CMU-VIDION: MODIFIED BLIP WITH AUDIO FOR VIDEO TO TEXT DESCRIPTION

Laura J Yao, Juncheng B Li*, Florian Metze*

{ljyao, junchenl, fmetze}@andrew.cmu.edu
Carnegie Mellon University

ABSTRACT

State-of-the-art vision-language pre-trained (VLP) models such as BLIP [1] are demonstrating impressive performance for zero-shot predictions of captions based on visual inputs. However, these VLP models are not aware of audio information. This paper describes our system submitted to TRECVID2022-VTT task [2] that leveraged a fine-tuned pretrained BLIP model trained on COCO, Visual Genome, three web datasets, Conceptual 12M, SBU Captions, and LAION [1] combined with a model that we trained on AudioSet [3] to account for the audio modality of the data. Our audio-visual system was able to generate more detailed captions based on audio auxiliary information compared to the vision baseline. We rank 3rd on CIDEr evaluation metric. Our runs differed in the keyframes used and whether the audio modality was considered. The runs submitted used the following approaches:

- VIDION.VIDION_cmu_1.vtt.run.txt (Run 1): middle frame ($\frac{1}{2}$)
- VIDION.VIDION_cmu_2.vtt.run.txt (Run 2): all three frames ($\frac{1}{3}$, $\frac{1}{2}$, $\frac{3}{4}$)
- VIDION.VIDION_cmu_3.vtt.run.txt (Run 3): middle frame and audio analysis
- VIDION.VIDION_cmu_4.vtt.run.primary.txt (Run 4): all three frames and audio analysis

***Index Terms*—** CLIP, BLIP, VLP, Audio-visual Description

1. INTRODUCTION

Video-To-Text description systems have seen huge improvements over the past 2 years with the advent of contrastive vision language pre-trained (VLP) models (e.g. CLIP [4]) trained with hundreds of millions of image-text pairs without supervision from high-quality annotation. CLIP [4] demonstrated the possibility of competitive zero-shot prediction performance compared to supervised baselines, but it also showed the limitation of simply scaling up the dataset: the poor performance of CLIP on Out-of-Distribution (OOD) data and on abstract tasks. BLIP [1] is one of the promising improvements over CLIP, which tackles the noisy image-text pair issues in web data through the caption filtering mechanism[1] and improves the task-specific performance by using a more expressive multimodal multitask encoder. Under the TRECVID2022 Video-To-Text (VTT) task setting, there is not enough annotated data for training, therefore, it is natural to leverage a VLP model and test it under the zero-shot setting. We adopt the BLIP model to take in single/multiple video frames in our system and generate the corresponding captions. The reason that we used an image-text model instead of a video-text model goes back to the scarcity of existing video-text annotations, as BLIP [1] demonstrated that they could outperform VideoCLIP [5] which retrieves text directly based on video features.

On the other hand, we observe the input contains three modalities including visual, audio, and text. However, the current state-of-the-art (SOTA) approach does not leverage all three of these modalities. This provides a natural motivation for us to explore more contextual information about a video from its audio. Following the same intuition of knowledge transfer, we train models on AudioSet [3] which is the largest existing general audio dataset, and use this model to extract audio information from the TRECVID test data.

In the end, we combine our visual-language backbone with our audio backbone to perform VTT description. We demonstrate in our ablation that extra modality, audio in this case, indeed boosts the performance, especially adding more audio contextual details to the generated captions.

2. OUR APPROACH

Our general approach to the task consisted of three main steps: extracting keyframes from the video, generating captions based on the visual features, and then concatenating audio-based captions to the previously-generated visual captions. A brief illustration of our framework is shown in Fig. 1.

For each video, we segmented them into frames and sam-

* Co-first author

[1] Caption filtering: uses a captioner to generate synthetic captions and a filter to remove the noisy ones.

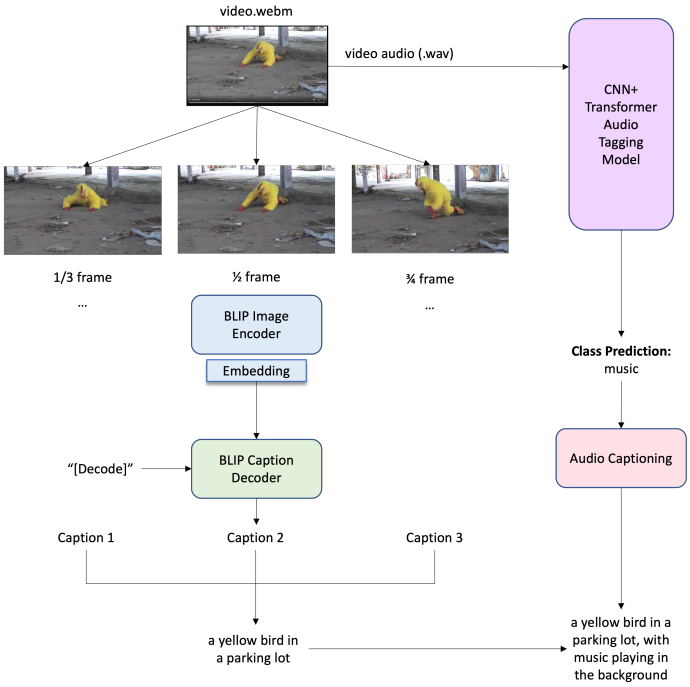

Fig. 1. The overall architecture of our system

pled the frames at the $\frac{1}{3}$ point into the video, at the $\frac{1}{2}$ point into the video, and at the $\frac{3}{4}$ point into the video to get an overview of the actions and visuals throughout the entire video.

The different frames were evaluated individually using the BLIP model to create the corresponding textual embeddings. We then utilized beam search with a beam width of 3 to generate synthetic captions with some flexibility [1]. These results were combined by averaging the Levenshtein distance from the other 2 predictions. This distance was used as a metric to determine which caption was the closest to the other 2 captions (creating a majority vote) for the "best" visual-based caption.

Concurrently, we use the best-performing pre-trained CNN+Transformer model described in [6] (43.1 mAP on AudioSet evaluation set) to process the audio input of the video data. All the audio inputs are first resampled to 16kHz, from which we extract logMel spectrogram features.[2] The pretrained CNN+Transformer model takes these audio features and predicts the corresponding label of the audio event and generates the corresponding audio caption based on a few custom templates: 1) who/what is making these sounds; 2) whether it is background music; 3) what instrument is playing.

[2]We use 64 Mel filters; frames of 1,024 samples (64 ms) are taken with a hop of 400 samples (25 ms); each frame is Hanning windowed and padded to 4,096 samples before taking the Fourier transform; The shape of the resulting feature is 64×400.

	SPICE	CIDEr	CIDErD	BLEU	METEOR
Run 1	0.073	0.595	0.098	0.0246	0.2118
Run 2	0.073	0.589	0.099	0.0258	0.2126
Run 3	0.077	0.607	0.108	0.0298	0.2209
Run 4	**0.077**	**0.611**	**0.113**	**0.0300**	**0.2219**

Table 1. Overall results of our system over automatic description metrics

	STS 1	STS 2	STS 3	STS 4	STS 5
Run 1	0.3967	0.3897	0.3939	0.3909	0.3903
Run 2	0.3984	0.3868	0.3928	0.3930	0.3889
Run 3	0.4065	0.3947	0.3996	0.3986	0.3955
Run 4	**0.4062**	**0.3952**	**0.3999**	**0.3985**	**0.3953**

Table 2. Semantic Similarity metric (STS) (Human Evaluation) with different ground-truth sets

3. EXPERIMENTS & RESULTS

3.1. Task Setup & Dataset

The TRECVID Video-To-Text task is to automatically annotate videos with natural language text descriptions. Specifically, the task asks participants to generate a single sentence that describes each video with videos ranging from 3-10 seconds long. These generated sentences are evaluated using automatic scoring metrics such as METEOR, BLEU, CIDEr, and SPICE as well as a semantic similarity metric (STS) to test how semantically similar the ground truth sentence is compared to the generation [2]. The TRECVID VTT task utilizes a dataset consisting of video segments from the V3C1 collection, which is part of the Vimeo Creative Commons Collection (V3C). V3C1 consists of 7475 Vimeo videos, which total about 1000 hours. These videos are divided into over 1 million segments. In the test set, a subset (around 2000) of these videos are used with each segment being from 3 to 10 seconds long [7].

3.2. Results & Discussion

Table 1 shows our results when run on the test set with a few different automatic description metrics. We ranked third for the CIDEr evaluation metric on run 4 (our primary run). From these results, we see that across all the metrics, run 4 had the best scores, closely followed by run 3. These results could indicate the importance of the audio modality in generating more detailed captions since both of these runs utilized the audio auxiliary information. The runs that utilize frames across the entire video also consistently outperform the runs that only utilize the middle frame of the video which suggests that the model benefits from more visual information across the video.

Table 2 shows our results with the semantic similarity metric that measures how the system generated description

Generated Caption	Ground Truth
a person sitting on the ground	young woman in a *black vest and pink tights and top* sitting on the *curb* in front of a blocked up red brick building on a *sunny day*.
a bride and groom walking down the aisle	A white woman in a bridal gown walking on a *grass lawn* near trees and a stage and seating on a *sunny day*.
a dog on the ground, with music playing in the background	With *eerie* music playing at dusk, a *black* dog sits majestically on a *concrete area* with white lettering. and then shadow of a lamp post.
a group of people running on the beach	Two men in *white shirts* are reaching the finish line in a running race, as others are still running, and *photographers* are taking pictures on a *sunny day* on the beach.

Table 3. Generated captions and their corresponding ground truth statements from the set used in the STS 5 metric

is semantically related to the ground truth captions across 5 different sets of captions [2]. We see that run 3 and run 4 still perform better than run 1 and 2, which further shows that the contextual information provided from audio is important to improving semantic similarity with ground truth captions.

From our qualitative analysis of the captions as well as the comparison to the run with the best metrics (RUCAIM3-Tencent) and ground truth statements [2], we observe that the captions we generated tend to have a more simplistic sentence structure and did not include key background descriptions or more specific color details on the objects within the videos. Table 3 contains several example captions our primary (4th) run generated compared to an example ground truth caption (STS 5 in Table 2) [2]. All of our captions follow the same sentence structure with the subject of the sentence followed by an action and then the setting as well as any additional audio captions. The ground truth statements have a more varied structure with the ordering of what happens in each caption as well as with the vocabulary used. Although our generated captions captured the key details of each video, they are not able to fully generate the context and details of the situation for each video. Details like what type of music is playing or what the weather is in the video are usually missing.

4. CONCLUSION & FUTURE DIRECTION

Our current framework loosely joins audio and video/visual input when generating descriptions for video, concatenating the output of two different models to generate the joint caption. This does not utilize the possible correlation between audio to textual or visual features in a video. From the results, we see that audio is important for boosting the captions generated to more closely match ground truth statements. We plan on exploring how to better add audio into CLIP-type models and improve the understanding of these audio correlations. This would allow for more accurate or detailed usage of the audio modality in generating descriptions for videos.

5. REFERENCES

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

[2] 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, "An overview on the evaluated video retrieval tasks at trecvid 2022," in *Proceedings of TRECVID 2022*. NIST, USA, 2022.

[3] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

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

[5] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer, "Videoclip: Contrastive pre-training for zero-shot video-text understanding," *arXiv preprint arXiv:2109.14084*, 2021.

[6] Juncheng B Li, Shuhui Qu, Florian Metze, et al., "Audiotagging done right: 2nd comparison of deep learning methods for environmental sound classification," *arXiv preprint arXiv:2203.13448*, 2022.

[7] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A Butt, "V3c–a research video collection," in *International Conference on Multimedia Modeling*. Springer, 2019, pp. 349–360.