RUC_AIM3 at TRECVID 2019: Video to Text

Yuqing Song, Yida Zhao, Shizhe Chen and Qin Jin
Renmin University of China
{syuqing, zyiday, cszhe1, qjin}@ruc.edu.cn

Abstract

In this paper, we present our solutions in the two sub-tasks of TRECVID 2019 Video to Text Challenge [1]. For both video-text matching and video description generation, it is important to understand videos from multiple modalities and generate video representations with rich semantic information. Therefore, we generate the video representation via fusing multi-modal features, including 2D, 3D visual features and audio features for both two sub-tasks. For matching and ranking, we employ the state-of-the-art video-semantic matching model to retrieve the best sentence with aforementioned multi-modal video features. For video description generation, in order to generate comprehensive, accurate and fluent descriptions for the video, we propose to integrate temporal and semantic attentions for the captioning model, and further boost the caption performance by providing specific fluency and relevancy rewards in reinforcement learning framework. Considering different models are complementary, we propose a late fusion strategy to ensemble different models to improve system generalization abilities.

1 Introduction

The TRECVID video-to-text task aims to describe the video content with a natural language sentence, which is one of the ultimate goals of video understanding. The solutions for such task can be generally divided into two categories: selecting sentences from the corpus through matching [2], and generating sentences by the captioning model [3]. It is important to generate video representations with rich semantic information in order to comprehensively understand the video. Since videos inherently contain multiple modalities, we generate the video representation via fusing the 2D, 3D visual features and audio features for both two sub-tasks.

For matching and ranking, we employ the state-of-the-art video-semantic matching model to match the video and sentences. For video description generation, it is more challenge because there is a huge gap between the video representation and language representation, and it's difficult for captioning model to generate comprehensive, accurate and fluent descriptions based on the video. To reduce such representation gap and capture multi-level aspects in the video, we propose to integrate both temporal and semantic attentions for video captioning. The temporal attention is employed to aggregate action movements in the video, while the semantic attention is employed to enhance video semantic representations.

We employ cross entropy loss to train the baseline video captioning model. In order to further boost captioning performance with respect to language fluency and relevancy aspects, we also fine-tune the model with reinforcement learning (RL). In RL, we not only utilize evaluation metrics such as CIDEr as the reward, but also design two specific reward functions to improve the fluency and relevancy of video descriptions. For the language fluency reward, we employ a language model which is pretrained on the fluent groundtruth video captions to evaluate the fluency score. For the visual relevancy reward, we employ a visual-semantic matching model to evaluate the semantic relevancy between video and generated descriptions.

Considering different models are complementary, we develop multiple captioning models and ensemble them by sentence reranking strategy in order to improve system generalization abilities. Our approaches achieve the best performance in the TRECVID 2019 VTT challenge on both two subtasks.

2 Methodology

In this section, we will introduce our models for two subtasks in details.

2.1 Description Generation

For the description generation subtask, our model is composed of four main parts, video semantic encoding, description generation with temporal and semantic attentions, reinforcement learning optimization and ensemble from multi-aspects.

Video Semantic Encoding. In order to comprehensively encode videos, we extract two types of video features for temporal and semantic attention respectively. In the temporal branch, we represent the video as a sequence of segment-level multi-modal features $V^T = \{v_0^t, \dots, v_n^t\}$. Each segment-level feature is the concatenation of video features from three modalities, including 2D (Resnext101), 3D (I3D) and audio (VGGish). In the semantic branch, in order to reduce the representation gap between video and language, we predict several visual concepts based on the video temporal feature V^T to enhance video semantic representations. The word embedding vector of the predicted concepts can be used as the video semantic feature $V^S = \{w_0^c, \dots, w_m^c\}$.

Description Generation with Temporal and Semantic Attentions. Based on the encoded video features, we can generate video descriptions with temporal and semantic attentions. The captioning model learns to focus on the relevant temporal frames and concepts to generate the word. The temporal and semantic context feature via attention mechanism to predict the t-th word can be represented as:

$$ctx_t^T = softmax(h_{t-1}W^T(V^T)^T)V^T \tag{1}$$

$$ctx_t^S = softmax(h_{t-1}W^S(V^S)^T)V^S \tag{2}$$

$$ctx_t = [ctx_t^T, ctx_t^S] \tag{3}$$

Therefore, the input of LSTM decoder in each time step is the concatenation of previous word embedding w_{t-1} and the context feature ctx_t.

$$h_t = f([w_{t-1}, ctx_t], h_{t-1}; \theta_d) \text{ for } t = 1, \dots, N_w \tag{4}$$

Reinforcement Learning Optimization. To generate fluent and accurate descriptions, we fine-tune the captioning model through reinforcement learning with fluency and visual relevancy rewards. We utilize a pre-trained language model to evaluate the language fluency of generated sentences. Disfluent sentences will be generated in high perplexities by the language model. Therefore, the fluency reward for the description $s = \{w_0, \dots, w_n\}$ can be represented as:

$$r_{flc}(s) = \frac{1}{n}\sum_{j=1}^{n} \log P(w_j | w_{0:j-1}; \theta_{lm}) \tag{5}$$

To further boost the caption performance on visual relevancy, we utilize the visual-semantic matching model in matching and ranking subtask to evaluate the relevancy of generated captions. The visual-semantic matching model is based on a cross-modal joint embedding space, and the embedding vectors of video and caption can be close to each other in this space if they are visual relevant. Therefore, we can utilize the cosine similarity of embedding vectors as the relevancy reward.

$$r_{rlv}(s) = cosine\_sim(E_v(v), E_c(s)). \tag{6}$$

Ensemble from Multi-aspects. Since different models are complementary, we can get the wisdom of crowd by captions reranking. We produce various captioning models with small component differences, such as the vanilla encoder-decoder captioning model, captioning model with different attentions, captioning model with reinforcement learning. Then, the aforementioned language model and visual semantic matching model can be used to evaluate the language fluency and visual relevancy of captions generated by different models. We can rerank these captions by the weighted sum of fluency score and relevancy score, and choose the best description for the video.

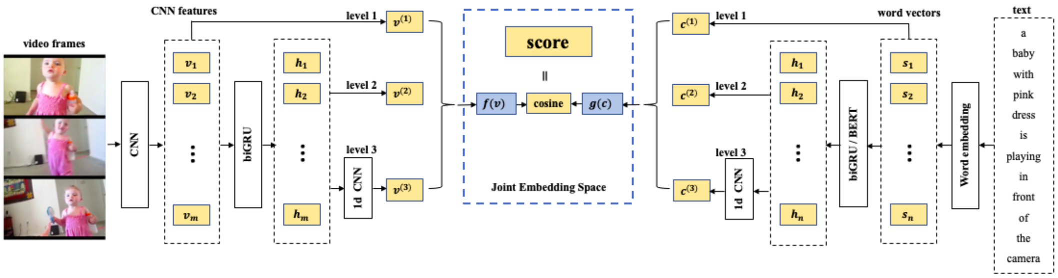

Figure 1: Illustration of the visual-semantic matching model for the matching and ranking subtask. There are three encoding branches for both video and sentence.

2.2 Matching and Ranking

For the matching and ranking subtask, our model is built based on the state-of-the-art dual encoding model [4]. Given a sequence of input features (video frame-level features or word embeddings of sentence), three branches are used to encode global, temporal and local information respectively. Figure 1 illustrates the framework of our visual-semantic matching model.

For the video encoder, the three encoding branches are the mean pooling of frame-level video features, bidirectional GRU (biGRU) encoding, 1-D CNN encoding of the output of second branch respectively. For the text encoder, we utilize the mean pooling of word embeddings which is initialized by GloVe, biGRU encoding and 1-D CNN encoding as the three branches. Furthermore, considering the excellent ability of BERT [5] to encode the long-context information, we also explore using BERT to replace the biGRU of the text encoder and fine-tune its last layer.

After encoding both video and sentence by the three branches, we concatenate the encoded features from three branches and map them into the joint embedding space through a fully connected layer followed with batch normalization and tanh activation. We employ the state-of-the-art ranking loss [2] which focuses on the hard negative samples to train this matching model.

3 Experiments

We employ the TGIF [6], MSRVTT [7], VATEX [8], TRECVID VTT 2016 & 2017 video captioning datasets as our training set, and TRECVID VTT 2018 as our validation set for both two subtasks. For the video representation, we extract features from different modalities including Resnext101, I3D, and VGGish.

3.1 Description Generation

For description generation subtask, we develop multiple captioning models, including the vanilla encoder-decoder captioning model [3], temporal attention model (TA), semantic attention model (SA), temporal and semantic attention model (TSA) and captioning models with reinforcement learning. To study the impact of training with different datasets on the model performance, we compare the performance of vanilla captioning model trained on various datasets. Table 1 shows that the captioning model trained on more datasets can be more generalized and achieves better performance.

Table 1: Performance comparison on TRECVID VTT 2018 of the vanilla model trained on different datasets.

Datasets	Bleu4	Meteor	Rouge	Cider
TGIF	11.59	13.37	32.54	16.55
TGIF+TRECVID	12.2	14.56	33.39	16.70
TGIF+TRECVID+MSRVTT	12.43	14.74	33.40	17.63
TGIF+TRECVID+MSRVTT+VATEX	**12.60**	**14.82**	**33.57**	**18.29**

Table 2: Performance comparison of different captioning models on TRECVID VTT 2018.

Datasets	Bleu4	Meteor	Rouge	Cider
Vanilla	12.60	14.82	33.57	18.29
TA	12.46	14.82	32.84	19.99
SA	13.20	14.96	33.64	19.74
TSA	12.83	15.13	33.31	20.70
Vanilla+Cider	12.90	14.84	33.25	20.42
TSA+fluency+relevancy	13.02	14.95	33.59	20.99
Rerank of all	**14.44**	**15.55**	**34.51**	**23.44**

We also compare the performance of different captioning models in table 2 to make ablation studies. The comparison in the first four lines shows that both the temporal and semantic attention can improve the performance of vanilla encoder-decoder captioning model, because the attention mechanism teaches the model to focus on specific regions (temporal frames or concepts) to generate each word. Furthermore, the temporal attentive model and semantic attentive model are complementary with each other since they focus on different aspects in the video. Therefore, combining the temporal and semantic attentions achieves additional gains on the model performance. The comparison between TSA and TSA+fluency+relevancy model shows the effectiveness of self-defined fluency and relevancy rewards. The last line shows a huge improvement on the single captioning model by captions reranking, and it demonstrates the evaluation ability of pretrained language model and visual-semantic matching model.

Finally, we submit four runs as following:

- Run 4: Ensemble of the captioning models trained with cross-entropy loss by captions reranking.
- Run 3: Ensemble of the captioning models trained by reinforcement learning via captions reranking.
- Run 2: The run3 optimized by caption generation length control.
- Run 1: Ensemble of run2 and run3 by captions reranking.

3.2 Matching and Ranking

For matching and ranking subtask, we submit four runs as following:

- Run 4: The basic three branches matching model as aforementioned.
- Run 3: The matching model replacing the biGRU of the text encoder with BERT.
- Run 2: Ensemble of six matching models. Three of them are the run4 with activation differences in the last fc layer, and the other three models are the run 3 with activation differences in the last fc layer.
- Run 1: The run2 with one-to-one matching optimization.

Table 3: Results of TRECVID 2019 VTT matching and ranking subtask.

Ours	SetA	SetB	SetC	SetD	SetE
Run 4	0.572	0.580	0.574	0.579	0.572
Run 3	0.569	0.581	0.575	0.579	0.575
Run 2	0.623	0.635	0.627	0.636	0.630
Run 1	**0.723**	**0.727**	**0.721**	**0.721**	**0.722**

Tabel 3 presents the performances of four submitted runs on TRECVID VTT 2019 test dataset. It shows that the improvement provided by BERT is limited because it's more important to enhance the semantic alignment between videos and sentences than memorize the internal information of sentences for video-sentence matching. The performance of run2 shows that the ensemble of models brings a significant improvement on basic matching model, which demonstrates the complementary of different models. Furthermore, the specific one-to-one matching optimization for this subtask brings more gains through a process of elimination.

4 Conclusions

In this section, we conclude our models for the two subtasks in TRECVID 2019 VTT Challenge. For video description generation, we propose to integrate temporal and semantic attentions for the captioning model to generate video descriptions comprehensively, and further boost the caption performance by providing specific fluency and relevancy rewards in reinforcement learning framework. For matching and ranking subtask, we employ the state-of-the-art visual-semantic matching model and improve the ranking by model ensemble and one-to-one matching optimization. Our approaches achieve the best performance in the TRECVID 2019 VTT challenge on both two subtasks.

5 Acknowledgement

This work was supported by National Natural Science Foundation of China (No. 61772535), and Beijing Natural Science Foundation (No. 4192028) and National Key Research and Development Plan (No. 2016YFB1001202).

References

[1] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, and Georges Quénot. Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search retrieval. In *Proceedings of TRECVID 2019*. NIST, USA, 2019.

[2] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.

[3] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*. IEEE, 2015.

[4] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *CVPR*, 2019.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. 2018.

[6] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016.

[7] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[8] Wang Xin, Wu Jiawei, Chen Junkun, Li Lei, Wang Yuan-Fang, and Wang William Yang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019.