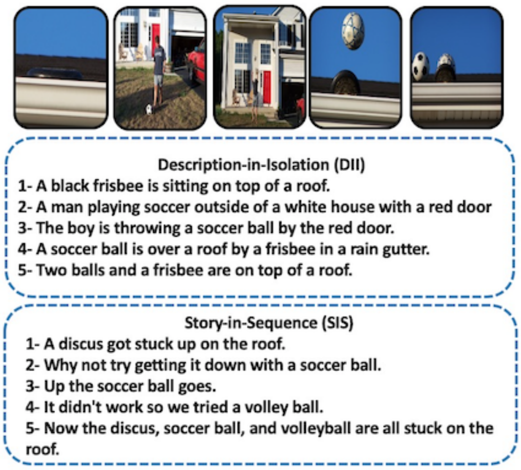
where $\odot$ is the Hadamard product, $x^{-1} = x$, $h^0_{prev} = h_{prev} = \zeta_i$ and r denotes the number of modulation rounds treated as a hyperparameter. Setting $r = 0$ represents the standard LSTM without any gated modulation at the input. Multiplication with a constant 2 ensures that the matrices W^i_{xh} and W^i_{hx} result in transformations close to identity.

3.4 Architecture Details

Our proposed architecture is based on deep learning model, which is combined in central two-parts, as shown in Figure 2. The first part is the encoder which is utilized a pre-trained Resnet-152 (He et al., 2016) as a CNN. The CNN is responsible for extracting the image features and feed all extracted features to bidirectional-LSTM. The bi-LSTM is utilized to reflect all context of the streaming images as story-like. Simultaneously, the bi-LSTM made up the vectors which are fed directly through a fully connected layer as word token into the decoder part. The decoder part is contained the Mogrifier LSTM with five steps. Then all decoded vectors are fed to LSTM tokenizer to generate visual stories. In the <start> sign, the tokenizer will start to receive the feature vectors from Mogrifier, and it tokenizes the sentence word until the LSTM meets the $< end >$ which is a complete sentence of the first image and so on for the whole story.

3.5 Model Training

All input images were resized to 256X256 pixels and their intensities were normalized to [0,1]. All the important words from the VIST dataset were extracted and used as words embedded with a diminution of 256 vectors. For the encoder part, we used the parameters settings suggested by the GLACNet model (Kim et al., 2018). The learning rate was set to 0.001 with weight decay equals to 1e-5, and used the Adam optimizer. We applied teacher-forcing algorithm for our LSTM training. To prevent overfitting and enhance the performance of the training, we set the batch size to 64.



Description-in-Isolation (DII)
1- A black frisbee is sitting on top of a roof.
2- A man playing soccer outside of a white house with a red door
3- The boy is throwing a soccer ball by the red door.
4- A soccer ball is over a roof by a frisbee in a rain gutter.
5- Two balls and a frisbee are on top of a roof.

Story-in-Sequence (SIS)
1- A discus got stuck up on the roof.
2- Why not try getting it down with a soccer ball.
3- Up the soccer ball goes.
4- It didn't work so we tried a volley ball.
5- Now the discus, soccer ball, and volleyball are all stuck on the roof.

Figure 4: VIST dataset is composed of two types of image sequences: Description In Isolation (DII), which is a collection of five images described in isolation, and Description In Sequence (SIS), which is a collection of five images captioned in sequence (Huang et al., 2016).

4 EVALUATION

Below are the details about the dataset and metrics which we used to evaluate our proposed model.

4.1 Visual Storytelling Dataset (VIST)

We evaluated our model on VIST dataset (Huang et al., 2016). It was the first dataset which was created for the problem of visual storytelling. The dataset contains 209,651 images creating 50,200 stories. There are two different types of sequence images: Description-in-Isolation (DII) and Story-in-Sequence (SIS). Both these types consist of the same photos, but the differences are in the description text. In DII, all sequence images are described as an image captioning, and they are not related to each other. In SIS, the images are described as a narrative story-like. Figure 4 demonstrate an example of DII and SIS description.

Table 1: Comparison of the efficiency of our proposed model with variants of GLAC Net (Kim et al., 2018) in terms of perplexity score, number of epochs and METEOR score. "-" means that the scores are not provided by the authors of the respective method. Results inside brackets represent best scores.

Model	Perplexity Score		METEOR	Number of
with Attention	Validation	Testing	Score	Epoch
LSTM Seq2Seq	21.89 %	22.18 %	0.2721	-
GLAC Net (-Cascading)	20.24 %	20.54 %	0.3063	-
GLAC Net (-Global)	18.32 %	18.47 %	0.2913	-
GLAC Net (-Local)	18.21 %	18.33 %	0.2996	-
GLAC Net (-Count)	18.13 %	18.28 %	0.2823	-
GLAC Net	18.13 %	18.28 %	0.3014	69
Ours (Encoder and Decoder Mogrifier)	17.11 %	16.77 %	0.3238	23
Ours - (Decoder Mogrifier only)	(16.67 %)	(15.89 %)	(0.335)	(22)

4.2 Evaluation Metrics

In the area of image description, researchers generally use several evaluation metrics to measure the quality of their proposed techniques. The first evaluation metric used in our study is *Bilingual Evaluation Understudy* (BLEU) which compares set of reference texts with the generated-text using n-grams. It is considered to be optimal for measuring the efficiency of techniques with short sentences and have different versions. We selected BLEU-1, BLEU-2, BLEU-3, and BLEU-4 in our study (Papineni et al., 2002). The second evaluation metric used is *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) which is used to compare three different types of text summaries such as n-grams, word sequences and word pairs. ROUGE has various subtypes, such as ROUGE-1, ROUGE-2, ROUGE-W. ROUGE-L and ROUGE-SU4, and we use ROUGE-L as it is suitable for measuring efficiency for single text document evaluation and short summaries (Lin, 2004). The third evaluation metric used is *Metric for Evaluation of Translation with Explicit ORdering* (METEOR) which considers the words' synonyms with its matching with the text reference. It is optimal in measuring efficiency at the sentence level (Banerjee and Lavie, 2005). The final evaluation metric used is *Consensus-based Image Description Evaluation* (CIDEr) which is designed for image captioning evaluation. It compares generated-text with various human captions (Vedantam et al., 2015). All of these selected evaluation metrics thoroughly evaluate our proposed methodology and help us to compare with state-of-the-art techniques.

Table 2: Human evaluation survey results for 10 ground truth stories and also 10 generated stories by our proposed model.

Story	Rank 1-5 (worst-best)	
Type	Relevance	Coherence
Ground Truth	3.45	3.44
Proposed model (CAMT)	(3.66)	(3.65)

5 RESULTS AND DISCUSSION

In this section, we compare the results of our proposed approach with state-of-the-art methods. We also discuss the human evaluation of our technique and performance of the model during training.

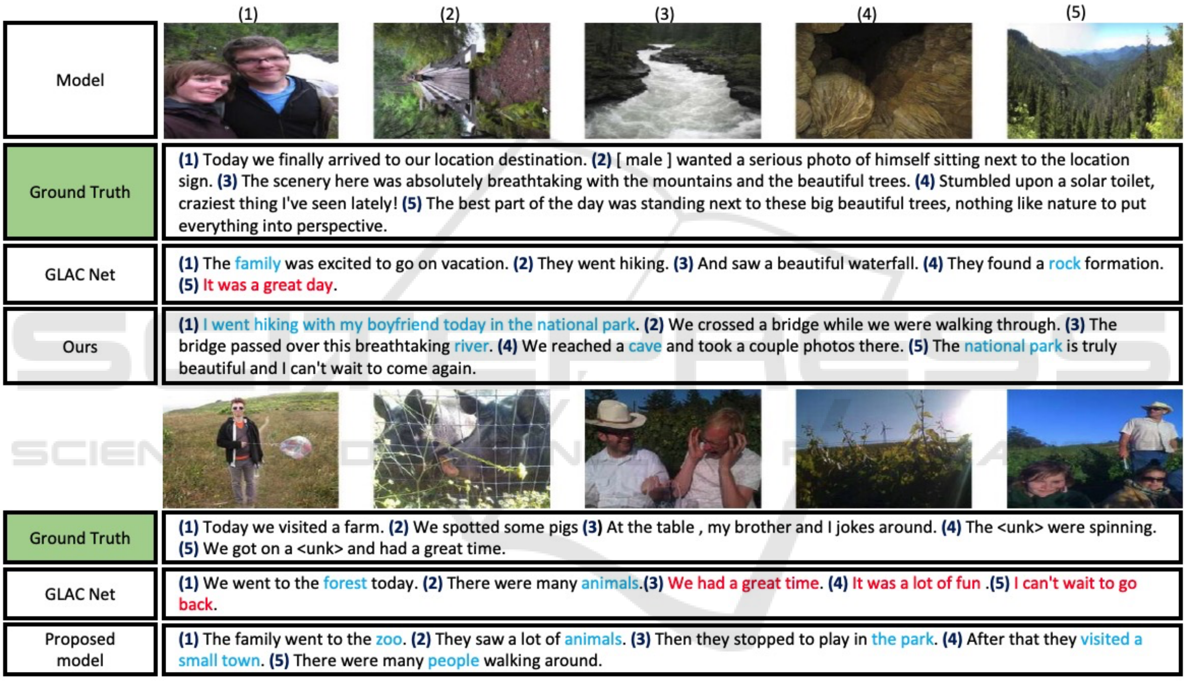
5.1 Quantitative Result

5.1.1 Perplexity Score Comparison

During training, we compare the proposed model's perplexity score with the model proposed by (Kim et al., 2018) which has different variants. We used two variants of our proposed model: using Mogrifier LSTM in both encoder and decoder parts, and using Mogrifier LSTM in decoder part only. Table 1 shows the perplexity scores, METEOR scores and number of epochs for all the selected models. Table 1 shows that using Mogrifier LSTM only in the decoder part improves the perplexity score to 16.67 and 15.89 for validation and test data respectively in only 22 epochs which is the best among the compared models. Similarly, we also found that our proposed model with Mogrifier LSTM in only the decoder part outperforms other models by achieving best METEOR score of 0.335. Therefore, we selected this variant of our model for comparison with other state-of-the-art techniques.

Table 3: Comparison of performance of our proposed model with state-of-the-art techniques using six automatic evaluation metrics. "-" means that the scores are not provided by the authors of the respective method. Results enclosed in brackets represent scores.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE-L
Show, Reward and Tell(Wang et al., 2018a)	0.434	0.213	0.104	0.516	(0.113)	-
AREL(baseline) (Wang et al., 2018c)	0.536	0.315	0.173	0.099	0.038	0.286
GLACNet(baseline) (Kim et al., 2018)	0.568	0.321	0.171	0.091	0.041	0.264
HCBNet (Al Nahian et al., 2019)	0.593	0.348	0.191	0.105	0.051	0.274
HCBNet(without prev. sent. attention) (Al Nahian et al., 2019)	0.598	0.338	0.180	0.097	0.057	0.271
HCBNet(without description attention) (Al Nahian et al., 2019)	0.584	0.345	0.194	0.108	0.043	0.271
HCBNet(VGG19) (Al Nahian et al., 2019)	0.591	0.34	0.186	0.104	0.051	0.269
VSCMR (Li et al., 2019b)	0.638	-	-	0.143	0.090	0.302
MLE (Hu et al., 2020)	-	-	-	0.143	0.072	0.300
BLEU-RL (Hu et al., 2020)	-	-	-	0.144	0.067	0.301
ReCo-RL (Hu et al., 2020)	-	-	-	0.124	0.086	0.299
Proposed model (CAMT)	(0.641)	(0.361)	(0.2011)	(0.1845)	0.042	(0.303)



Ground Truth: (1) Today we finally arrived to our location destination. (2) [male] wanted a serious photo of himself sitting next to the location sign. (3) The scenery here was absolutely breathtaking with the mountains and the beautiful trees. (4) Stumbled upon a solar toilet, craziest thing I've seen lately! (5) The best part of the day was standing next to these big beautiful trees, nothing like nature to put everything into perspective.

GLAC Net: (1) The family was excited to go on vacation. (2) They went hiking. (3) And saw a beautiful waterfall. (4) They found a rock formation. (5) It was a great day.

Ours: (1) I went hiking with my boyfriend today in the national park. (2) We crossed a bridge while we were walking through. (3) The bridge passed over this breathtaking river. (4) We reached a cave and took a couple photos there. (5) The national park is truly beautiful and I can't wait to come again.

Ground Truth: (1) Today we visited a farm. (2) We spotted some pigs (3) At the table , my brother and I jokes around. (4) The <unk> were spinning. (5) We got on a <unk> and had a great time.

GLAC Net: (1) We went to the forest today. (2) There were many animals.(3) We had a great time. (4) It was a lot of fun .(5) I can't wait to go back.

Proposed model: (1) The family went to the zoo. (2) They saw a lot of animals. (3) Then they stopped to play in the park. (4) After that they visited a small town. (5) There were many people walking around.

Figure 5: Comparison of stories generated by our proposed model and GLAC Net model with ground truth for a sequence of five images. The text in red represents repetition while text in blue represents relevance to the information in its subsequent image.

5.1.2 Comparison with State-of-the-Art Methods

We compare our proposed model with the following state-of-the-art models: Show, Reward and Tell method (Wang et al., 2018a); AREL[1], a method for an implicit reward with imitation learning (Wang et al., 2018c); GLACNet[2], an approach to learn attention cascading network (Kim et al., 2018); HCBNet, an approach of using image description as a hierarchy for the sequence of images (Al Nahian et al., 2019); VSCMR, a method of cross-modal rule mining (Li et al., 2019b); and ReCo-RL, an approach of designing composite rewards (Hu et al., 2020). All these methods achieve high scores on VIST visual storytelling dataset. The performance of all models are compared using the selected automatic evaluation metrics which include BLEU-1, BLEU-2, BLEU-3, BLEU-4, CIDEr, METEOR, and ROUGE-L. The evaluation metric script was published by (Hu et al.,

[1] https://github.com/eric-xw/AREL.git

[2] https://github.com/tkim-snu/GLACNet

2020)[3]. Table 3 presents the results for all the mentioned models, which clearly shows that our proposed model outperform other models on all the selected evaluation metrics except CIDEr score.

5.1.3 Human Evaluation

To evaluate the performance of our model in real world scenarios, we conducted a survey where participants were asked to rank the stories generated by our model in terms of *relevance* and *coherence*. We selected 10 stories generated by our model and 10 ground truth stories. All with their associated images and asked the participants to rank them on a scale of 1 to 5 (worst to best) in terms of relevance and coherence. Table 2 presents the results which shows that the participants found the stories generated by our model to be more coherent and relevant than the provided ground truth with the dataset.

5.2 Qualitative Analysis

Due to the nature of the visual storytelling problem, we thoroughly analysed the stories generated by our model. We also compared our stories with ground truth and stories generated by GLAC Net model (Kim et al., 2018). Figure 5 presents two sets of five images from VIST dataset along with the provided ground truth stories followed by stories generated by GLAC Net and our proposed model. The text in red mentions the repetition of the story while the text in blue mentions the relevance of the story with the images. In both stories generated by GLAC Net, we can observe the repetition of the information and non-relevance to the images while stories generated by our proposed model were more coherent and relevant to the generated images. For instance, the last three sentences in the story generated by GLAC Net for the lower set of images are “We had a great time”, “It was a lot of fun”, and “I can’t wait to go back”. All the three sentences do not contribute in the story and repeating the information. On the other hand, our proposed model generated story which is more coherent and relevant to the images.

5.3 Discussion

The main purpose of our study was to build a simple and computationally efficient model which can generate stories form images that are more relevant to the images and coherent in structure. From the results, it can be observed that our proposed model achieves both objectives in addition to achieving better performance than the existing state-of-the-art techniques. As mentioned in Table 1, our proposed model is time-efficient in training as it is trained over 22 epochs only. In addition, our proposed model is much simpler than existing models such as (Hu et al., 2020) which uses more complex architectures involving two RNNs. Whereas, our proposed model is designed on a simple encoder-decoder methodology achieving competitive results with state-of-art techniques.

6 CONCLUSION

In this paper, we introduced a novel encoder-decoder technique for visual description as storytelling. Our framework is straightforward and comprises a ResNet 152 and Bi-LSTM as an encoder, and Mogrifier LSTM with five steps as a decoder. In between, we utilize an attention mechanism which allows the model to generate a novel sentence according to a specific image region while maintaining the overall story context. Our proposed model outperforms state-of-the-art methods on automatic evaluation metrics. In future work, we aim to extend our model to generate different types of stories such as fiction. In addition, we aim to to extend our model to be able to track objects that re-appear in subsequent images and incorporate their re-appearance and movement into the story.

REFERENCES

Aafaq, N., Akhtar, N., Liu, W., Gilani, S. Z., and Mian, A. (2019a). Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *Proceedings of IEEE CVPR*, pages 12487–12496.

Aafaq, N., Akhtar, N., Liu, W., and Mian, A. (2019b). Empirical autopsy of deep video captioning frameworks. *preprint arXiv:1911.09345.*

Al Nahian, M. S., Tasrin, T., Gandhi, S., Gaines, R., and Harrison, B. (2019). A hierarchical approach for visual storytelling using image description. In *International Conference on Interactive Digital Storytelling*, pages 304–317. Springer.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Das, P., Xu, C., Doell, R. F., and Corso, J. J. (2013). A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object

[3] https://github.com/JunjieHu/ReCo-RL

stitching. In *Proceedings of the IEEE CVPR*, pages 2634–2641.

Ding, S., Qu, S., Xi, Y., Sangaiah, A. K., and Wan, S. (2019). Image caption generation with high-level image features. *Pattern Recognition Letters*, 123:89–95.

do Carmo Nogueira, T., Vinhal, C. D. N., da Cruz Júnior, G., and Ullmann, M. R. D. (2020). Reference-based model using multimodal gated recurrent units for image captioning. *Multimedia Tools and Applications*, pages 1–21.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE CVPR*, pages 2625–2634.

Feng, Y., Ma, L., Liu, W., and Luo, J. (2019). Unsupervised image captioning. In *Proceedings of the IEEE CVPR*.

Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., and Deng, L. (2017). Semantic compositional networks for visual captioning. In *Proceedings of the IEEE CVPR*, pages 5630–5639.

Hao, W., Zhang, Z., and Guan, H. (2018). Integrating both visual and audio cues for enhanced video caption. In *Proceedings of the AAAI*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hsu, C.-C., Chen, Y.-H., Chen, Z.-Y., Lin, H.-Y., Huang, T.-H., and Ku, L.-W. (2019). Dixit: Interactive visual storytelling via term manipulation. In *The World Wide Web Conference*, pages 3531–3535.

Hu, J., Cheng, Y., Gan, Z., Liu, J., Gao, J., and Neubig, G. (2020). What makes a good story? designing composite rewards for visual storytelling. In *AAAI*, pages 7969–7976.

Huang, L., Wang, W., Chen, J., and Wei, X.-Y. (2019a). Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4634–4643.

Huang, Q., Gan, Z., Celikyilmaz, A., Wu, D., Wang, J., and He, X. (2019b). Hierarchically structured reinforcement learning for topically coherent visual story generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8465–8472.

Huang, T.-H., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., et al. (2016). Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

Kim, G., Moon, S., and Sigal, L. (2015). Ranking and retrieval of image sequences from multiple paragraph queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1993–2001.

Kim, T., Heo, M.-O., Son, S., Park, K.-W., and Zhang, B.-T. (2018). Glac net: Glocal attention cascading networks for multi-image cued story generation. *arXiv preprint arXiv:1805.10973*.

Kojima, A., Tamura, T., and Fukunaga, K. (2002). Natural language description of human activities from video images based on concept hierarchy of actions. *IJCV*, 50(2):171–184.

Krishnamoorthy, N., Malkarnenkar, G., Mooney, R. J., Saenko, K., and Guadarrama, S. (2013). Generating natural-language video descriptions using text-mined knowledge. In *Proceddings of the AAAI*, volume 1, page 2.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Li, G., Zhu, L., Liu, P., and Yang, Y. (2019a). Entangled transformer for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8928–8937.

Li, J., Shi, H., Tang, S., Wu, F., and Zhuang, Y. (2019b). Informative visual storytelling with cross-modal rules. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2314–2322.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Liu, S., Ren, Z., and Yuan, J. (2020). Sibnet: Sibling convolutional encoder for video captioning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 1–1.

Melis, G., Kočiskỳ, T., and Blunsom, P. (2020). Mogrifier lstm. In *Proceedings of the ICLR*.

Mogadala, A., Shen, X., and Klakow, D. (2020). Integrating image captioning with rule-based entity masking. *arXiv preprint arXiv:2007.11690*.

Nayyer, A., Mian, A., Liu, W., Gilani, S. Z., and Shah, M. (2019). Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, 52(6):1–37.

Pan, B., Cai, H., Huang, D.-A., Lee, K.-H., Gaidon, A., Adeli, E., and Niebles, J. C. (2020). Spatio-temporal graph for video captioning with knowledge distillation. In *IEEE CVPR*.

Pan, Y., Yao, T., Li, H., and Mei, T. (2017). Video captioning with transferred semantic attributes. In *IEEE CVPR*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Park, C. C., Kim, Y., and Kim, G. (2017). Retrieval of sentence sequences for an image stream via coherence recurrent convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):945–957.

Pei, W., Zhang, J., Wang, X., Ke, L., Shen, X., and Tai, Y.-W. (2019). Memory-attended recurrent network for video captioning. In *Proceedings of the IEEE CVPR*, pages 8347–8356.

Phukan, B. B. and Panda, A. R. (2020). An efficient technique for image captioning using deep neural network. *arXiv preprint arXiv:2009.02565*.

Shen, X., Liu, B., Zhou, Y., Zhao, J., and Liu, M. (2020). Remote sensing image captioning via variational autoencoder and reinforcement learning. *Knowledge-Based Systems*, page 105920.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Wang, J., Fu, J., Tang, J., Li, Z., and Mei, T. (2018a). Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Wang, J., Wang, W., Huang, Y., Wang, L., and Tan, T. (2018b). M3: Multimodal memory modelling for video captioning. In *Proceedings of the IEEE CVPR*, pages 7512–7520.

Wang, J., Wang, W., Wang, L., Wang, Z., Feng, D. D., and Tan, T. (2020). Learning visual relationship and context-aware attention for image captioning. *Pattern Recognition*, 98:107075.

Wang, X., Chen, W., Wang, Y.-F., and Wang, W. Y. (2018c). No metrics are perfect: Adversarial reward learning for visual storytelling. *arXiv preprint arXiv:1804.09160*.

Wang, X., Chen, W., Wu, J., Wang, Y.-F., and Yang Wang, W. (2018d). Video captioning via hierarchical reinforcement learning. In *Proceedings of the IEEE CVPR*, pages 4213–4222.

Wiessner, P. W. (2014). Embers of society: Firelight talk among the ju/'hoansi bushmen. *Proceedings of the National Academy of Sciences*, 111(39):14027–14035.

Xiong, Y., Dai, B., and Lin, D. (2018). Move forward and tell: A progressive generator of video descriptions. In *Proceedings of the ECCV*, pages 468–483.

Xu, J., Yao, T., Zhang, Y., and Mei, T. (2017). Learning multimodal attention lstm networks for video captioning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 537–545.

Yan, C., Tu, Y., Wang, X., Zhang, Y., Hao, X., Zhang, Y., and Dai, Q. (2019). Stat: spatial-temporal attention mechanism for video captioning. *IEEE Trans. on Multimedia*, 22:229–241.

Yang, X., Tang, K., Zhang, H., and Cai, J. (2019). Auto-encoding scene graphs for image captioning. In *Proceedings of IEEE CVPR*.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., and Courville, A. (2015). Describing videos by exploiting temporal structure. In *Proceedings of the IEEE ICCV*, pages 4507–4515.

Yu, H., Wang, J., Huang, Z., Yang, Y., and Xu, W. (2016). Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE CVPR*, pages 4584–4593.

Zhang, J. and Peng, Y. (2020). Video captioning with object-aware spatio-temporal correlation and aggregation. *IEEE Trans. on Image Processing*, 29:6209–6222.

Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., and Zha, Z.-J. (2020). Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE/CVF CVPR*.

Zheng, Q., Wang, C., and Tao, D. (2020). Syntax-aware action targeting for video captioning. In *Proceedings of the IEEE/CVF CVPR*.