RUCAIM3-Tencent at TRECVID 2022: Video to Text Description

Zihao Yue, Yuqi Liu, Liang Zhang, Linli Yao, Qin Jin[†]
School of Information, Renmin University of China
{yzihao, yuqi657, zhangliang00, linliyao, qjin}@ruc.edu.cn

Abstract

This report presents our solution for the Video to Text Description (VTT) task of TRECVID 2022 (Awad et al., 2022). We leverage a vision-language pre-training model pre-trained on large-scale image-text datasets for video captioning. With an effective pseudo-label-based data augmentation method to expand the fine-tuning data and well-designed re-ranking strategies to automatically select better descriptions from the candidates, we boost our system performance to a new level. Our submission ranks *1st* in all evaluation metrics including BLEU, METEOR, CIDER, SPICE, and STS, and achieves the best CIDEr score of 60.2, a relative improvement of 67.2% over the best result of last year.

1 Introduction

The video-to-text description task requires a model to automatically generate a single-sentence description in natural language for a given video. The dominant approaches use deep neural networks with an encoder-decoder framework, where the encoder encodes the video input into a visual representation and the decoder generates a caption conditioned on the encoder output.

Recently, large-scale vision-language pre-training (VLP) models including UniVL (Luo et al., 2020), Oscar (Li et al., 2020) and CLIP (Radford et al., 2021) have been validated to achieve superior performance on vision-language tasks such as image captioning. From large-scale image-text data covering a wide range of domains, VLP models can learn very effective representations that can benefit various downstream tasks. However, the amount of video-text data is not as large as that of image-text data, which limits the pre-trained models based on video-text data. We therefore consider leveraging image-text pre-trained models for video tasks.

[†]Corresponding author. Code is available at https://github.com/yuezih/BLIP4video.

Specifically, we build our system based on the BLIP model (Li et al., 2022), a VLP model with its large version pre-trained on a bootstrapped dataset with 129M images and paired captions. We also propose our data augmentation strategies to utilize the provided video description data in the VTT challenge, and our post-processing strategies to rerank the generated caption candidates.

2 Related Works

The video description generation task, also known as video captioning, is one of the key benchmarks for visual understanding. Early works adopt pre-designed templates for filling in the recognized visual objects (Kojima et al., 2002). In recent years, sequence-to-sequence models have become mainstream schemes, some of which are based on recurrent neural networks (RNN) (Venugopalan et al., 2015), and later dominated by Transformer-based models (Chen et al., 2018). Zhang et al. (2021a) achieve the multimodal inputs encoding and captions generation with only encoders, using multi-task pre-training and fine-tuning based on next-token language modeling to win the VTT challenge at TRECVID 2021 (Awad et al., 2021).

3 Approach

3.1 BLIP for Video

We propose BLIP for video (**BLIP4video**), a framework that shares the same model structure as BLIP (Li et al., 2022) but supports the input of frame sequence. Concretely, selected frames of a video are sequentially fed into the ViT-based visual encoder of BLIP, to transform the frame sequence into the embedding sequence, which is then concatenated as the visual representation of the video, as shown in Fig. 1. For the video captioning task, the visual representation is injected into the cross-attention layer of the image-grounded text decoder of BLIP for caption generation; for the video-text match-

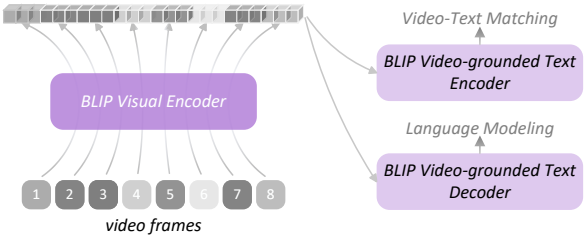

Figure 1: The overall architecture of BLIP4video.

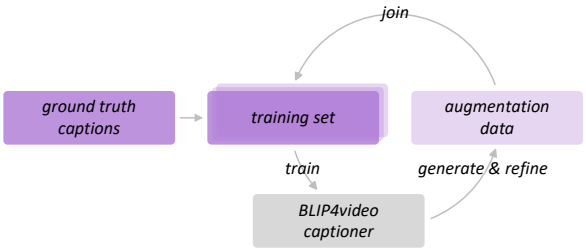

Figure 2: Our data augmentation pipeline.

ing task, the image-grounded text encoder of BLIP calculates a matching score, which is further explained in 3.3 Candidates Re-ranking. In this way, BLIP4video inherits all the parameters of BLIP, supporting the loading of its pre-training weights, which are then fine-tuned on the VTT data.

3.2 Data Augmentation

By fine-tuning on the provided data, the BLIP4video model as a competitive captioner enables the generation of high-quality caption pseudo labels. For each video v_i in the video training set V, we generate 5 captions via beam search decoding as its first pseudo label batch ${p_{v_i}}^1$ to obtain an augmentation collection P^1. To improve the quality of P^1, different from CapFilt in BLIP (Li et al., 2022), which uses a filter to tell whether a text matches an image, we calculate CIDEr scores for sentences in each ${p_{v_i}}^1$ with the given ground truth c_{v_i} as the reference, and set a threshold according to our empirical observation to filter out the ones with low CIDEr scores. Then we merge the refined $P^{1'}$ from P^1 into the training set C, to acquire an augmented dataset, named *Aug1*. With the augmented data as the new training set, we make this a recursive procedure where the newly trained model can additionally scale up the training data, as illustrated in Fig. 2.

3.3 Candidates Re-ranking

We apply post-processing strategies such as candidates re-ranking to further improve our system performance. With a finalized model, we make

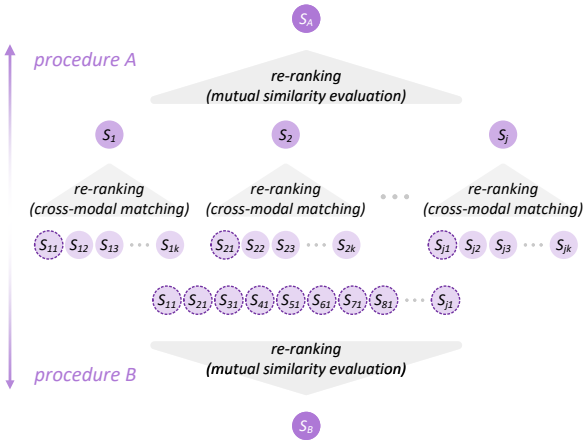

Figure 3: The proposed two re-ranking procedures.

j inferences $I_{1,2,...,j}$ with different randomly selected frames as input, with k sentences $S_{i1,i2,...,ik}$ for each inference (i denotes the inference index) by beam search decoding (top-k beams), to get $j \cdot k$ candidate captions per video. Then, we adopt two measures to estimate the quality of candidate captions for re-ranking, including Cross-modal Matching (CMM) and Mutual Similarity Evaluation (MSE).

Cross-modal Matching requires the BLIP4video model, which is fine-tuned with video-text contrastive (ITC) and video-text matching (ITM) objectives, to calculate matching scores between a video and each of its caption candidates, and these matching scores are used to evaluate the relevance of the video and captions.

Mutual Similarity Evaluation divides n candidates of a video into 1 prediction and $n-1$ references, and a mutual similarity score between the prediction and references is calculated via n-grams matching-based algorithms. We perform re-ranking based on the similarity scores of all candidates to eliminate those differing significantly from other candidates. Following CLIP4Caption++ (Tang et al., 2021), we use the CIDEr metric for n-grams matching evaluation.

We design two procedures, procedure *A* and *B*, as the final re-ranking solutions based on the generated candidates and two proposed measures, as illustrated in Fig. 3. Procedure A conducts intra-inference CMM re-ranking and inter-inference MSE re-ranking while procedure B directly takes the highest-score beam as the intra-inference winner for further inter-inference MSE re-ranking.

Table 1: Details about the data augmentation models. (n) under the datasets refers to the number of captions in the dataset; $C > m$ indicates that the augmentation data is filtered with CIDEr$> m$.

Model					Augmentation Data				
Model Name	Training Data			CIDEr	Aug. Name	Video Source	Caption Num.		
	stage-1	stage-2	stage-3				*all*	$C > 50$	$C > 60$
Aug-Model-1	Ext (574,321)	VTT17-19 (25,121)	VTT17-19 (25,121)	51.8	Aug-1	VTT16-20 (9,443)	47,215	17,939	14,258
Aug-Model-2	-	VTT16-20, Aug-1$_{C>50}$ (55,374)	VTT16-20 (37,435)	52.8	Aug-2	VTT16-20 (9,443)	47,215	18,784	14,767

Table 2: Experimental settings of our 4 submission runs.

Run Name	Model					Re-ranking procedure
	Name	Dataset			CIDEr	
		Training stage-2	Training stage-3	Validation		
run1 run2	Final-Model-1	VTT16-20, Aug1-2$_{C>50}$ (74,158)	VTT16-20 (37,435)	VTT21 (8,385)	53.9	A B
run3 run4	Final-Model-2	VTT16-21, Aug1-2$_{C>60}$ (74,845)	VTT16-21 (45,820)	-	-	A B

Table 3: Validation and submission performance of our 4 runs.

Run Name	Validation Performance				Submission Scores				
	CIDEr	BLEU@4	METEOR	ROUGE_L	CIDEr	BLEU@4	METEOR	SPICE	STS
run1	**54.9**	28.8	**22.4**	46.4	59.4	**13.5**	41.2	18.2	53.0
run2	54.1	**29.5**	22.2	**46.6**	57.5	13.2	40.9	18.0	52.8
run3	-	-	-	-	**60.2**	13.3	**41.5**	**18.4**	**53.4**
run4	-	-	-	-	59.2	**13.5**	41.4	18.3	53.0

4 Experiments

4.1 Implementation Details

We first obtain the final models through 2 rounds of fine-tuning and data augmentation. For the pre-training weights, we choose BLIP (with ViT large) fine-tuned on MSCOCO (Lin et al., 2014) with the image captioning task provided by Li et al.[1]. For the first round, we adopt a 3-stage fine-tuning procedure to get *Aug-Model-1*, first on extended datasets including VTT17-19 (VTT data released from 2017 to 2019), MSRVTT (Xu et al., 2016) and VATEX (Wang et al., 2019) (*stage-1*), then on VTT17-19 only, with language modeling objective to decode texts given videos (*stage-2*) and SCST (Rennie et al., 2017) implemented as in VinVL (Zhang et al., 2021b) (*stage-3*). For the second round, the fine-tuning *stage-1* is dropped and *Aug-1* produced by *Aug-Model-1* is merged into the training set of *stage-2*, as detailed in Table 1. We evaluate the two data-augmentation models on VTT21, based on which we produce caption pseudo label collections *Aug-1* and *Aug-2*, respectively.

With the VTT data and augmentation data, we obtain *Final-Model-1* and *Final-Model-2* for generating final caption results, as detailed in Table 2. For *Final-Model-1*, the fine-tuning process is early stopped according to the validation performance. For *Final-Model-2*, we train it with all development data, using both the training and validation sets, and select checkpoints in the two-stage training process with empirical experience[2]. For both models, our proposed two candidates re-ranking procedures are adopted to get 4 runs for submission.

Models are fine-tuned on 4 NVIDIA RTX A6000 GPU nodes with a batch size of 48 (*stage-1,2*) and 8 (*stage-3*). We adopt AdamW optimizer with the weight decay set to 0.05, and a learning rate ini-

[1] https://github.com/salesforce/BLIP

[2] 5 epochs at training *stage-2* and 3 epochs at training *stage-3*. Note that to make training more stable, we increase the filter threshold in the augmentation processing, in order to control the quality of the augmented data.

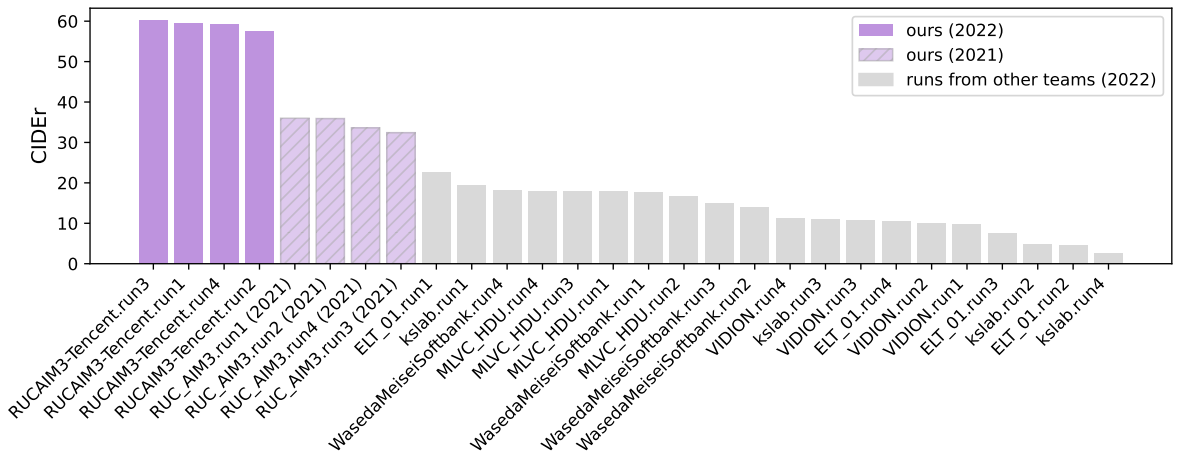

Figure 4: Results comparison of all teams in VTT challenge 2022 and the best results in VTT challenge 2021.

tialized at $2e-6$ (*stage-1,2*) and $1e-7$ (*stage-3*) with cosine scheduling and a reduction to 0 in 10 epochs. For the video input, we get 8 frames of each video with TSN sampling (Wang et al., 2016) during training, which divides the video equally into k segments and selects one frame from each randomly, while with uniform sampling during inference. Frames are resized to 224×224 and divided into 16×16 patches. For texts, the maximum length of each caption is set to 40, and generation length L is retrained as $10 < L < 32$. For re-ranking, 10 inferences are conducted with 5 beams for each video, leading to 50 caption sentences per video.

4.2 Results

The evaluation results of our augmentation models and final models are shown in Table 1 and Table 2, respectively. Our *Final-Model-1* sets a new performance record with the benefit of the pseudo labels generated by augmentation models. The performance of the submitted 4 runs with corresponding re-reranking strategies on the validation set VTT21 and test set VTT22 are shown in Table 3. Re-ranking further improves the performance and both A and B strategies excel in different evaluation metrics respectively. In our submission results (Table 3, right), *run3* and *run4* produced by our final model trained with both training and validation data outperform *run1* and *run2*. Our submission ranks *1st* on all metrics, and a comparison with other teams' and our 2021's results on the CIDEr score is shown in Fig. 4.

We also perform ablation studies on the key components of our system, including data augmentation, SCST at training *stage-3* and candidates re-ranking. Table 4 shows that data augmentation contributes a lot and SCST is also helpful. Both re-ranking procedures are beneficial while procedure *A* works better on CIDEr.

Table 4: Ablation study of key components of our system.

Row	Data Augmentation		Training Strategy	Re-ranking		Perform.
	Aug-1	Aug-2	SCST	A	B	CIDEr
1						48.2
2	✓					52.8
3	✓	✓				53.6
4	✓	✓	✓			53.9
5	✓	✓	✓	✓		**54.9**
6	✓	✓	✓		✓	54.1

5 Conclusion

In this report, we present our solutions for the VTT challenge of TRECVID 2022. We build a strong baseline by fine-tuning a VLP model pre-trained on large-scale image-text data on the VTT task, and demonstrate the effectiveness of our proposed data augmentation and candidate re-ranking strategies. Our systems set a new performance record in the VTT challenge.

6 Limitations

6.1 Methodological Limitations

In our BLIP4video, each frame is encoded separately to a final visual representation, lacking inter-frame dynamics encoding that we believe is important for deeper video understanding. In future

works, we will explore better video representation learning, especially extending the visual understanding capabilities of VLP models from images to videos, or performing video-text pre-training.

6.2 Benchmark

From our experimental analysis, we observe that many videos in the VTT data are informatively static, as shown in Figure 5, implying that an image captioning system may be easily competent to describe these videos. We argue that in a video description challenge, more consideration should be given to properties that distinguish videos from images such as temporal variability (we tentatively name it videoness). A more challenging benchmark could perhaps be constructed by filtering cases with more videoness.

- *An Asian man playing an electronic guitar in an indoor setting.*
- *An Asian guitarist plays his guitar inside a bar with colorful posters on the wall.*
- *An Asian man in a black jacket playing a guitar indoors.*
- *A man wearing a black jacket is playing a guitar indoors.*
- *Young man with black hair wearing black leather jacket plays electric guitar inside a dark room surrounded by posters.*

Figure 5: A data example from the VTT22 dataset. Texts are manually annotated captions.

6.3 Evaluation Metrics

The higher evaluation scores of our system in the VTT Challenge lead us to consider whether the existing official evaluation metrics will continue to support the development of better models. We conduct a shallow evaluation to explore how the ground truth captions (GTs) score on these metrics. We divide the GTs into 5 groups by one caption per video, and then calculate each group's score in turn with the other four groups as the reference, in comparison with our submission (*run3*). As shown in Table 5, captions annotated by human experts score much less than automatically generated ones. It could be because that captions from different annotators are diverse, and a semantically correct description may get low scores due to inconsistency in language with other GTs, while model-generated captions are more general to match with GTs.

To some extent, existing metrics increasingly fail to properly measure the accuracy of generated descriptions, let alone other aspects, including fluency and diversity that are also worth considering. Although some representation learning-based metrics including CLIP score may help, there are still limitations such as poor interpretability. We leave it to future work to explore better evaluation metrics for the video captioning task.

Table 5: Comparison of the ground truth captions and model-generated captions on CIDEr.

Group	GT		Prediction	
	CIDEr	BLEU@4	CIDEr	BLEU@4
1	36.8	16.7	51.3	25.6
2	36.3	16.4	50.4	25.5
3	36.0	16.2	50.1	25.6
4	37.5	16.9	49.8	25.7
5	36.9	17.0	50.9	25.5
avg.	36.7	16.6	50.5	25.6

7 Acknowledgement

This work was partially supported by the National Key R&D Program of China (No. 2020AAA0108600) and the National Natural Science Foundation of China (No. 62072462).

References

George Awad, Asad A Butt, Keith Curtis, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, et al. 2021. Evaluating multiple video understanding and retrieval tasks at trecvid 2021. In *2021 TREC Video Retrieval Evaluation*.

George Awad, Keith Curtis, Asad A. Butt, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, Lukas Diduch, Jeffrey Liu, Yvette Graham, , and Georges Quénot. 2022. An overview on the evaluated video retrieval tasks at trecvid 2022. In *Proceedings of TRECVID 2022*. NIST, USA.

Ming Chen, Yingming Li, Zhongfei Zhang, and Siyu Huang. 2018. Tvt: Two-view transformer network for video captioning. In *Asian Conference on Machine Learning*, pages 847–862. PMLR.

Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. 2002. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhaleving, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mingkang Tang, Zhanyu Wang, Zhaoyang Zeng, Fengyun Rao, and Dian Li. 2021. Clip4caption ++: Multi-clip for video caption. *CoRR*, abs/2110.05204.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liang Zhang, Yuqing Song, and Qin Jin. 2021a. Ruc_aim3 at trecvid 2021: Video to text description. In *Proceedings of TRECVID 2021*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. Vinvl: Making visual representations matter in vision-language models. *CoRR*, abs/2101.00529.