Boosting Video Captioning with Dynamic Loss Network

Nasib Ullah  
Electronics and Communication Sciences Unit  
Indian Statistical Institute  
Kolkata, India  
Email: nasibullah104@gmail.com

Partha Pratim Mohanta  
Electronics and Communication Sciences Unit  
Indian Statistical Institute  
Kolkata, India  
Email: ppmohanta@isical.ac.in

Abstract—Video captioning is one of the challenging problems at the intersection of vision and language, having many real-life applications in video retrieval, video surveillance, assisting visually challenged people, Human-machine interface, and many more. Recent deep learning based methods [1]–[3] have shown promising results but are still on the lower side than other vision tasks (such as image classification, object detection). A significant drawback with existing video captioning methods is that they are optimized over cross-entropy loss function, which is uncorrelated to the de facto evaluation metrics (BLEU, METEOR, CIDER, ROUGE). In other words, cross-entropy is not a proper surrogate of the true loss function for video captioning. To mitigate this, methods like REINFORCE, Actor-Critic, and Minimum Risk Training (MRT) have been applied but have limitations and are not very effective. This paper proposes an alternate solution by introducing a dynamic loss network (DLN), providing an additional feedback signal that reflects the evaluation metrics directly. Our solution proves to be more efficient than other solutions and can be easily adapted to similar tasks. Our results on Microsoft Research Video Description Corpus (MSVD) and MSR-Video to Text (MSR-VTT) datasets outperform previous methods.

I. INTRODUCTION

Video captioning is the task of describing the content in a video in natural language. With the explosion of sensors and the internet as a data carrier, automatic video understanding and captioning have become essential. It can be applied in many applications such as video surveillance, assisting visually challenged people, video retrieval, and many more. Despite having many applications, jointly modeling the spatial appearance and temporal dynamics makes it a difficult task.

Motivated by machine translation [4] and image captioning [5], [6], the encoder-decoder architecture has been adapted for the video captioning task [1], [2], [7]–[9]. On the encoder side, different visual features are extracted using 2D and 3D convnets. The encoder’s combined visual features are sent to the decoder to generate the caption, one word at a time. So basically, the decoder is a conditional language model, and a variant of recurrent neural networks (LSTM, GRU) is the most popular and successful. Recent improvements on the encoder-decoder baseline have happened in mainly three areas: (i) incorporation of better visual feature extraction modules at the encoder side, (ii) addition of external language models to guide the decoder, (iii) better frame selection strategy. Despite the improvements, a potential drawback with these methods is that the training signal does not align with the standard evaluation metrics such as BLEU [10], METEOR [11], ROUGE-L [12], CIDER [13]. As a result, even low
training and validation loss can lead to poor metric scores and vice versa, as shown in Fig.1(b). Furthermore, direct optimization over metric function is not possible due to the non-differentiable nature of the network. Alternate solutions from Reinforcement learning (REINFORCE, Actor-Critic) and Minimum Risk Training (MRT) have been applied to machine translation and image captioning. However, they have not proved to be very successful in the case of video captioning. To this end, we propose a dynamic loss network (DLN), a transformer-based model that approximates metric function and is pre-trained on external data using a self-supervised setup. Although the proposed DLN can be utilized to approximate any metric function, in our case, we approximate the BLEU, METEOR, and CIDER scores. Once trained, the DLN can be used with the video captioning model in an end-to-end manner, as shown in Fig.1(a).

Finally, we demonstrate that the feedback signals from our proposed model align with the evaluation metric, as shown in Fig.1(c).

II. RELATED WORK

A. Video Captioning.

The main breakthrough in video captioning happened with the inception of encoder-decoder based sequence to sequence models. The encoder-decoder framework for video captioning was first introduced by MP-LSTM \cite{7}, which uses mean pooling over-frame features and then decodes caption by LSTM. Although MP-LSTM \cite{7} outperformed its predecessors, the temporal nature of the video was first modeled by S2VT \cite{11} and SA-LSTM \cite{8}. The former shares a single LSTM model and forces it to capture better representation. Moreover, from the DLN back propagates through the encoder-decoder setup, whereas at the second stage, the trained DLN is used and is pre-trained on external data using a self-supervised transformer-based model that approximates metric function and is pre-trained on external data using a self-supervised setup. Although the proposed DLN can be utilized to approximate any metric function, in our case, we approximate the BLEU, METEOR, and CIDER scores. Once trained, the DLN can be used with the video captioning model in an end-to-end manner, as shown in Fig.1(a).

B. Training on evaluation metric function.

There are mainly three approaches to optimize the sequence to sequence model on the non-differentiable objective function: (i) Ranzato et al. \cite{19} use the REINFORCE algorithm \cite{20} to train an image captioning model directly on BLEU score and Rennie et al. \cite{21} use the Actor-critic method \cite{22}. Both methods use the reward signal, but these methods are not applicable for video captioning due to the sparse nature of the reward. (ii) Optimization on differentiable lower bound where Zhukov et al. \cite{23} propose a differentiable lower bound of expected BLEU score and Casas et al. \cite{24} reported poor training signal corresponding to their formulation of differentiable BLEU score \cite{10}. (iii) Shi Qi Sheng et al. \cite{25} use Minimum risk training (MRT) instead of Maximum likelihood estimation for neural machine translation, and Wang et al. \cite{26} shows Minimum Risk Training (MRT) helps in reducing exposure bias. Unlike previous works, we leverage successful Transformer based pre-trained models to approximate the evaluation metrics.

III. METHOD

Our proposed method follows a two-stage training process. At the first stage, the DLN is trained in a self-supervised setup, whereas at the second stage, the trained DLN is used along with the existing video captioning model. The entire process flow is in the Fig.2. During the second stage, the loss from the DLN back propagates through the encoder-decoder model and forces it to capture better representation. Moreover, the proposed loss network can be combined with different encoder-decoder architectures for video captioning. Below we describe each component of our model.

A. Visual Encoder

We uniformly sample N frames \(\{f_i\}_{i=1}^{N}\) and clips \(\{c_i\}_{i=1}^{N}\) from a given input video, where each \(c_i\) is a series of clips surrounding frame \(f_i\). We extract appearance features \(\{a_i\}_{i=1}^{N}\) and motion features \(\{m_i\}_{i=1}^{N}\) using pre-trained 2D convnets \cite{27}, \(Phi_a\) and \(3D\) convnets \cite{28}, \(Phi_m\), with \(a_i = Phi_a(f_i)\) and \(m_i = Phi_m(c_i)\), respectively. Apart from appearance \(\{a_i\}_{i=1}^{N}\) and motion \(\{m_i\}_{i=1}^{N}\), we extract object characteristics \(\{o_i\}_{i=1}^{N}\) through a pre-trained object detection module \(Phi_o\), where \(o_i = Phi_o(f_i)\). We select prominent items from each frame based on the objectness threshold v and average their features. The appearance and motion characteristics aid in comprehending the video’s global context and motion information. By contrast, object characteristics are more localized, which aids in the comprehension of fine-grained information.

B. Dynamic Loss Network (DLN)

As shown in Fig.1(a), the proposed DLN is built on top of the encoder-decoder and provides an additional training signal aligned with the evaluation metric. The proposed DLN approximates the evaluation metric BLEU \cite{10}, METEOR \cite{11}, and CIDER \cite{13}, which involves mapping from a pair of sentences to numerical values. Motivated by the tremendous success in vision and natural language processing (NLP),
a pre-trained transformer network \cite{29,32} is used as the backbone for the proposed DLN.

The training of the DLN is achieved in a self-supervised manner. The training data and its ground truth to train the DLN (Fig. 2(a)) are generated following two strategies: (i) we take MSCOCO \cite{33} caption corpus and perturb each sentence randomly with a p% probability to generate (candidate \(C_i\), reference \(R_i\)) pair. For the perturbation, deletion and swapping are done over the word(s). (ii) we train a standard encoder-decoder based video captioning model and gather the predicted and ground truth caption as (candidate, reference) pair at different epochs on MSVD \cite{34} data. In both cases, ground truth (BLEU, METEOR, and CIDER) is generated using the NLTK \cite{35} library and the COCO evaluation \cite{36} server.

The self-attention layer in the transformer network (to be more specific, transformer network with the word as input) calculates the attention score between words. This characteristic makes the transformer network \cite{29} a natural choice to model the metric score function (since BLEU, METEOR, and CIDER are precision and recall based formulas on the n-gram overlap). Although BERT \cite{30} and GPT \cite{31} are state-of-the-art pre-trained transformer architecture, they are not suitable to model metric scores due to subword input tokenization. Instead, we use TransformerXL \cite{32} architecture, which works with standard word input (similar to the LSTM decoder). A regression head has been added on top of the standard TransformerXL \cite{32} network and trained by minimizing the mean square loss between the true and predicted BLEU, METEOR, and CIDER values. The output of DLN is,

\[ t_i = W \Upsilon(C_i, R_i) + b \]  

where, \( t_i = (t_{i,BLEU}^i, t_{i,METEOR}^i, t_{i,CIDER}^i) \), \( \Upsilon \) is transformerXL model, \( W \) and \( b \) are the learnable parameters corresponding to regression head. \( R, C \) are reference and candidate sentences, respectively.

Once trained, the DLN is combined with the standard encoder-decoder network at the second stage of training. The proposed DLN is applied only at the training stage, so there is no run-time overhead during inference. As shown in Fig. 2(b), the DLN takes inputs from the output of the decoder and ground truth caption. During the backward pass, the output....
value of DLN is added to cross-entropy loss, and the model is trained on the combined loss function.

C. Language Decoder

The decoder generates the caption word by word based on the features obtained from the visual encoder. A recurrent neural network is utilized as the backbone of the decoder because of its superior temporal modeling capability. In the proposed system, the decoder is designed using LSTM [7], whose hidden memory at time step \( t \) can be expressed as

\[
h_t = LSTM(C_t, h_{t-1})
\]

Where \( C_t \) is the concatenation of appearance, motion, and object features from the visual encoder and \( h_{t-1} \) is the hidden memory of time step \( t-1 \). To predict the word probability, a linear layer followed by a Softmax layer is added on top of the hidden layers of the LSTM.

\[
P(s_t|V, s_1, s_2, ..., s_{t-1}) = Softmax(V_t h_t + b_h)
\]

where \( s_t \) is the \( t \)th word in the caption and \( V \) and \( b \) are the learnable parameters and biases, respectively.

D. Parameter Learning

Along with the typical cross-entropy loss, we train our model with two extra losses: Loss from DLN and Coherent loss.

1) Language Decoder: The cross-entropy or negative log-likelihood function is the typical loss function for an encoder-decoder based video captioning model. For a mini-batch, the loss can be expressed as

\[
L_{LD} = -\sum_{i=1}^{B} \sum_{t=1}^{T} \log p(s_t|V, s_1, s_2, ..., s_{t-1}; \theta)
\]

where \( \theta \) is learnable parameters, \( V \) is the video feature, \( s_t \) is the \( t \)th word in the sentence of length \( T \), and \( B \) is the mini-batch size.

2) DLN Loss: The proposed DLN works in two stages. We train the DLN to predict BLEU, METEOR, and CIDER scores first. We use the Mean square error loss function as the objective for this task, and for a mini-batch, it can be expressed as,

\[
L_{DLN} = \sum_{i=1}^{B} [\lambda_1^i(y_i^{BLEU} - \hat{y}_i^{BLEU}) + \lambda_2^i(y_i^{METEOR} - \hat{y}_i^{METEOR}) + \lambda_3^i(y_i^{CIDER} - \hat{y}_i^{CIDER})]
\]

where, \( y_i \) is the ground truth and \( \hat{y}_i \) is the model prediction. \( \lambda_1, \lambda_2, \lambda_3 \) are hyperparameters to control the relative importance of three different losses.

The DLN predicts BLEU, METEOR, and CIDER score at the second stage and uses it to optimize the encoder-decoder model. For a mini-batch, the loss is

\[
L_{DLN} = -\sum_{i=1}^{B} [\lambda_1 y_i^{BLEU} + \lambda_2 y_i^{METEOR} + \lambda_3 y_i^{CIDER}]
\]

where, \( y_i^{BLEU}, y_i^{METEOR}, y_i^{CIDER} \) are the predicted BLEU, METEOR and CIDER scores from the DLN respectively and \( \lambda_1, \lambda_2, \lambda_3 \) are the hyperparameters.

3) Coherent Loss: A video’s successive frames are exceedingly repetitious. As a result, the encoding of subsequent frames should be comparable. We use the coherence loss to constrain subsequent frames’ embeddings to be comparable. Coherent loss has been used before to normalise attention weights [2]; however, unlike Pei at al. [2], we use the coherence loss to appearance, motion, and object aspects. For a mini-batch, the total coherence loss is,

\[
L_C = \lambda_f L_f^o + \lambda_{mc} L_{mc}^o + \lambda_a L_a^o + \lambda_{ac} L_{ac}^o
\]

where \( \lambda_f, \lambda_{mc}, \lambda_a \) and \( \lambda_{ac} \) are hyperparameters corresponding to appearance coherent loss \( L_f^o \), motion coherent loss \( L_{mc}^o \), object coherent loss \( L_a^o \) and attention coherent loss \( L_{ac}^o \) respectively.

The individual coherent losses are calculated as,

\[
\Psi(f) = \sum_{i=1}^{B} \sum_{t=1}^{T} \sum_{n=2}^{N} |f_{n,t}^{(i)} - f_{n-1,t}^{(i)}|
\]

At the early training phase, cross entropy acts as a better training signal, so we rely more on cross entropy loss. On the other hand, we rely more on loss from the proposed loss network at the later phase of training. The total loss for a mini-batch is

\[
L = L_{LD} + L_{DLN} + L_C
\]

IV. EXPERIMENTS AND RESULTS

We have conducted experiments to evaluate the proposed DLN-based video captioning performance on two benchmark datasets: Microsoft Research-Video to Text (MSRVTT) [42] and Microsoft Research Video Description Corpus (MSVD) [34]. In addition, We have compared the performance of our method with the state-of-the-art video captioning methods. Adding DLN provided significant gain to the captioning performance in all metrics.

A. Datasets

1) MSVD: MSVD contains open domain 1970 Youtube videos with approximately 40 sentences per clip. Each clip contains a single activity in 10 seconds to 25 seconds. We have followed the standard split [2, 7, 8] of 1200 videos for training, 100 for validation, and 670 for testing.

2) MSRVTT: MSRVTT is the largest open domain video captioning dataset with 10k videos and 20 categories. Each video clip is annotated with 20 sentences, resulting in 200k video-sentence pairs. We have followed the public benchmark splits, i.e., 6513 for training, 497 for validation, and 2990 for testing.
TABLE I
PERFORMANCE COMPARISON ON MSVD AND MSR-VTT BENCHMARKS. B4, M, R, AND C DENOTE BLEU-4, METEOR, ROUGE_L, AND CIDER, RESPECTIVELY.

B. Implementation Details

We have uniformly sampled 28 frames per video and extracted 1024D appearance features from Vision Transformer [27], pre-trained on ImageNet [43]. The motion features are 2048D and extracted using C3D [28] with ResNeXt-101 [44] backbone and pre-trained on Kinetics-400 dataset. We use Faster-RCNN [45] pre-trained on MSCOCO [33] for object feature extraction. Appearance, motion, and object features are projected to 512D before sending to the decoder. At the decoder end, the hidden layer and the size of the word embedding are both set as 512D. The dimension of the attention module is set to 128D. All the sentences longer than 30 words are truncated, and the vocabulary is built by words with at least 5 occurrences. For the DLN, we use 16 multi-head and 18 layers TransformerXL [32] pre-trained on WikiText-103. A regression head composed of three fully connected (FC) layers is added on the top of the TransformerXL [32]. During both stages of training, the learning rate for DLN and the end-to-end video captioning model is set to 1e-4. Adam [46] is employed for optimization. The model selection is made using the validation set performance. The greedy search is used for the caption generation at the test time. The coherent loss weights $\lambda_{ac}$, $\lambda_{fc}$, $\lambda_{mc}$, and $\lambda_{oc}$ are set as 0.01, 0.1, 0.01, and 0.1, respectively. All the experiments are done in a single Titan X GPU.

C. Quantitative Results

We have compared our proposed model with the existing video captioning models on MSVD and MSRVTT datasets, as shown in Table I. All four popular evaluation metrics, including BLEU, METEOR, ROUGE_L, and CIDER, are reported. From Table I, we can see that our proposed method significantly outperforms other methods, especially in the CIDER score. It is to be noted that CIDER is specially designed to evaluate captioning tasks. Compared to current methods (ORG-TRL [3], STG-KD [15], SAAT [41]), which uses more complex object-relational features, our method only takes mean object localization features for simplicity and to prove the effectiveness of the DLN.

D. Ablation Studies

In order to validate the effectiveness of the proposed DLN and prove that improvement is not because of the other components of the model, we perform ablation studies. We added the DLN on top of the methods mentioned in Table II under the same settings provided by the original paper’s authors. We report the METEOR and CIDER scores with and without the DLN.

TABLE II
ABLATION STUDIES ON MSVD BENCHMARK. M AND C DENOTE METEOR AND CIDER RESPECTIVELY.
without the DLN on MSVD dataset. From Table II we can see that adding the DLN has significantly boosted performance.

| Models          | RL M | C     | MRT M | C     | DLN M | C     |
|-----------------|------|-------|-------|-------|-------|-------|
| SA-LSTM [8]     | 32.1 | 76.7  | 32.3  | 76.1  | 33.1  | 77.2  |
| RecNet [9]      | 34.3 | 81.1  | 34.1  | 80.7  | 34.4  | 81.3  |
| M3 [14]         | 33.7 | -     | 33.4  | -     | 34.9  | -     |
| PickNet [17]    | 33.5 | 77.8  | 33.3  | 77.1  | 34.5  | 78.7  |
| MARN [2]        | 35.4 | 93.5  | 35.0  | 92.5  | 35.7  | 93.4  |

TABLE III
ABLATION STUDIES ON MSVD BENCHMARK. M AND C DENOTE METEOR AND CIDER RESPECTIVELY.

The comparison of the performance of the DLN with its competitors on direct metric training is shown in Table III. The experiments are done on the above-mentioned methods under the original settings for a fair comparison. Table III shows that our method outperforms its other counterparts. We report METEOR and CIDER scores for all the comparisons since these two are the most important metric to evaluate captioning tasks.

E. Study on the training of the DLN.

The training of the DLN is performed to predict BLEU, METEOR, and CIDER. When it comes to ROUGE modeling, DLN is not as effective as other measures. Also, the signal from ROUGE is not helpful to boost the model performance. The novel idea of the DLN is proposed in this paper, so no benchmark results are available for this task. Hence, the qualitative analysis is performed by comparing histograms of the ground truth and the predicted values on the test set, as shown in Fig. 3. We have given the BLEU results, whereas the METEOR and CIDER stage-1 training outcomes are also similar.

F. Qualitative Results

The Fig. 4 shows the captions generated by our model and MARN [2]. From the figure, we can see that our proposed model performs better than MARN in detecting objects and actions. Also, the captions generated by our model are more grammatically sound.

Groundtruth: a cat and turtle are playing.
MARN: a cat is playing with a cat.
Ours: a cat is laying next to a turtle.  

(a)

Groundtruth: a woman and a man are playing a drama on a stage.
MARN: a women are dancing.
Ours: three people perform a drama.  

(b)

Groundtruth: a deer is on a woman’s back.
MARN: a man is riding a bicycle.
Ours: two men are walking in the snow.  

(c)

Groundtruth: a man opening a box with a knife.
MARN: a man is playing a man.
Ours: a man is opening a parcel.  

(d)

Fig. 4. Qualitative comparison of Captions generated by our model and MARN [2].

V. CONCLUSION

This work addresses the training signal evaluation metric alignment mismatch problem of existing video captioning models and proposes a dynamic loss network (DLN), which models the evaluation metric under consideration. The training is performed in two stages, and the experimental results on the benchmark datasets show superior performance than current state-of-the-art models. Also, our approach shows better performance than other existing non-differentiable training strategies for video captioning and can be easily adaptable to similar tasks. Future studies could investigate the effectiveness of our method on other tasks such as image captioning and machine translation.
Transformers for image recognition at scale," in 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018. Salt Lake City, UT, USA. June 18-22, 2018. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 6546–6555.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need” in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.

J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., “Language models are unsupervised multitask learners,” OpenAI blog, vol. 1, no. 8, p. 9, 2019.

Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," in Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019. Florence, Italy, July 28 - August 2, 2019, Volume 1: Long Papers, A. Korhonen, D. R. Traum, and L. Márquez, Eds. Association for Computational Linguistics, 2019, pp. 2978–2988. [Online]. Available: https://doi.org/10.18653/v1/p19-1285

T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: common objects in context,” in Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, ser. Lecture Notes in Computer Science, D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8693. Springer, 2014, pp. 740–755. [Online]. Available: https://doi.org/10.1007/978-3-319-10602-1_48

D. L. Chen and W. B. Dolan, “Collecting highly parallel data and evaluation server,” CoRR, vol. abs/1504.00325, 2015. [Online]. Available: http://arxiv.org/abs/1504.00325

N. Xue, “Steven bird, evan klein and edward loper. Natural Language Processing with Python. o’reilly media, inc 2009. ISBN: 978-0-596-51649-9,” Nat. Lang. Eng., vol. 17, no. 3, pp. 419–424, 2011. [Online]. Available: https://doi.org/10.1017/S1351324910003060

X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, “Microsoft COCO captions: Data collection and evaluation server,” CoRR, vol. abs/1504.00325, 2015. [Online]. Available: http://arxiv.org/abs/1504.00325

H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, “Video paragraph captioning using hierarchical recurrent neural networks,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA. June 27-30, 2016. IEEE Computer Society, 2016, pp. 4584–4593. [Online]. Available: https://doi.org/10.1109/CVPR.2016.596

J. Song, L. Gao, Z. Guo, W. Liu, D. Zhang, and H. T. Shen, “Hierarchical LSTM with adjusted temporal attention for video captioning,” in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, C. Sierra, Ed. ijcai.org, 2017, pp. 2737–2743. [Online]. Available: https://ojs.aaai.org/index.php/IJCAI/article/view/17581

N. Aafaq, N. Akhtar, W. Liu, S. Z. Gillani, and A. Miian, “Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning,” in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA. June 16-20, 2019. Computer Vision Foundation / IEEE, 2019, pp. 12 487–12 496.

B. Wang, L. Ma, W. Zhang, W. Jiang, J. Wang, and W. Liu, “Controllable video captioning with POS sequence guidance based on gated fusion network,” in 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019. Seoul, Korea (South), October 27 - November 2, 2019. IEEE, 2019, pp. 2641–2650. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00273

Q. Zheng, C. Wang, and D. Tao, “Syntax-aware action targeting for video captioning,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. Computer Vision Foundation / IEEE, 2020, pp. 13 093–13 102.

J. Xu, T. Mei, T. Yao, and Y. Rui, “MSR-VTT: A large video description dataset for bridging video and language,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 2016, pp. 5288–5296. [Online]. Available: https://doi.org/10.1109/CVPR.2016.571

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, 2015. [Online]. Available: https://doi.org/10.1007/s11263-015-0816-y

S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 2017, pp. 5978–5995. [Online]. Available: https://doi.org/10.1109/CVPR.2017.634

S. Ren, K. He, R. B. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2577031

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980