Exploring the Role of Contextualization in Dynamic Contextual Word Embeddings

Anubhav Agarwal, Rohan Jinturkar, Henrique Schechter Vera[1]

[1]Princeton University

Objectives

This project aims to verify the two fundamental assumptions of the Dynamic Contextualized Word Embeddings (DCWE) paper:

- BERT is an adequate modeling framework to capture the benefit derived from contextual embeddings.
- Contextual embeddings are better than non-contextual embeddings in capturing meaning, especially when paired with dynamic components.

Notably, the authors address the latter of these issues in their related work, by proving these statements true in cases separate from dynamic embeddings. However, in the scope of their ablation studies, they neither consider other contextualizers, nor verify that their dynamic system "prefers" contextual embeddings to non-contextual ones.

Introduction

Hofman, Pierrehumbert, and Schutze [1] propose a novel mechanism for capturing two disparate factors in the modeling of language: the context of the sentence in which the word occurs, and the variability of words in different temporal and social contexts. In order to do so, they inject a temporal and social element to vectors that are derived contextually using the BERT model as their primary means of contextualization. To verify (1) the importance of this contextual element, and (2) the ability of BERT to accomplish it, we train a variety of non-contextual (Word2Vec and GLoVe) and contextual (GPT-2 and RoBERTa) models.

PRINCETON UNIVERSITY

Methods

In addition to the BERT baseline, we evaluate the DCWE layer on Word2Vec, GloVe and RoBERTa for a downstream sentiment analysis task.

- **BERT** Devlin et al. [2] introduce BeRT as a contextual transformers model, pretrained with two objectives: MLM and next sentence prediction.
- **Word2Vec** Mikolav et al. [3] propose Word2Vec as a two-layer neural network model to learn global word associations from text, where each word is represented by a unique vector.
- **GloVe** Pennington et al. [4] develop GloVe, an unsupervised algorithm to develop word representations. Training is performed on aggregated global word co-occurrence matrices, using the observation that co-occurrence probabilities encode some meaning.
- **RoBERTa** Liu et al. [5] present RoBERTa, a bidirectional contextual transformers model using a MLM objective directly in pretraining.
- **GPT-2** Radford et al. [6] build GPT-2 as a self-supervised bidirectional transformers model that was trained using causal language modeling (predict next word in sentences).

DCWE Architecture

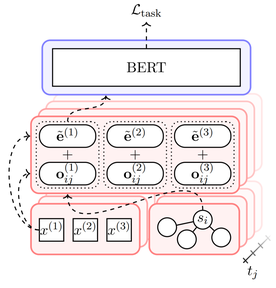

Figure 2: Model architecture. Words are mapped to dynamic embeddings by the parts of the dynamic component ( ), which are then contextualized by the contextualizer ( ). The output of the contextualizer is used to compute the task-specific loss $\mathcal{L}_{\text{task}}$.

[1]

Experimental Setup

The model uses the contextualizer to extract an initial set of embeddings, which is usually done using the pretrained in-model look-up table. It then calculates a dynamic embedding. These are then used as inputs to the whole contextualizing model.
First, we run the baseline in order to compare the performance on sentiment analysis, using primarily the F1-score. Then, we run the model without the contextualizer, to understand the quality of the non-contextual embeddings. Finally, we replaced the contextualizer with 2 different models, to verify the quality of BERT.
All code was based on the existing codebase, with modifications in the creation of the model. All models are trained on the Yelp dataset. Code was ran on a GPU-accelerated Colab Pro instance.
GitHub: github.com/rjintu/cos484final/

Results

The table containing the initial development and test F1-scores for both the GPT-2 and BERT models is below. These results were obtained after training on one epoch of data, through 452,000 examples of reviews in each epoch. These are the scores for when the model uses a dynamic layer as well.

Model	Dev F1	Test F1
GPT-2	0.5024	0.5023
BeRT	0.8813	0.8799

Table 1:F1-Scores on Sentiment Analysis Task, Yelp Dataset

Conclusion

DistilBeRT seems to be more performant than distilled GPT-2. This may be a result of GPT-2's autoregressive nature, which prevents the model from creating bidirectional embeddings. However, GPT-2 does have more parameters, so its performance may be hampered by the limited training time.

References

[1] Valentin Hofmann, Janet B. Pierrehumbert, and Hinrich Schütze.
Dynamic contextualized word embeddings, 2020.

[2] 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.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean.
Efficient estimation of word representations in vector space, 2013.

[4] Jeffrey Pennington, Richard Socher, and Christopher D Manning.
Glove: Global vectors for word representation.
In *EMNLP*, volume 14, pages 1532–1543, 2014.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.
Roberta: A robustly optimized bert pretraining approach.
ArXiv, abs/1907.11692, 2019.

[6] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever.
Language models are unsupervised multitask learners.
2019.

Acknowledgements

Many thanks to Karthik Narasimhan for his feedback on the project proposal and Valentin Hofmann for providing additional context on the DCWE paper.