Statement of Purpose

Leilan Zhang
zhangleilan@gmail.com

My primary research interests lie in the field of representation learning, topic modeling, and knowledge extraction. As for my career goal, I plan to continue research after my Ph.D. as a researcher in industry or in academia.

1 Motivation

The neural topic model is my main focus during my master's study. Topic models describe documents using two distribution matrices: the document-topic distribution and the topic-word distribution. The statistical topic model Latent Dirichlet Allocation (LDA)[1] performs approximate inference by variational Bayesian or collapsed Gibbs sampling method. More recently, neural networks have been used as topic models with the emergence of Variational Auto-Encoder (VAE)[2], in which the inference can be carried out easily via a forward pass of the recognition network. These models consist of an encoder network that maps the Bag-of-Words (BoW) input of a document to a document-topic vector in latent space and a decoder network that maps the vector to a discrete distribution over the words in the vocabulary.

Empirically, the representations of the documents in the latent space would affect the performance of topic extracting and inference. I have replaced the prior distribution of WLDA[3] with Gaussian mixture distribution, which proved to be more effective on short dialogue corpus. This success aroused my interest in further explorations of representation learning in a more general case.

The pre-trained language model BERT[4] has brought the representation learning of text into a new era. Recent researches have given some interpretations of BERT, e.g. BERT's intermediate layers encode a rich hierarchy of linguistic information, with lexical features at the bottom, syntactic features in the middle and semantic features at the top. However, some questions to BERT remain unrevealed. For example, (1) what exact level of information is contained in each layer of BERT? (2) What is the distribution of a language in BERT? (3) How to improve BERT's representation ability except for larger capacity? (4) Can the BERT representation be disentangled to enhance its interpretability? I am also curious about how to utilize the knowledge learned by BERT. For example, (5) how to utilize the language knowledge of BERT for advanced NLP tasks, e.g. semantic analysis and coreference resolution, (6) how to extract the knowledge (real-world knowledge) from BERT for knowledge graph construction.

2 Proposed Research Plan

Some studies have shown that the direct use of feature vectors of BERT without fine-tuning could result in low performances in tasks like semantic similarity computing[5]. This problem attracts me and I tend to assume that it is the unnormalized distribution of BERT's latent representation space that leads to this problem. Figure 1 displays the visualization result of latent space of BERT's layer-1 and layer-12 respectively I conducted on THUCNews[6], which is a short news text dataset and the categories are indicated by colors. The figure indicates that the latent distribution is complicated and unnormalized among layers even for a simple text

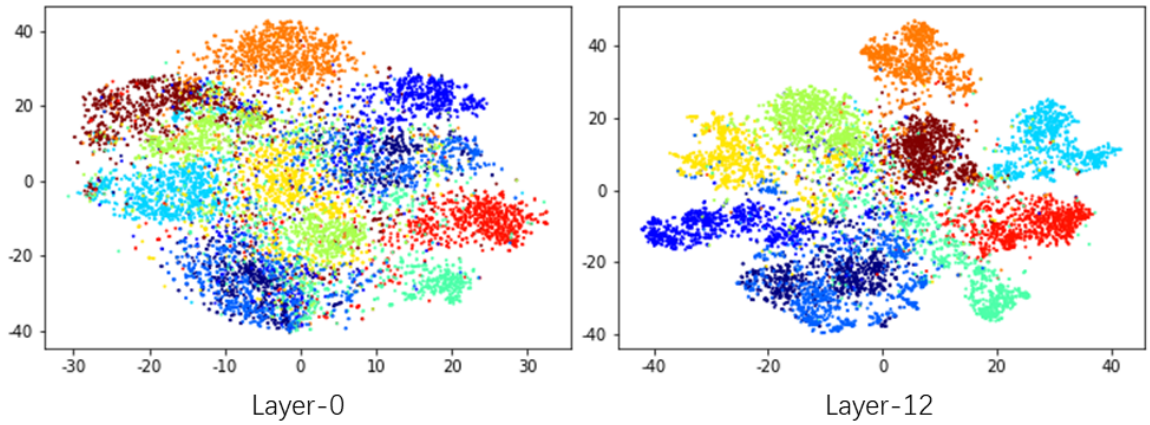

Figure 1: Latent Space Visualization of BERT on THUCNews

dataset. Then the successful use of the flow-based model in speaker recognition tasks inspired me when I am tackling this problem.

Flow is a generative model, which calculates the probability by directly computing the integral $P_G(x) = \int p(x|z)p(z)dz$. As Figure 2 shows, flow model maps data $\mathbf{x}$ to some specific distribution in latent space $\mathbf{z}$ through a series of invertible transformations, which meanwhile preserves the dimension and likelihood[7][8][9].

The speaker recognition task usually adopts cosine similarity to measure the distances between the feature vectors of speakers. However, the performance of cosine similarity drops sharply when the latent distribution is unnormalized or complex. Flow model can normalize the latent distribution to a standard Gaussian distribution, which is more suitable for cosine similarity. Inspired by this, my hypothesis is that by normalizing the complicated latent distribution of BERT to Gaussian distribution by the Flow model, the performance of the tasks involving semantic similarity will be improved. Moreover, since the standard Gaussian distribution is conducive for disentangling the representation to factors, my approach will be exploring possible interpretations of these factors with the linguistic concepts in semantics, or, to put it another way, whether the disentangled dimensions correspond to different aspects of semantic characteristics such as sentiment and topic. My further approach will be exploring the effective way of controlling these factors in generative tasks.

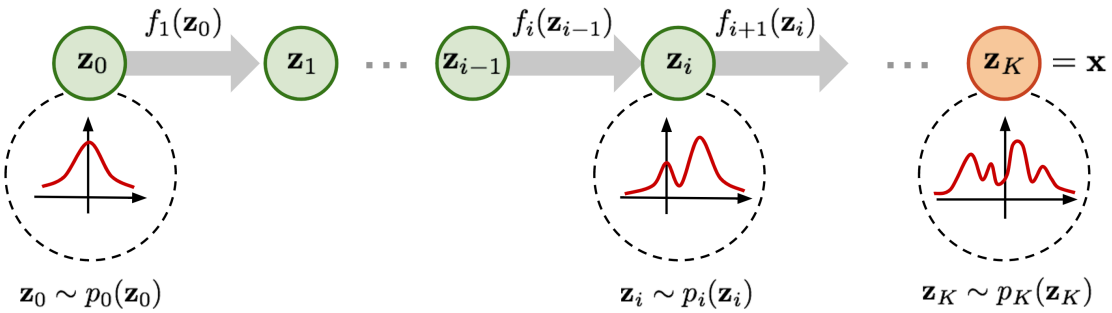

Figure 2: Normalization Flow

My next step plan is to adopt a Gaussian mixture as the prior distribution for the Flow model. Since the original Flow model utilizes standard Gaussian as prior, and it will transform the BERT's representation into a single-modal Gaussian distribution, which is not suitable for discriminative tasks like classification or clustering. Therefore, I want to explore the flow model with Gaussian mixture prior, introducing categorical information as prior knowledge in a supervised or semi-supervised way to form a discriminative generative model. In this way, clusters in the latent space can be formed according to the prior categories. Previously, I

have constructed the Wesserstein Topic Model (WTM) based on the Gaussian mixture prior. Compared with the neural topic model employing the single Gaussian prior, the GMM-based WTM extracted topics with significantly improved diversity and coherence. For similar reasons, I believe that the representation produced by this flow model of Gaussian mixture prior will be more conducive to downstream clustering or classification tasks.

Although the multi-modal Gaussian distribution is good for the discrimination of categorical text, it could result in a poor performance for semantic similarity, since the between-class distribution has no prior and it may be very complicated. Therefore, the next step is to explore the way to constrain the between-class distribution with a prior to from the entire latent space in a Gaussian distribution. In this way, the discriminative ability and similarity of the semantic representation of the text can be ensured at the same time.

For application, I plan to apply the improved representation to unsupervised word segmentation on ancient Chinese. It is not a trivial task since Chinese characters are continuous in a sentence, which requires cut into words before further processing. If an algorithm that can accurately detect the boundary of words relies only on pre-trained representation, it will undoubtedly save lots of time for the annotation of data as well as has a wide range of application.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[3] Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. Topic modeling with Wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6345–6381, Florence, Italy, July 2019. Association for Computational Linguistics.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[5] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019.

[6] Maosong Sun Jingyang Li. Scalable term selection for text categorization. *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 774-782, 2007.

[7] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10215–10224. Curran Associates, Inc., 2018.

[8] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[9] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. *CoRR*, abs/1605.08803, 2016.