Elyadata TRECVID 2022 VTT Model

Haroun Elleuch
haroun.elleuch@elyadata.com

Fethi Bougares
fethi.bougares@elyadata.com

Ahmed Badri
ahmed.badri@elyadata.com

Abstract

This paper describes Elyadata's first participation in the TRECVID [1] 2022 evaluation. We participated in the Video to Text Description (VTT) task. We experimented with various approaches using a combination of a Vision-Language Pre-trained framework and spatio-temporal Transformer architecture. The evaluation results show that, out of 6 teams, our system achieved fifth place on METEOR, second place on average STS and CIDEr-D and fourth place on BLEU, CIDEr and SPICE.

Keywords: TRECVid, Video Captioning, Video to Text Description.

1 Introduction

The TRECVID [1] VTT task requires proposing a system which automatically generate a single sentence that best describes a given input video using natural language. A total of four runs were submitted, exploring different transformer combinations and types of data. We compared the use of an image captioning model versus the use of multiple frames using spatial features only. Lastly, the introduction of spatio-temporal transformer [2] was explored as an alternative to the spatial frame encoder. The image captioning variant performed the best across all reported evaluation metrics.

This paper is organized as follows. In Section 2, we describe the different approaches adopted for this task. In Section 3, we describe the datasets used to train our models. In Sections 4, 5, and 6, we detail the models underlining these approaches. In Section 7, we describe the model training and parameters used. In Section 8, we discuss the obtained results and in Section 9, we conclude this paper and present future work.

2 Method

To obtain captions from a video clip, as required in the description generation task, three approaches were adopted. The first is centred around the Bootstrapping Language Image Pretraining (BLIP) model [3], taking advantage of its heavy pretraining on image content and excellent results in the field of image-text understanding tasks. The second is an iteration upon the previous approach by using whole clips, instead of a single sampled image, during the training process. Finally, the third approach aimed at using another state-of-the-art model in conjunction with BLIP.

The first approach consisted in creating an image dataset from the provided VTT (Video to Text) dataset and then fine-tune BLIP using it. The performance of the resulting model was then assessed using the internal evaluation splits (See section 3.1)

The second approach also used the BLIP model. This time, instead of using a single frame at a time during training, the model is trained on a tensor of sampled frames from the clip, thus multiplying the number of image-caption couples available in the image training dataset. This model is still considered an image captioning model.

The third approach was to merge the TimeSformer model [2] with BLIP and use it as a video encoder. The resulting model architecture was called TimeSBLIP. Similarly to the second approach, this model is trained on clips and the frame sampling is performed according to the chosen strategy during training. The main idea behind the use of the TimeSformer encoder is the obtention of a system that models the temporal information between the frames, unlike BLIP, which treats them as separate shots having the same caption.

3 Training Data

3.1 VTT dataset

The development dataset along with past editions ground truths annotations for this task are publicly and freely available, under the Vimeo Creative Commons Licence [4], [1]. This dataset consists of 10862 video clips between 3 and 10 seconds in length. 6475 of them are from the Twitter Vine dataset. 4387 videos are directly hosted on the NIST website under the Creative Commons licence coming from the Flickr and Vimeo Creative Commons (V3C) datasets. All clips are annotated with two to five captions and have the same 480 x 480 pixels resolution.

This dataset was divided internally into three splits: a train split consisting of 6362 videos, a validation split of 2120 videos, and a test split of 2121 videos. In total, 10603 videos were used. The remaining 259 files were either corrupt or unobtainable.

Since the sound modality is inconsistent in quality with multiple languages spoken, in cases where speech is present, using sound was deemed inefficient.

3.2 MSR-VTT

Created as an answer to an ever-growing need for a larger dataset aimed at video captioning, MSR-VTT [5] (standing for Microsoft Video To Text) is a large collection carefully curated from the YouTube platform comprised of 10,000 clips making up 41.2 combined hours of length and more than 200,000 clip-caption pairs. Like the other Microsoft dataset used in this work, each clip in MSR-VTT is accompanied by twenty captions. The standard provided splits divides the clips as follows: 6,513 training videos, 497 for validation and the remaining 2,990 were set aside for testing.

4 Bootstrapping Language-Image Pre-training

Bootstrapping Language-Image Pre-training (BLIP) [3] is a Video-Language Pre-training (VLP) framework designed to be more flexible and used for understanding-based tasks (question answering, for example) and generation-based tasks (like image captioning).

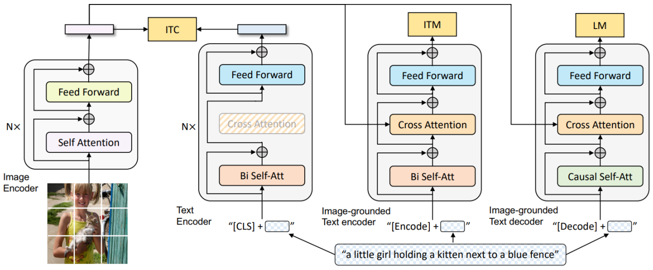


Figure 1: The multimodal mixture of encoder and decoder (MED) architecture used by the BLIP framework.

The advantage of the BLIP framework lies in both its pre-training process through the CapFilt [3] process, which can be considered a variant of knowledge distillation and the feature alignment step. During the first process, two unimodal encoders, a multimodal image-grounded encoder and the image-grounded text decoder are trained (see figure 1) by initializing a *Captioner* and a *Filter* from the same MED model. The captioner is a multimodal image-grounded text encoder and the filter is an image-grounded text decoder. A large dataset consisting of images collected from the interned labelled with their alternative text is used in conjunction with synthetic captions generated by the captioner. The filter removes noisy captions (web-sourced or synthetic) from the dataset. Knowledge distillation is achieved by having each module performing its respective task.

The *feature alignment* step is performed through the image-text contrastive learning loss (ITC) [3], [6]. By aligning the textual and visual features during pre-training, the semantic gap between both modalities can be reduces and similar embeddings with close semantic meanings can be obtained across modalities. This operation significantly improves the performance of the model in downstream tasks.

For the specific task at hand, video to text description, only the image encoder and the image-grounded text decoder are used to generate captions (see figure 2). This model is obtained by fine-tuning those modules for caption generation. The image encoder is a Vision Transformer (ViT) [7] which works by feeding patches of an image as a sequence to a BERT [8] encoder. The decoder is also a BERT block with an added cross-attention module after the self-attention in order to inject the visual information.

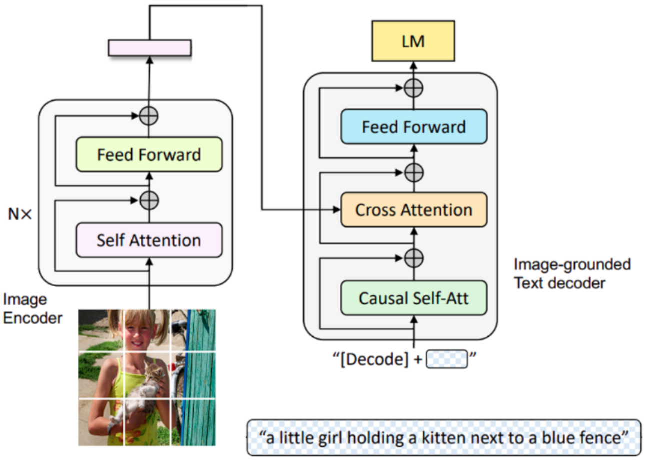


Figure 2: Architecture of the BLIP model for image caption generation when fine-tuning.

5 TimeSformer

The TimeSformer architecture -which stands for *time-space* transformer- builds on the original Transformer [9] in conjunction with the Vision Transformer [7]. The aim is to make a model that not only does away with convolutions, as is the case with the ViT, but can also be used with video content. The motivation behind the idea is that 3D convolutions do not take into account the temporal dimension of video data. there are four TimeSformer variants, each having a different time-space attention scheme. The most notable attention mechanism is the *divided attention* (See the block on the left in figure 3) in which temporal attention is performed before the spatial attention as it resulted in the best scores overall in the original work [2].

TimeSformers sample N frames from the input clip and decompose each frame into patches (usually 16) covering the entire frame without overlap.These patches are then linearly embedded and fed to the subsequent layers, as is the case in the ViT. The difference lies in the attention mechanisms ahead (See the block on the left in figure 3).

In this paper, when referring to TimeSformer, the implied variant is the Divided Space-Time Attention mechanism, as it is the one used to implement the TimeSBLIP model.

6 TimeSBLIP

The model used for the third approach is TimeSBLIP. As the name suggests, this model is a combination of the TimeSformer and BLIP models. The ViT encoder is replaced by a TimeSformer module. The idea behind TimeSBLIP is to conceive a model capable of leveraging and understanding the temporal information existing in the clips. Thus, moving from a model that sees clips as a set of independent frames sharing the same caption to a model that can represent the temporal dependencies between them.

Previous work on the past editions of the video to text description task [10] suggests that temporal features outweigh the spatial ones when it comes to caption quality. This is especially the case for action that have a direction through time, like going up or down stairs, for example.

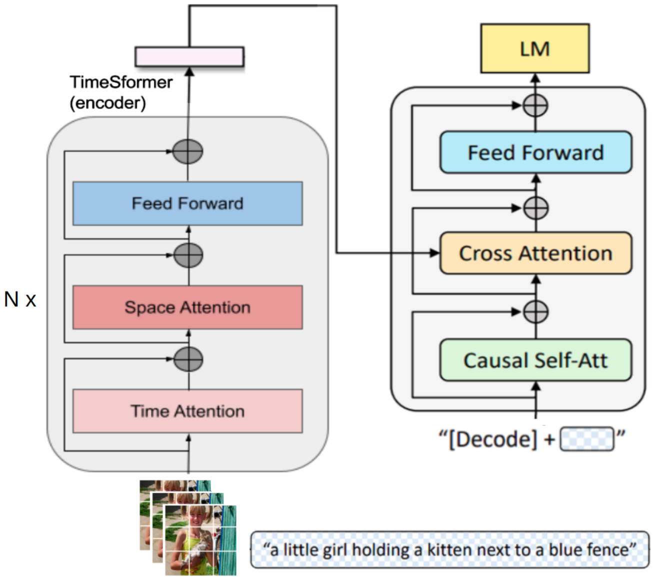


Figure 3: TimeSBLIP architecture obtained by combining the TimeSformer encoder and the multimodal image-grounded text decoder.

A high spatial resolution module (*TimeSformer-HR*) pretrained on the Kinetics 600 dataset using $16 \times 448 \times 448$ clips was grafted onto the caption generation MED model in place of the ViT encoder (see figure 3). The obtained model was then fine-tuned on the MSR-VTT dataset and the provided TRECVid VTT development dataset. Since the best compromise between performance and training cost is achieved with the Divided Space-Time Attention [2], it was the adopted attention scheme for TimeSBLIP.

In order to connect both modules of TimeSBLIP, the output shape of the encoder must match the expected input shape of the cross-attention in the text decoder responsible for injecting the visual information. To achieve this shape compatibility, the output tensor of the encoder was divided into two halves (one for time and another for spatial representation) that were then summed. This approach has the benefit of being computationally cheap and fast. For a resolution of 480×480 pixels, the standard output shape of the TimeSformer would be $B \times 1801 \times D$, where B is the batch size and D is the embedding dimension. The expected cross-attention input shape is $B \times 901 \times D$. Excluding the special BERT <*CLS*> token prepended to the embeddings, it is noticeable that there are double the embeddings with TimeSformer than what was previously achieved with a ViT. This is intuitively attributed to the addition of the temporal attention operation.

7 Experiments

Three models were trained in total, each corresponding to a run. The fourth run is a combination of the previous three by selecting the caption having the highest confidence score for each clip among those generated. Thus, the following runs were submitted:

- Run 1: **BLIP** model fine-tuned on images.
- Run 2: **BLIP** model fine-tuned on multiple clip frames.
- Run 3: **TimeSBLIP** trained on the TRECVid 22 dataset.
- Run 4: Selection of the caption with the highest **confidence score**.

All models were trained on a single Nvidia Quadro RTX 6000 graphics card and implemented using PyTorch [11].

7.1 Image captioning with BLIP

The first run consists of an *image captioning* BLIP model fine-tuned on the TRECVid 2022 dataset. To that end, frames corresponding to the exact half of the duration of each clip were sampled. The obtained dataset was then used to train BLIP. The captions were generated using the *beam search* decoding strategy with a beam size of 4, which yielded the best results, as opposed to the *nucleus sampling* strategy which gave much lower validation scores. This proved true for all the models trained. Table 1 shows the parameters used for training.

	Parameter	Value
Data loading	Batch size	16
	Image size	480
Training	Learning rate	5×10^{-6}
	Weight decay	0.05
Captioning	Minimum length	5
	Maximum length	25
	Beam size	4

Table 1: Parameters of the best BLIP model trained for the image captioning task

7.2 Video captioning with BLIP

The second run is also a fine-tuned BLIP model. The difference, compared to the first one, lies in the dataset used. This instance was trained using the full clips with a random frame sampling strategy. The parameters used are the same as those of the first run (see Table 1) except for the beam size which was lowered to 3, since it gave better results, in this case.

This model achieved lower validation scores than BLIP for image captioning. This can be attributed to the frame sampling strategy. Another consideration is the fact that a training epoch takes more than *13 hours* to complete, compared to the 2 hours duration of its image captioning counterpart's training epoch. This also resulted, overall, in less training for this BLIP variant, which is reflected in the results.

7.3 Video captioning with TimeSBLIP

The third run is a TimeSBLIP model trained on the MSR-VTT dataset and fine-tuned on the TRECVid 2022 dataset. A lower learning rate was used to avoid overfitting. Eight frames were sampled per clip, for a batch size of 16, while keeping the original resolution of the dataset. Table 2 the training parameters used for TimeSBLIP.

The validation scores of TimeSBLIP were much lower than those of the other runs. This can be attributed to the loss of the embedding alignment between the textual and visual modalities, since the ITC learning optimization is no longer performed in this model.

	Parameter	Value
Data loading	Batch size	16
	Number of frames	8
	Frame size	480
Training	Learning rate	10^{-7}
	Weight decay	0.05
Captioning	Minimum length	5
	Maximum length	25
	Beam size	3

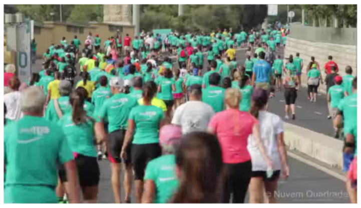
Table 2: Parameters for fine-tuning TimeSBLIP on the TRECVid dataset.

8 Results

The four runs discussed in Section 7 were submitted and yielded the results reported in Table 3. The results corroborate what was observed on the validation scores during training: the best performing model is BLIP for image captioning, thanks to its pre-training and feature alignment process. Video captioning systems performed much worse, for different reasons each, as discussed in Section 7.

Figure 4 shows a sample of the different systems outputs when captioning a video. While the first system performs best and is semantically close to the ground truth annotation, it is less confident in its caption than second system. This caused the fourth system to select the caption generated by the second system. Captions generated from systems 2 and 3 have poor quality, which explains the lower results in table 3.

Overall, the first system performed well, especially in the CIDEr-D and STS metrics, placing among the best performing submissions.



s1: many people in green shirts are running on the street
s2: in the daytime, two men are playing basketball on an outdoor court
s3: on group people in all shirts, and green their along the street
s4: in the daytime, two men are playing basketball on an outdoor court

ref: Outside on a sunny day on a highway, a multitude of people, most wearing teal t-shirts, are walking ready to start a walking or running marathon.

Figure 4: An video sample taken from the TRECVid VTT task test set. Captions s1 through s4 are generated by our submissions. Whereas the reference (ref) is the first of the five ground truth annotations.

The TimeSBLIP model shows promise and given the reintroduction of a multimodal feature alignment mechanism, performance could improve. Other fields of improvement include the connection between the TimeSformer encoder and the text decoder: Instead of simply summing-up the spatio-temporal features, other strategies that permit their isolation could be tried. This work is left for the future.

The fourth run performed worse than the first. This shows that even though the image captioning system obtained the highest scores in all the evaluations, it is less confident about some of its captions than other models.

A more detailed examples of captions generated by our submitted systems are available in figure 5 in appendix A. We present five randomly selected videos from the test set, each represented by frame accompanied by its corresponding captions generated by our systems and the ground truth reference for comparison purposes.

9 Conclusion

This report presents the systems submitted by Elyadata for the Video To Text description (VTT) task for the 2022 edition of TRECVid. All systems are BLIP-based. The image captioning variant performed best, whereas both video captioning models, although more confident in their captioning in some instances, performed far worse. On average our best run placed second out of six, when considering the best run from each participating team. These models were either a direct conversion of BLIP [3] for video captioning or a modification of the latter, consisting in the replacement of its ViT [7] encoder by a TimeSformer [2] module.
The TimeSBLIP model performed far worse than the BLIP model for image captioning. Future work will focus on improving the former to make it competitive and able to correctly model temporal features.

Run	BLEU@4	METEOR	CIDEr	CIDEr-D	SPICE	STS 1	STS 2	STS 3	STS 4	STS 5
1	**6.936**	**24.84**	**50.70**	**22.60**	**10.20**	**42.11**	**41.89**	**41.99**	**41.91**	**41.51**
2	1.298	17.83	10.30	4.50	4.30	23.57	24.01	23.02	23.78	23.86
3	1.403	16.92	24.30	7.60	6.20	35.70	34.13	33.51	33.61	36.27
4	3.414	19.41	23.40	10.50	6.40	30.83	30.34	29.94	30.34	30.73

Table 3: Submission results for the four runs on the TRECVid 2022 dataset.

References

[1] G. Awad, K. Curtis, A. A. Butt, *et al.*, "An overview on the evaluated video retrieval tasks at trecvid 2022," in *Proceedings of TRECVID 2022*, NIST, USA, 2022.

[2] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," 2021. arXiv: `2102.05095 [cs.CV]`.

[3] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," *CoRR*, vol. abs/2201.12086, 2022. arXiv: `2201.12086`. [Online]. Available: `https://arxiv.org/abs/2201.12086`.

[4] "Index of tv_vtt_data." (), [Online]. Available: `https://ir.nist.gov/tv_vtt_data/`. (accessed: 25.02.2022).

[5] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016. [Online]. Available: `https://www.microsoft.com/en-us/research/publication/msr-vtt-a-large-video-description-dataset-for-bridging-video-and-language/`.

[6] J. Li, R. R. Selvaraju, A. D. Gotmare, S. R. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *CoRR*, vol. abs/2107.07651, 2021. arXiv: `2107.07651`. [Online]. Available: `https://arxiv.org/abs/2107.07651`.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. arXiv: `2010.11929`. [Online]. Available: `https://arxiv.org/abs/2010.11929`.

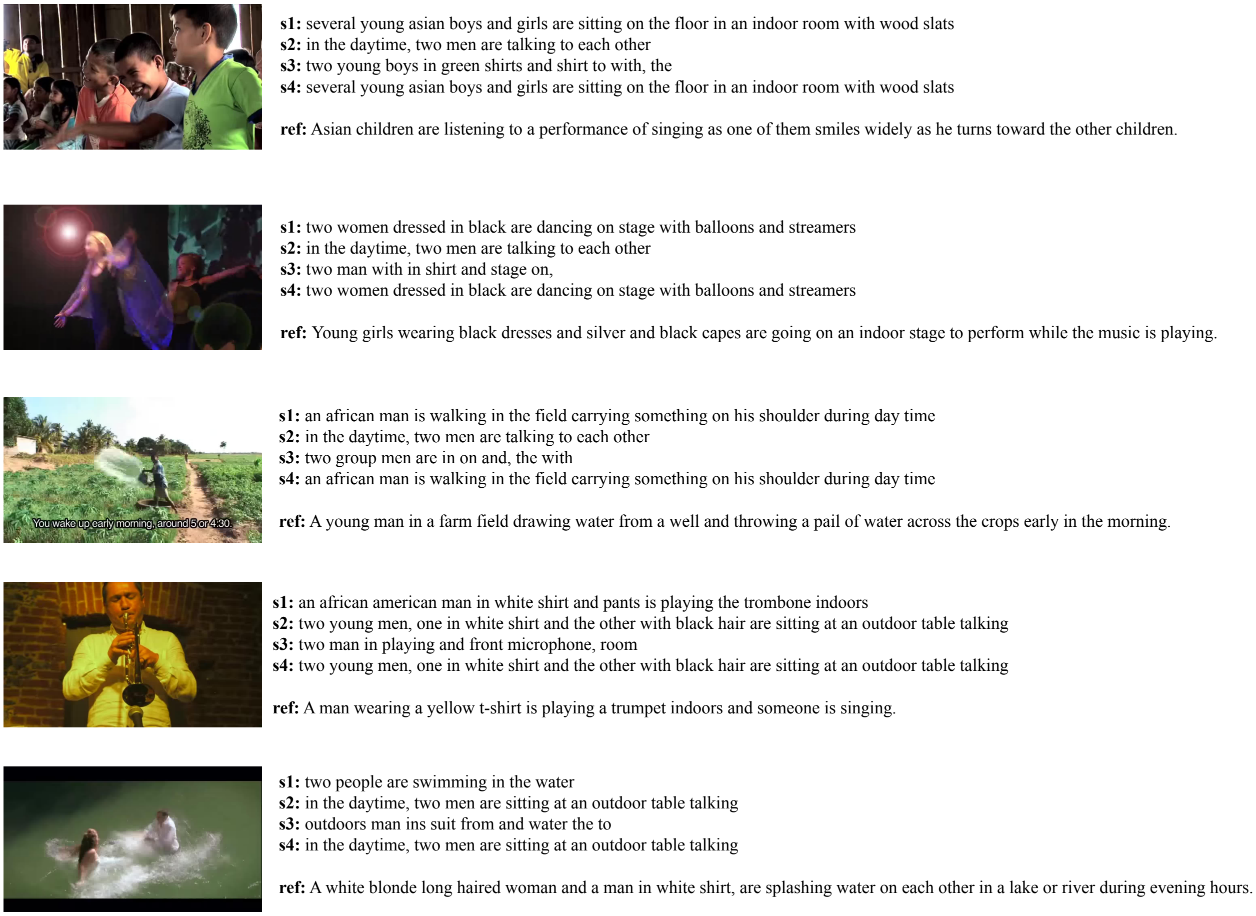
[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. arXiv: `1810.04805`. [Online]. Available: `http://arxiv.org/abs/1810.04805`.

[9] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. arXiv: `1706.03762`. [Online]. Available: `http://arxiv.org/abs/1706.03762`.

[10] Y. Song, Y. Zhao, S. Chen, and Q. Jin, "Ruc_aim3 at trecvid 2020: Ad-hoc video search & video to text description," in *Proceedings of TRECVID 2020*, NIST, USA, 2020.

[11] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *CoRR*, vol. abs/1912.01703, 2019. arXiv: `1912.01703`. [Online]. Available: `http://arxiv.org/abs/1912.01703`.

A Caption generation example



s1: several young asian boys and girls are sitting on the floor in an indoor room with wood slats
s2: in the daytime, two men are talking to each other
s3: two young boys in green shirts and shirt to with, the
s4: several young asian boys and girls are sitting on the floor in an indoor room with wood slats

ref: Asian children are listening to a performance of singing as one of them smiles widely as he turns toward the other children.

s1: two women dressed in black are dancing on stage with balloons and streamers
s2: in the daytime, two men are talking to each other
s3: two man with in shirt and stage on,
s4: two women dressed in black are dancing on stage with balloons and streamers

ref: Young girls wearing black dresses and silver and black capes are going on an indoor stage to perform while the music is playing.

s1: an african man is walking in the field carrying something on his shoulder during day time
s2: in the daytime, two men are talking to each other
s3: two group men are in on and, the with
s4: an african man is walking in the field carrying something on his shoulder during day time

ref: A young man in a farm field drawing water from a well and throwing a pail of water across the crops early in the morning.

s1: an african american man in white shirt and pants is playing the trombone indoors
s2: two young men, one in white shirt and the other with black hair are sitting at an outdoor table talking
s3: two man in playing and front microphone, room
s4: two young men, one in white shirt and the other with black hair are sitting at an outdoor table talking

ref: A man wearing a yellow t-shirt is playing a trumpet indoors and someone is singing.

s1: two people are swimming in the water
s2: in the daytime, two men are sitting at an outdoor table talking
s3: outdoors man ins suit from and water the to
s4: in the daytime, two men are sitting at an outdoor table talking

ref: A white blonde long haired woman and a man in white shirt, are splashing water on each other in a lake or river during evening hours.

Figure 5: Samples of videos taken from the TRECVid VTT task test set. Captions s1 through s4 are generated by our submissions. Whereas the reference (ref) is the first of the five ground truth annotations.