Journal & Proceedings of the Royal Society of New South Wales, vol. 157, part 2, 2024, pp. 301–302. ISSN 0035-9173/24/020301-02

Thesis abstract

Label efficient video and language representation learning and applications

Dongxu Li

Abstract of a thesis submitted to the Australian National University

Video and language research aims to model and analyse the two communication modalities and their connections. Learning effective *video and language representation* is pivotal in facilitating a wide spectrum of applications, such as content-based video retrieval, multimedia content generation, and video-based assistive technology.[1]

Modern deep learning-based video-language models require a large amount of data for supervised training. However, obtaining accurately-annotated video and language data is laborious and expensive, especially for tasks requiring domain expertise. Consequently, existing works usually show compromised results with the limited access to annotations. To this end, this thesis devises *label-efficient algorithms for video and language understanding*, aiming to learn good video and language representations with only a few and/or weak labels. To demonstrate the practical importance of these techniques, we also study extensively their applications on automated video sign language understanding, where annotations are scarce due to the costly domain knowledge required. The main contributions of this thesis are summarised as follows.

First, we present a generic video and language pre-training framework (ALPRO), which learns effective multimodal representations from video-text pairs. Instead of fully-annotated video-text pairs, we use those easily accessible from the web to reduce the demand for human labelling efforts. Specifically, our method aims at capturing alignment between video and text inputs. This is achieved by contrastively aligning unimodal video-text features at the instance level, as well as enhancing the fine-grained alignment between visual regions and textual entities. When transferring to downstream tasks, such as video-and-text retrieval and video question answering, our pre-trained model surpasses previous methods by a significant margin, while using orders of magnitude less training data.

We then describe our efforts in the development of techniques and resources for automated sign language understanding and generation, a typical video and language task where labels are expensive to acquire.

In particular, we study the problem of word-level sign language recognition from videos, aiming at classifying gestures of sign language “words" from videos. Training recognition models for such a task requires video samples with large variations in signer appearance; therefore, scalable datasets with labels are not commonly existent. To tackle this issue, we propose to utilise sign language news videos on public video sharing platforms as an auxiliary data source with weak labels, leading to a self-training

1 ANU 2024 JG Crawford Prize for STEM

framework. We are motivated by the observation that important visual concepts are shared across domains and propose to learn domain-invariant visual descriptors that benefit the recognition.

Our method obtains significant improvement across multiple public datasets, including the largest Word-level American Sign Language recognition dataset (WLASL) developed by ourselves. Apart from showing quantitative advantages over previous works, we also compile the developed techniques into an automatic sign language dictionary, GlossFinder, and demonstrate that such technology and resources help significantly to reduce the learning barriers for sign language learners.

We then study the task of glossification, of which the aim is to transcribe natural language sentences for the deaf to ordered sign language glosses. While the task has important applications in automated sign language video generation, the glossification models suffer from limited gloss annotations. To utilise more efficiently the gloss annotations, we propose to exploit the task priors when designing the glossification model. In particular, we observe that despite different grammar, glosses effectively simplify sentences for the ease of deaf communication, while sharing a large portion of vocabulary with sentences. This has motivated us to implement glossification by executing a collection of editing actions, *e.g.*, word addition, deletion, and copying, called *editing programs*, on their natural spoken language counterparts. Specifically, we design a new neural agent that learns to synthesise and execute editing programs, conditioned on sentence contexts and partial editing results. The agent is trained to imitate minimal editing programs while exploring more widely the program space via policy gradients to optimise sequence-wise transcription quality. Results show that our approach outperforms previous glossification models by a large margin, improving the transcription metrics significantly.

Finally, we study the problem of sign language translation, aiming to translate continuous sign language videos into natural language sentences. Previous works require annotations of ordered signing gestures in the videos, called *glosses*, in order to infer the boundaries between consecutive signing gestures. However, gloss annotations require in-domain sign language expertise and can be time-consuming to obtain. Instead, our model directly learns translation models from sign language videos to natural language sentences without glosses required, exhibiting the potential to extend to a wider range of sign language resources, such as subtitled news videos. We achieve this by proposing a novel sign video segment representation, which takes into account multiple temporal granularities. The model then uses the proposed inter-scale and intra-scale attention modules to adaptively select meaningful gesture segments. In this way, we avoid segmenting gesture boundaries explicitly and alleviate the annotation burdens on glosses and obtain superior results than prior works.

Dr. Dongxu Li
College of Engineering, Computing and Cybernetics
Australian National University

E-mail: dongxuli1005@gmail.com

URL: http://hdl.handle.net/1885/292348