CS4650 Midway Report: Interpret, Visualize BERT

Xueren Ge
gexueren@gatech.edu

Haibei Zhu
hzhu389@gatech.edu

1 Abstract

In semantically similar sentence-level classification, BERT often neglects the subtle semantic differences in sentences and performs poorly because of its inherent anisotropy regarding semantic space. Our project proposed to froze some of the BERT's layers and use contrastive learning to solve the anisotropy problem. We will visualize BERT to see the above changes could let BERT to capture the subtle differences between semantically similar sentences.

2 Goal

Previously, Our goal was to find the "anistorpy" problem within BERT. Visualize it and further find ways to solve it. Now, we have done some part of the visualization and found out the result from (Ethayarajh, 2019) is correct. And we validated that word embedding generated from BERT is reasonably contextualized. And since we identified it, there is an interesting finding about layer 12(In Preliminary Results). We slightly changed our goal and will use the pre-trained BERT model from Hugging Face. In this way, we could freeze some layers and further identify methods to understand BERT better, like what layer are important.

- Tackle "Anisotropy" problem within BERT
- Understand BERT: Explore the importance of each layer in BERT; Does BERT really learns reasonable contextualized word embedding?
- Build better BERT from data and structure perspectives Structure

3 Progress made

3.1 Data

The statistics on dataset is shown on Table 1.

SNLI: SNLI corpus (version 1.0) is a collection of 570k human-written English sen-tence pairs manually labeled for classification tasks with the labels entailment, contradiction,and neutral (can be seen as positive, some-what positive, negative sentences pair). It is a widely used benchmark for Natural Language Inference. We can use SNLI to better analyze the visualization result. It serve both as a benchmark for evaluating representational systems for text, especially including those induced by representation-learning methods, as well as a resource for developing NLP models of any kind [1].

MedWeb: Sentences regarding different disease labels using Chinese and English[2]. The datasets provides pseudo-Twitter mes-sages (in Japanese, English, and Chinese) withlabels for eight diseases/symptoms such asinfluenza, diarrhea/stomachache, hay fever,cough/sore throat, headache, fever, runnynose, and cold. The reason for choosing this dataset is that it's a multilinguistic dataset so that we will have better ideas on how BERT performs on the different languages. (We have not yet use this dataset but we plan to investigate multilingual data in the future of our project)

Data	Classes	#Dataset	#Test
SNLI	3	$500k$	$70k$
MedWeb(C)	8	2560	640
MedWeb(E)	8	2560	640

Table 1: Basic statistics of the datasets. Dataset means the size of the dataset. Here the MC is short for MedWeb Chinese, MEis short for MedWeb English

The sentence examples shown on Table 2.

- Entailment: the hypothesis is a sentence with a similar meaning as the premise.
- Contradiction: the hypothesis is a sentence with a contradictory meaning.

[1] https://nlp.stanford.edu/projects/snli/

[2] http://research.nii.ac.jp/ntcir/permission/ntcir-13/perm-ja-MedWeb.html

- Neutral: the hypothesis is a sentence with mostly the same lexical items as the premise but a different meaning[3].

3.2 Method

We use to methods that stated in the paper (Ethayarajh, 2019) and reproduce it to fit our datasets.

3.2.1 self-similarity

The word w in layer l is the average cosine similarity between its contextualized representations across its n unique contexts. That means if the word has more contextualized representations, the self-similarity score should be lower.

$$SelfSim_l(w) = \frac{\sum_j \sum_{k \neq j} cos(f_l(s_j, i_j), f_l(s_k, i_k))}{n^2 - n} \quad (1)$$

3.2.2 Intra-sentence similarity

It is the method to measure the average cosine similarity between its word representations and the sentence vector. If $IntraSim_l(s)$ is high but $SelfSim_l(w)$ is low that means the words in the sentence are less contextualization. We want to see our words have both lower $IntraSim_l(s)$ and $SelfSim_l(w)$.

$$IntraSim_l(s) = \frac{\sum_i cos(\overrightarrow{s}_l, f_l(s, i))}{n} \quad (2)$$

where

$$\overrightarrow{s}_l = \frac{\sum_i f_l(s, i))}{n}$$

3.2.3 Maximum explainable variance

It indicates the proportion of variance in w's contextualized representations for a given layer that can be explained by their first principal component. If MEV_l is equal to 0, it means that static embedding will be poor. If MEV_l is equal to 1, static embedding can replace for contextualized representations.

$$MEV_l(w) = \frac{\sigma_1^2}{\sum_i \sigma_i^2} \quad (3)$$

3.3 Preliminary Results

In Figure(1), We randomly select 2 words and calculate the cosine similarity, and find that as layer goes deeper, any words cosine similarity grows, it indicates Any 2 random words have similar representation, which means, the whole vector space tend to be a cone instead of ball. We have validated that BERT do has problem of 'anisotropy'.

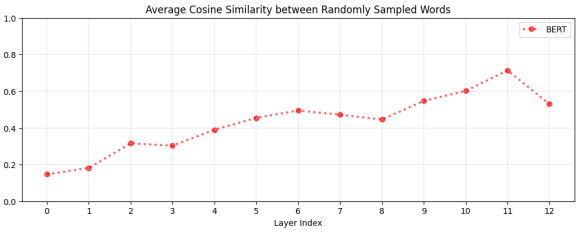

Figure 1: Anisotropy

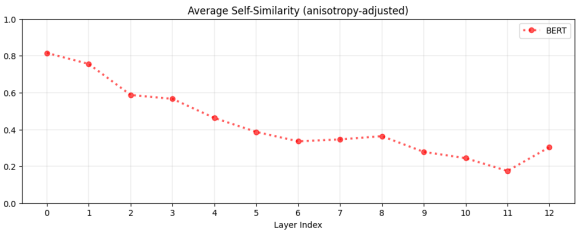

Figure 2: Self-Similarity

In Figure(2), we found that the higher the layer, the lower the self-similarity, suggesting that contextualized word representations are more context-specific in higher layers. It validates that for each word, its corresponding word embedding generated from BERT is contextualized. This is what BERT intends to do.

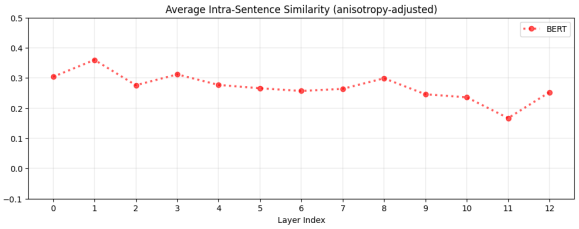

Figure 3: InstraSim

In Figure(3), we found that As layer goes deeper, 2 words in the same sentence have lower Intra-sentence similarity, which implies that 2 words don't have the same meaning even when they share the same context. It also shows BERT's contexualization for each word is reasonable.

In Figure(4), we randomly select samples and plot its MEV, for example, 'beverage' is almost 1, which means in all context , in all layers its meaning is almost the same. And can be replaced by static embedding. While 'go', 'school' have low

[3] https://trishalaneeraj.github.io/2017-12-22/semantic-entailment

Table 2: A few example pairs taken from the development portion of the corpus. Each has the judgments of five mechanical turk workers and a consensus judgment.

Text	Judgments	Hypothesis
Two women are embracing while holding to go packages.	neutral	The sisters are hugging goodbye while holding to go packages.
Two women are embracing while holding to go packages.	entailment	Two woman are holding packages.
Two women are embracing while holding to go packages.	contradiction	The men are fighting outside a deli.

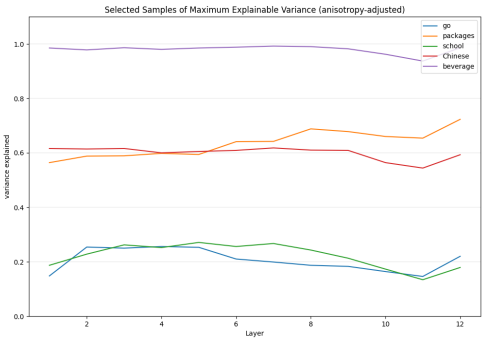

Figure 4: Selected Sample Words from Datasets

MEV which implies multiple meaning in different context.
There are a interesting finding, in Figure(2) and Figure(3), the $IntraSim_l(s)$ and $IntraSim_l(s)$ both increase, which indicates that word embedding in layer 12 will make words less contextualized. And also, 2 different words in the same sentence tends to have the same word embedding. We doubt the rationality of layer 12 based on our findings and will do more experiments.
In conclusion, I think we addressed some part of project goal, for example, we visualiza the problem of anisotropy in BERT and validate that as layers go deeper, BERT do make each word more contextualization. Besides, we also start to doubt the some layers(layer 12) may do harm to the whole performance.

3.4 Work Division

Xueren, Ge was in charge of implementing metrics defined in the paper, drew plots figures using matplotlib. Haibei, Zhu was in charge of letting the BERT read the dataset. Both team members analyzed the final result and discussed it together.

4 Plan to complete the project

4.1 Future tasks

We have several threads to explore further.
The first thread is to solve anisotropy problem in BERT. And we still plan to think how to combine contrastive learning to solve this problem.
The second thread is that We will further to understand why BERT work, especially in understanding each layer's importance, it will includes 2 parts,

- Layer 12
  Right now, through our analysis and argument, we think layer 12 may introduce bad performances, we decide to use Hugging Face to freeze layer 12 and do experiments to see how metrics (self-similarity, intra-sentence similarity and maximum explainable variance) we defined before changes.

- Other layers
  We will use Hugging Face to freeze these layers separately and see their importance to BERT.

The third thread is to explore how to better BERT, we plan to consider this problem from the perspective of dataset and architecture.

- Biased Dataset
  We will try to create a biased dataset based on current Stanford Natural Language Inference (SNLI) Corpus and also MedWeb. And then see how it will influence BERT's performance. Doing this experiment is meaningful because it will help us to know what kind of dataset we should collect before training.

- Architecture
  In this section, we will use what we find in thread 2, and then try to adjust some layers to see whether it will help to improve BERT performances

4.2 Project changes

We have several changes related to the data sets and the way to analyze BERT. Originally, we planned not to froze any layer of BERT because we thought both layers are important. But since we found out layer 12 may led to bad performances we plan to freeze some layers to see if it could led to a better performance of BERT and validate its reasonability.

4.3 Work division

Xueren, Ge's future work will mainly focus on freeze some layer of BERT and using datasets to visualize BERT's performance. Besides, he will focused on how to create biased-dataset and continued to see how bias infects BERT. Haibei, Zhu's future work will mainly focus on contrastive learning. It is a new methods that stated to be able to solve the anistropy problem. We will work together to anaylze the result and fine-tune our model.

5 Appendix

our code and paper can be found from this github

References

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.