Year 3 BAIR Commons Update: Low-Data Learning for Assistive Video Description (August, 2021)

BAIR: David Chan (davidchan@berkeley.edu), John Canny (canny@berkeley.edu)
Google: David Ross (dross@google.com), Austin Myers (aom@google.com), Sudheendra Vijayanarasimhan (svnaras@google.com), Bryan Seybold (seybold@google.com)

Project Outline

Globally, over 285 million people suffer from some form of visual impairment, of which more than 40 million are fully blind[1]. In many cases, traditional media has remained accessible to the visually impaired through Descriptive Video Services (DVS), an additional narration track for videos which is intended to make video content available to visually impaired users. DVS is a time-intensive manual process requiring annotators to watch videos, decide which elements of the video are important to convey the visual information in the scene, and write/perform a script which captures that visual information. Such annotation is possible, but expensive (~$20/min). Online video media has therefore remained inaccessible to those with visual impairments due to the cost and scalability of current DVS annotation procedures. With online interaction becoming a societal norm, the task of making video content available to all users has become an even higher priority.

The goal of the proposed project is to **continue the development of a human-in-the-loop system for automated DVS (ADVS)**. Designing such a system requires investigation into a number of complex tasks including: understanding useful description formats, developing data collection methods such as active learning systems, semi-supervised and unsupervised video to text translation, video understanding and modeling, and evaluation of video descriptions. This project has the potential to significantly expand the development of ADVS, with the aim of developing user-testable end-to-end systems which can watch, understand, and describe videos.

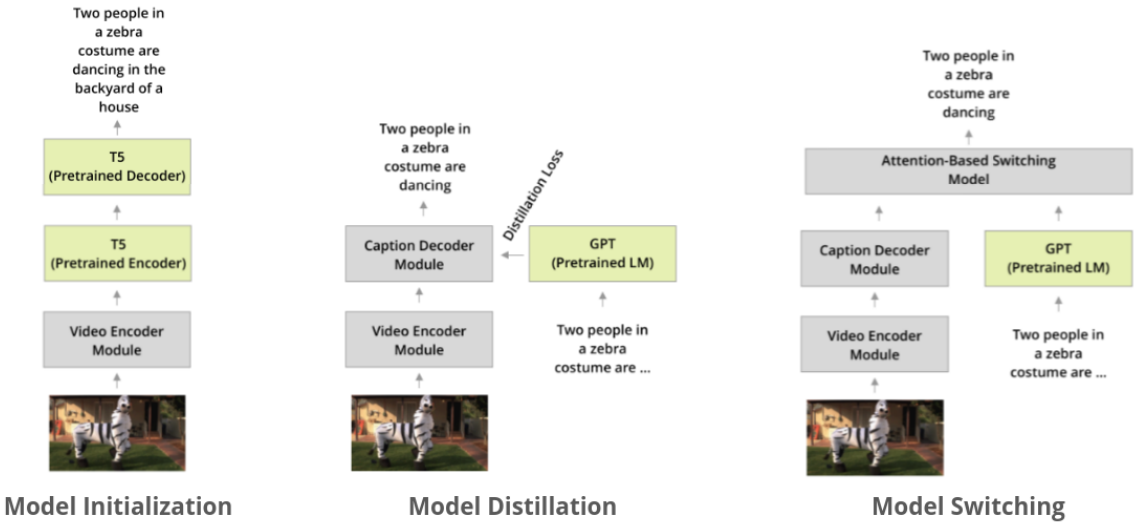
Currently, the development of ADVS is heavily limited by the availability of high-quality training data. Professionally annotated datasets are primarily focused on long-form media such as TV shows and movies, with difficult semantic syntax only containing a few hundred videos. Large-scale data is available, but relies on inexperienced mechanical turk users or automated speech recognition, leading to descriptions that are inadequate for visually impaired users.

Project History & Results

In year one of this project, we focused on reducing the impact of the data collection process for ADVS and we demonstrated an active-learning based approach to video descriptions which achieved performance parity with the current state of the art using only 25% of the available data (Presented at ACCV 2020). For more detailed information regarding the first year of the project, see the report here: https://bit.ly/2WqxFKk.

In year two of this project, we have focused on the reuse of large-scale vision and language models leveraging self-supervised learning. In the first part of the year, we focused on potential extensions to the text-generation component of the model, developing three novel methods for integrating pre-trained language models into the training pipeline.

NLP in Captioning: 3 Paradigms



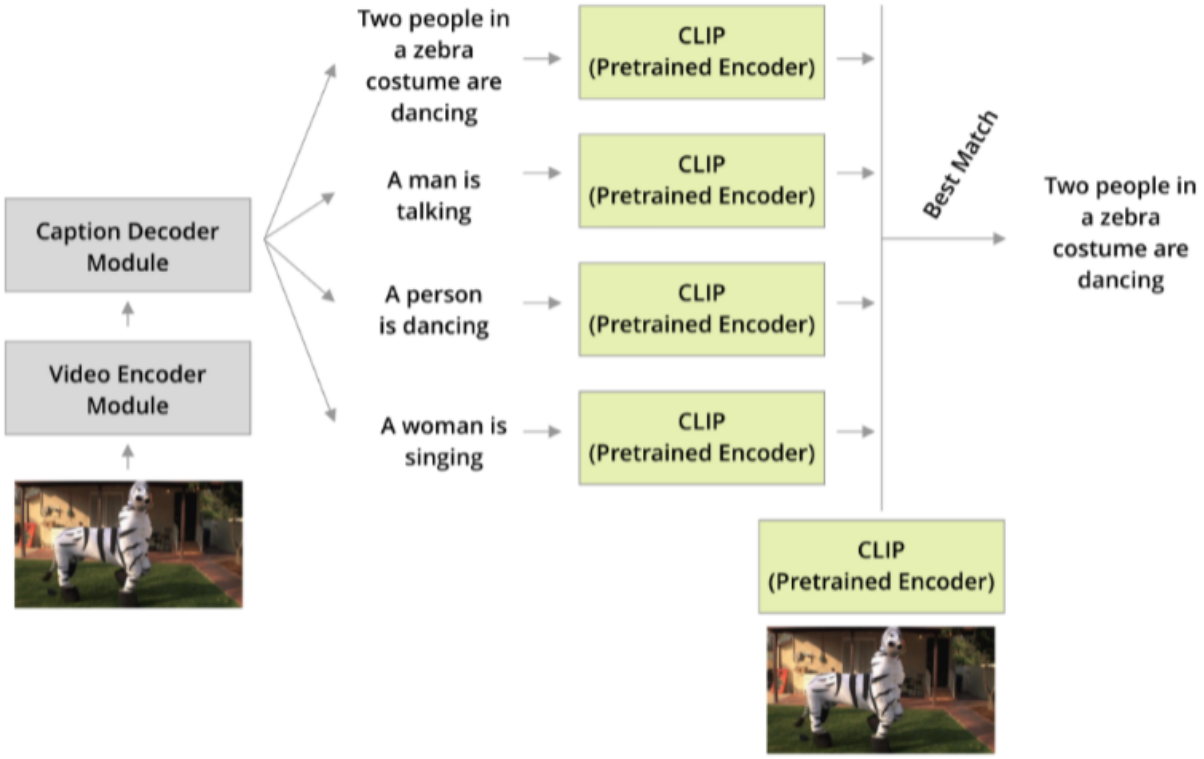
Model Initialization **Model Distillation** **Model Switching**

Using these three paradigms, we found that we were able to achieve improvements over the baseline model of up to 10%, demonstrating that exploiting language-only pre-training can provide significant benefits over traditional from-scratch learning.

Results (Automated, MSR-VTT Dataset)

	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGE	CIDER
Baseline	0.712	0.601	0.501	0.409	0.261	0.582	0.420
Initialization	0.754	0.609	0.481	0.370	0.262	0.578	0.407
Distillation	0.732	0.590	0.468	0.368	0.259	0.563	0.358
Switching	**0.763**	**0.641**	**0.522**	**0.412**	**0.268**	**0.595**	**0.439**

Additionally, we explored the use of large-scale vision and language alignment models such as OpenAI's CLIP[2] model, as a post-processing technique as shown in the figure below:



We found that such models provide strong post-processing benefits, increasing the performance of captioning models by up to 8% over baselines under automated evaluation. We are preparing both of these results for a submission later this year.

Project Future Goals and Targets

Even though we have demonstrated that large-scale language modeling has the potential to greatly improve the performance of ADVS systems, evaluation of the results has demonstrated that current approaches consistently struggle to accurately ground descriptions to the underlying video data. Models often produce language which is overly generic, i.e. “A person is talking”, or factually incorrect. **In year three, we aim to target this problem of specificity by leveraging a hierarchical structured data approach based on the idea of abstract scene graphs to reuse information learnable from datasets and models trained on similar, but not identical vision and language tasks.**

Scene graphs are a structured data representation containing agents, objects, and their relationships in an image/video. Several research groups[3,4,5,6] have demonstrated that ground-truth scene graphs have the potential to improve specificity and general caption quality in the image description domain. Such representations, however, have yet to be deeply explored in video description - primarily due to the lack of an analogous structure to the scene graph in video data[8].

Therefore, while the generation of “video scene graphs” is an important and promising research direction, it remains a challenging and open problem. We hypothesize that this may be addressed by bootstrapping from pseudo scene-graph representations, derived from structured text based on actions,

objects, and scenes in the video, extracted using current SOTA computer vision techniques, to improve the performance of video description by acting as a pre-training objective and a downstream captioning task input. Preliminary results in the image captioning domain are promising, with such linguistic-based techniques approaching state of the art performance even in the *unsupervised* scenario[7]. In the video description domain, where such pseudo scene-graph techniques could be based on very large object detection and action recognition datasets such as Kinetics-700 (with over 600K videos), and TrackingNet (with over 14M object detection annotations) with automated cross-labeling, we expect such techniques to be extremely performant when compared to supervised learning on under 50K videos.

Semi/Self-supervised ADVS is an exciting, emerging research area that aligns with Google Research's priorities to advance weakly/unsupervised and cross-modal vision+language learning methods. By evaluating our pseudo scene-graph method using automated approaches on external research datasets, as well as taking advantage of Google resources to perform human evaluations of the models in live environments, we hope to strongly encourage research in the important area of video description and video understanding. These results may not be applicable to video description alone, and could serve as strong inspiration in several low-data downstream tasks, each of which can improve the accessibility of information to the visually impaired in a rapidly expanding internet universe.

References

[1] “Global Data on Visual Impairment.” World Health Organization, World Health Organization, 8 Dec. 2017, www.who.int/blindness/publications/globaldata/en/#:~:text=Globally the number of people,blindness is cataract (51%).

[2] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." arXiv preprint arXiv:2103.00020 (2021).

[3] Laina, Iro, Christian Rupprecht, and Nassir Navab. "Towards unsupervised image captioning with shared multimodal embeddings." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[4] Yang, Xu, et al. "Auto-encoding scene graphs for image captioning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[5] Li, Xiangyang, and Shuqiang Jiang. "Know more say less: Image captioning based on scene graphs." IEEE Transactions on Multimedia 21.8 (2019): 2117-2130.

[6] Milewski, Victor, Marie-Francine Moens, and Iacer Calixto. "Are scene graphs good enough to improve Image Captioning?." arXiv preprint arXiv:2009.12313 (2020).

[7] Gu, Jiuxiang, et al. "Unpaired image captioning via scene graph alignments." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019

[8] Ji, Jingwei, et al. "Action genome: Actions as compositions of spatio-temporal scene graphs." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.