Tech Review of BERT (Pre-training of Deep Bidirectional Transformers for Language Understanding)

Author: asankar2@illinois.edu

1. Introduction

BERT[1] is a new language representation model created by Google in 2018 which achieved state of the art results on several natural language processing tasks. Pre-Training a language model on a large corpus of data has been shown to be a good starting point for solving many NLP problems. Once a model is pre trained (semi-supervised learning), it can be fine-tuned for specific data sets for better performance (supervised-learning). Older models such as ELMo[2] used a deep bi-directional LSTM (Long Short Term Model) approach to do *feature learning* to accomplish this task. Newer ones such as GPT[3] used the *fine-tuning* approach with the transformer architecture to improve accuracy of narrow tasks. BERT uses a combination of ideas from both papers and uses bidirectional transformers to train the model, which showed big accuracy gains in a variety of tasks.

2. ELMo

Word embeddings is the name given to the vector representation of tokens in a document that are trained by models and are used in the neural network to classify words and compute probabilities. Traditionally, word embeddings were computed by training a model on tokens reading left to right. With the sentence "*I love natural language processing",* the embedding (vector) of the probabilities of the word *language* would be computed against the probabilities of the words "*I love natural"* before it. (ex. What is the probability the next word is the language if the first three words are "*I love natural*")

The ELMo paper showed that there is a significant advantage to train your language model bidirectionally, by reading a sentence right-to-left AND left-to-right. In the previous example, the LTSM model would be trained by looking at the probabilities before and after the token in question to build a forward and backward LM. Then, it tries to maximize the log likelihood of summing the two probabilities of seen tokens. This approach was a breakthrough and improved the state of the art considerably.

Forward LM

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_1, t_2, \ldots, t_{k-1}).$$

Backward LM

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_{k+1}, t_{k+2}, \ldots, t_N).$$

3. Transformers and GPT

After the release of the Transformers[4] paper, research on neural nets for NLP shifted to using this new architecture. In the "Attention is all you Need" paper, the Google team shows that

a NN architecture with an encoder step (layer of networks to create embeddings) and a decoder step improves accuracy even further. This is different from the historic approach of using LTSMs, like the ELMo paper. This approach can be applied to both RNNs and CNNs, reducing both the cost of training and the time taken to fully train the models.

Open AI used this approach in their language representation model GPT. After training the model in a large corpus of data, it is fine-tuned on specific tasks on labeled data sets. The transformer model really helped here as the number of parameters for this model approached 1.5B so efficiency of training and reduced costs were very beneficial. This was also one of the first successful applications of transfer learning in the NLP space. It is important to note that unlike ELMo, GPT was trained on only a forward language model.

4. BERT

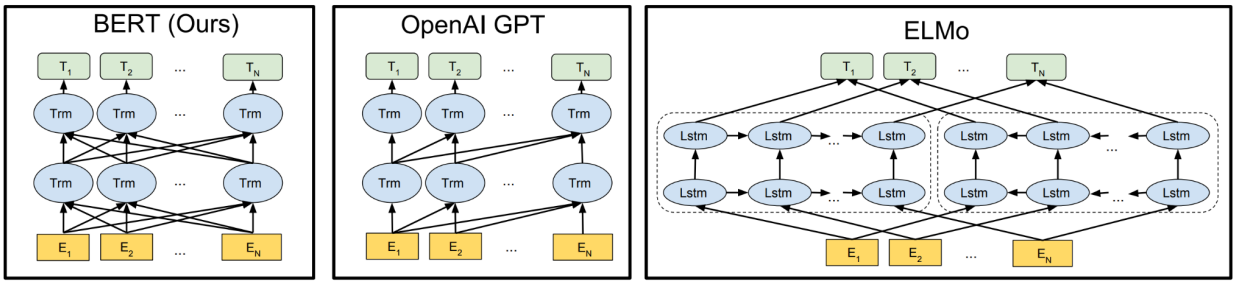
BERT is the next generation language model created by Google. It combines a lot of concepts from previous papers to create a cutting edge model that beat out both GPT and ELMo in 11 classical NLP tasks.

BERT is a pre-trained language model on a large corpus of data that is later fine tuned on a smaller labeled dataset. Like ELMo, BERT is trained *bidirectionally* by jointly conditioning on both left and right contexts. Similar to GPT, BERT uses transformers to train the model, so there are encoder and decoder neural networks. This presents an issue, since the hidden layers of the left and right context (inputs to the NN) might leak information about the token that is getting trained.

BERT solves this problem by using a masked language model. This means that 15% of the input words are “masked”. With 80% probability, the word will be omitted, 10% probability the word will be replaced with another word and 10% probability the word will be left alone. Then, we train the model to fill in the gaps and optimize the language model like so. Note that this is taking unlabeled data and transforming it into a supervised learning task.

After creating a pre-trained language model with bidirectional transformers, BERT will fine tune on datasets similar to GPT. Fine tuning is shown to be relatively less expensive than pre-training and can be used on a variety of different sources. The paper shows various data sets and claims that adding a single layer to the NN should be sufficient to fine tune the data for each task.

Model Representations



5. Future

After creation of BERT, researchers from Google AI team and Carnegie Mellon University noticed a weakness in the model where it requires left and right context to determine

a token, which they deem as a corrupted input. Their proposed solution to this problem was to train on various permutations of a sentence and their new model performed even better than BERT on a variety of NLP tasks. This is the XLNet paper. [5]

GPT-3 is the newest language model from Open AI which has also performed much better than BERT. This new model has 175B(!) parameters and is much larger than the BERT model in the paper (nearly 400x the size). This model is considered state of the art. Note that this model is not open sourced, while the others are.

6. Conclusion

The BERT language model was a very influential paper in the natural language processing space and was state of the art for its time. It used various ideas from its predecessors such as Bi-Directional training from ELMo and Transformer architecture from GPT. Future state of the art models like XLNet rely on BERT too and BERT still has some advantages over the current state of the art models, such as fast tuning, relatively small model sizes, and the benefits of being an open source model. Even though BERT may not be the most accurate model in modern day, it remains a benchmark for future models to beat, which speaks to its influence on the NLP research space.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Peters, Matthew et al. “Deep Contextualized Word Representations.” Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (2018): n. pag. Crossref. Web.

[3] A. Radford, et al., “Improving Language Understanding by Generative Pre-Training,” 2018. [Online].

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010

[5] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pre-training for language understanding. arXiv preprint arXiv:1906.08237, 2019.