Does each layer and mechanism in BERT is necessary? Let's identify!

Xueren Ge
gexueren@gatech.edu

Haibei Zhu
hzhu389@gatech.edu

1 Motivation

Text classification methods have attracted significant interest because of their many applications. Deep learning, utilizing big data to eliminate human definitions in rule design and feature extraction, has rapidly become the methodology of choice for text classification in the past ten years. Recently, transformer models, such as BERT (Devlin et al., 2019) and it's variations tBERTa (Peinelt et al., 2020) and ALBERT (Lan et al., 2020) have utilized contextualized word representations and achieved excellent results over a wide range of NLP tasks. However, far too little attention has been paid to classifying semantically similar texts. Due to close semantic similarities, labels that ought to be different are classified under the same label name. To reslove this issue, we will address BERT's anisotropy problem (Ethayarajh, 2019) and analysis of hard examples.

BERT's inherent anisotropy attribute can cause similar words' representations to greatly differ from each other due to their occurrence frequency in the batch set. Through empirical probing over the embedding we argue that the cosine similarity cannot represent semantic similarity directly. Anisotropy makes word embedding appear like a cone in the vector space. The uniformly sampled sentences' cosine similarities were close to 1, where sentence representations from a particular layer were anisotropic. Therefore, it is inexact to use only cosine similarity for distinct representations of sentences.

2 Goal

We will analyze each layer of BERT and provide detail visualization and newly technologies to alleviate and further solve anistropy problem.

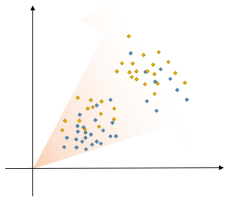

Figure 1: The sentences embedding space of BERT. "Anisotropic" makes word embedding present like a cone in the vector space

3 Plan

- Visualization & Measure of Contextuality
  (1) In paper "BERT, ELMo, GPT-2: How contextual are contextualized word representations?", author proposed new 3 definitive measure of contextuality, Self-Similarity (Self-Sim), Intra-Sentence Similarity (IntraSim), Maximun Explainable Variance (MEV), we are planning to use these three measurements as representation of isotropy of embeddings. To be more specific, we are going to see how will the anisotrpy change in different layers of BERT, then visualize and interpret the anisotrpy problem.
  (2) We will also visualize the hidden features in each layer by using feature transformation method like PCA, Autoencoders and etc. We will consider other methods to do visualization and compare them with each other. And based on the visualization result, we are planning to change some mechanisms in BERT, like the attention module. For example, if it uses channel attention and position attention, we first planned to remove channel attention and see how visualization changes. Then, we will drop some layers or add some "noise" to the data sets. Doing this will help us understand the role of each module/layer in

BERT architecture plays. By fine-tuning the BERT, our aim is to identify the best module to achieve overall performance.

- Contrastive learning
  Contrastive learning aims to learn effective representation by pulling semantically close neighbors together and pushing away semantically different pairs. We plan to incorporate contrastive learning models as a solution to the anisotropy problem. In SimCSE, the author incorporated annotated pairs from natural language inference datasets into contrastive learning framework by using "entailment" pairs as positives and "contradiction" pairs as hard negative. By combining SimCSE, we will visualize whether the measurement metrics of contextuality will change better or not, which means could it solve the anisotropy problem as it described.

- Dataset
  First, We plan to use NTCIR-13 MedWeb dataset, the task provides pseudo-Twitter messages (in Japanese, English, and Chinese) with labels for eight diseases/symptoms such as influenza, diarrhea/stomachache, hay fever, cough/sore throat, headache, fever, runny nose, and cold. The reason for choosing this dataset is that it's a multi-linguistic dataset.
  And we also planned to use another dataset, that is Stanford Natural Language Inference (SNLI) Corpus. SNLI corpus (version 1.0) is a collection of 570k human-written English sentence pairs manually labeled for classification tasks with the labels entailment, contradiction, and neutral (can be seen as positive, somewhat positive, negative sentences pair). It is a widely used benchmark for Natural Language Inference. We can use SNLI to better analyze the visualization result.

- Computing Resource
  First, we planned to use our own PC as the main computing resource, our compute is armed with Intel Core i7, NVIDIA GeForce RTX 3060 Laptop GPU 6GB GDDR6, 16GB RAM, 512 GB SSD. And in addition, if our computing resources is not enough, we planned to rent Google Colab as our secondary resource.

- Labor Division
  For dataset, Haibei Zhu will mainly for collecting both dataset mentioned above.
  Xueren Ge will take the responsibility for visualization part, mainly including implementing Self-Sim, IntraSim, MEV in each layer and also design dimension reduction methods to visualize what's going on in each layer.
  Haibei Zhu will take the responsibility for SimSCE part, mainly investigating in how to combine contrastive learning into BERT and validate that anisotropy problem can be solved by using contrastive learning.
  Both of us will play an important role in analysizing different module of BERT, validate the necessity of each module, maybe come up better ideas to improve some module.

- Backup strategy
  Suppose we fail in the experiment with BERT for lack of computing resources, unreliable visualization results, etc. We will incorporate small language models like biLSTM and Elmo and tried to validate our thinking: what's going on in each layer of the architecture, and does every module play a part?

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference of Learning Representations*, pages 1909–1942.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tBERT: Topic models and BERT joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055.