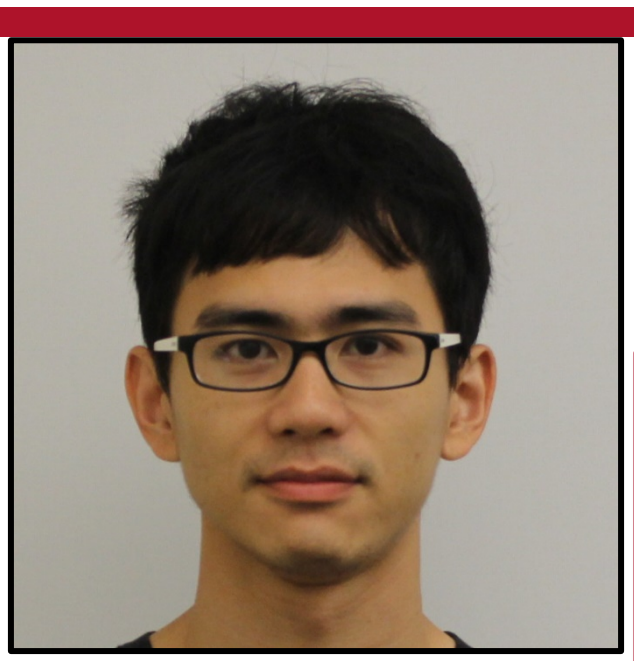
Language Technologies Institute

Thesis Proposal

Monday, May 10, 2021| 10:00 am



Disentangled Representations beyond Vectors for Multimedia Content

Ting-Yao Hu

COMMITTEE:

Alexander G. Hauptmann, (chair)

Alan W Black

Kris Kitani

Yu Tsao, (Academia Sinica, Taiwan)

Abstract

In recent years, a tremendous amount of multimedia data is being generated and published on a variety of platforms such as Instagram, Podcast, Clubhouse, and YouTube. This phenomenon inspires the research works of large-scale multimedia analysis, including the foundation of analysis methodology, and some specific downstream applications (e.g. recognition, retrieval, and information extraction). Particularly, representation learning of multimedia is one of the most crucial research directions. A good feature representation for a multimedia data instance provides interpretability and generality, improving the performance and efficiency of downstream tasks. It is challenging to obtain a good representation of multimedia content due to its richness and noisiness. For instance, in the task of speech processing, human speech utterances contain linguistic information, and other factors such as speaker identity, speaking style and background noise. In this case, we need a type of representation that captures the information from all these factors, and recovers the useful factors for downstream applications. Most of the mainstream techniques exploit a feature vector to represent each instance in a training dataset, and optimize the feature extractor by conducting a pretraining task. However, vector based representation is not enough to preserve the richness and handle of the noisiness of multimedia data. Also, common pretraining procedures, such as the ImageNet classification task, only focus on a single type of discriminative information, which might be insufficient for certain applications. Thus, in this thesis, I explore two research directions addressing these issues.

In the first part of this thesis, I develop two new types of representation: a probability distribution and a linear subspace, for multimedia content. Compared with vector based representation, both of them are capable of dealing with the richness and noisiness of multimedia. To leverage the two types of representation in downstream tasks, it is essential to design particular algorithms and training strategies. In this part of the thesis, I introduce methods incorporating distribution and subspace representations with deep neural network architectures, which can be optimized in an end-to-end manner. The experiment results on downstream tasks show that two proposed representations yield better performance compared to mainstream vector based methods.

In the second part of this thesis, I investigate style and content disentanglement techniques, which explicitly preserve different factors within multimedia content during the representation learning process. The disentangled representation provides better intepretability, and enables the manipulation of hidden factors in data synthesis scenarios. Based on this motivation, I propose two methods to effectively separate the hidden factors in multimedia data. The first method models the relation between style and content as a simple matrix operation in hidden feature space. The second method minimizes the mutual information between two hidden factors by formulating an adversarial training criterion. The advantages of the two proposed methods are evaluated in qualitative and quantitative experiments of data synthesis/generation tasks. Besides, I propose to demonstrate the applicability of these methods by conducting supervised training tasks with generated data.

https://drive.google.com/file/d/1r8fDmoLbhEVdk_FpYpinP2lTP4Ln8UAl/view