Research Overview

Jie Lei, jielei@cs.unc.edu *Adobe Fellowship Application*

My primary research interest lies in the intersection of computer vision and natural language processing, particularly video and language understanding. It is an important area to study, as it reflects the real world, where people communicate through language, and where many computational systems like robots would ultimately have to operate. Many videos, like tutorial videos[1] for Adobe products and user generated videos (such as those edited using Adobe Premiere Pro), come with dialogues. Meanwhile, people often interact with videos via language (e.g., search videos related to a natural language query). Video and language understanding is challenging as it involves visual and language semantic understanding, spatio-temporal grounding, and commonsense understanding. My long-term goal is to equip systems with the ability to interact with people in various environments using language, including online video platforms and real-world environments. To achieve this goal, my current research focuses on designing tasks, datasets, and algorithms for video and language understanding at two levels: understand events at present and predict events in the future.

The first level is to make sense of the observed facts, involving tasks like video question answering [5, 6, 3], video captioning [7, 4, 2], and text-to-video retrieval [7, 3, 2]. It requires systems to recognize objects, actions, and people from videos (and or dialogue), and reason over their contextual relationships (e.g., figuring out who is holding a cup). It also requires connecting query words with events in video frame pixels and dialogue text in space and time. For example, given a video corpus, retrieving relevant moments to a user query 'people discussing marriage in a coffee shop.' Beyond returning the most related videos, it also requires temporally locating specific moments inside these videos that have high correspondence with the user query, which helps to improve user experience. This particular problem was studied in [7], where we developed a dataset (task) for moment retrieval at the corpus-level together with a fast and accurate system for addressing it. The developed system and functionality could potentially be integrated into online video platforms (e.g., Adobe Stock[2]) and video editing tools (e.g., Adobe Premiere Pro) to better find short moments of interest from a set of online or locally stored video footages using a language query.

Given a good grasp of the present, the second level of understanding is anticipating or predicting possible events in the near future. It goes beyond basic visual recognition and language semantic understanding, requiring systems to incorporate commonsense world knowledge to make sound and practical judgments about the future. This is essential as it would eventually empower systems to predict people's intent and possible future actions, thus, seamlessly assist people in various tasks – imagine a home robot that can remind you to carry an umbrella when it senses that you are about to go out on a rainy day. In [8], we proposed a new dataset and task *video-and-language future event prediction* to study this problem. Predicting possible future events is challenging as the required commonsense knowledge is difficult to acquire at scale with human annotation. My near-term research goal is to build

[1] https://helpx.adobe.com/creative-cloud/tutorials-explore.html
[2] https://stock.adobe.com/video

systems that can automatically mine this knowledge from unlabeled videos.

During the course of my research, I noticed a few key challenges that underlying video and language learning. One of them is the disconnection in *representation learning* and *task-specific modeling*. The canonical approaches to video and language learning tasks (e.g., text-to-video retrieval) dictate a neural model to learn from offline-extracted dense video features from vision models and text features from language models. These feature extractors are trained independently and usually on tasks different from the target domains (e.g., features learned for action recognition from human activity videos are incongruently applied to question answering on generic-domain animated GIF videos), rendering these fixed feature representation sub-optimal for downstream tasks. Meanwhile, learned representations from different modalities are independent of each other. For instance, action recognition models are typically trained from pure video data without textual input, yet are applied to video and language tasks. Moreover, due to the high computational overload of dense video features, it is often difficult (or infeasible) to plug feature extractors directly into existing approaches for easy finetuning. In [3], we provide a remedy to this dilemma, where only a single or a few sparsely sampled short clips from full-length videos are used to make predictions at each training step. This strategy greatly reduces the memory and computation cost of modeling long videos, enabling efficient end-to-end training for video and language tasks.

Another challenge is the lack of large-scale high-quality annotations. Past success in video and language has mostly relied on supervised learning, where models are learned on manually labeled data for a particular task. However, manually annotating video and language data is very expensive, hence limiting the scale of such datasets, and consequently also limiting the performance of models trained on the datasets. One way to mitigate this issue is to transfer knowledge from existing high-quality and less-expensive image and language datasets. In [3], we explored a path where a model is firstly pre-trained on large-scale image and language corpus, then adopted to smaller-scale downstream video and language tasks. Another approach is to leverage unlabeled web videos, such as instructional videos, for self-supervised pre-training, followed by task-specific finetuning. In [2], we showed a method that learns from noisy, usually semantically (the narration words are sometimes irrelevant to the video content, e.g., credits) and temporally (people might talk about something before or after they actually demonstrate it) misaligned instructional videos.

Recent breakthrough in AI has partly been driven by large-scale standardized benchmarks (e.g., ImageNet [1]). My research has contributed multiple benchmark datasets and tasks across various problems in the video and language, including TVQA [5] and TVQA+ [6] for spatio-temporal question answering, TVR [7] for moment retrieval, TVC [7] for multimodal captioning, and VLEP [8] for future event prediction. To encourage diversity and inclusivity, I am also working on extending these datasets to multiple languages. I hope these resources, along with our solutions to these identified challenges above, would facilitate video and language research, and help make steps towards the ultimate goal of building systems that can seamlessly and accurately interact with people using language. I envision that the advance in this field would benefit a range of Adobe products that involve multimodal components (i.e., video and language, and potentially image, language, and audio, etc.).

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[2] Z. Tang, **J. Lei**, and M. Bansal. Improved pre-training from noisy instructional videos via dense captions and entropy minimization. *Under Review*, 2020. 1, 2

[3] **J. Lei**, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. *Under Review*, 2020. Preprint. 1, 2

[4] **J. Lei**, L. Wang, Y. Shen, D. Yu, T. L. Berg, and M. Bansal. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *ACL*, 2020. 1

[5] **J. Lei**, L. Yu, M. Bansal, and T. L. Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 1, 2

[6] **J. Lei**, L. Yu, T. L. Berg, and M. Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *ACL*, 2020. 1, 2

[7] **J. Lei**, L. Yu, T. L. Berg, and M. Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020. 1, 2

[8] **J. Lei**, L. Yu, T. L. Berg, and M. Bansal. What is more likely to happen next? video-and-language future event prediction. In *EMNLP*, 2020. 1, 2