Cody Webster
October 26, 2020

Tech Review: BERT

The purpose of this review is to produce an overview of a relatively new language representation known as Bidirectional Encoder Representations and Transformers (BERT). In research literature BERT has been described as “conceptually simple and empirically powerful” (Devlin et al., 2019). It has achieved this distinction through its ability to handle a wide range of natural language processing (NLP) tasks through minor modifications to the original model. The performance of a BERT model has been empirically shown to exceed the performance of current state-of-the-art models and algorithms in common benchmarks.

As the research and institutional knowledge of language representation has progressed it has become ever more apparent that pre-training is an effective way to improve performance on natural language processing models (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). The two current most prevalent methods of pre-training, feature-based and fine-tuning, both rely on unidirectional language models to learn the representation of a language. The purpose of the BERT method is to alleviate the restrictions that these common approaches have. The available choices in pre-training architecture is one of the major limitations of the unidirectional language models because it limits the tasks that the models can be trained to perform.

BERT attempts to remove the constraints on unidirectional language model by using a masked language model (MLM). The MLM attempts to train the model by randomly removing words for the training data and requiring the model to fill in the missing text through context analysis of the surrounding terms. The BERT models also introduces the “next sentence prediction” step that attempts to train the model on the relationships between pairs of text within the training data.

The BERT model contains two main steps, pre-training and fine-tuning. Pre-training of a BERT model consists of running the model over unlabeled data and through various tasks set forth by the developer. The fine-tuned parameters of a BERT model are unique to every potential implementation of the model. Even though each variation of fine-tuning is unique they all use the same initial parameters that are generated from the pre-training step. There are only minor changes between the pre-trained model the generated fine-tuned model that is used for the actual NLP task.

The architecture of a BERT model is multi-layered. This multi-layered model is a bidirectional Transformer encoder and its implementation as released in the tensor2tensor library. There are two model sizes, base and large, described within the paper (Devlin et al., 2019). The base model has 12 layers with a hidden size of 768 and 12 self-attention heads. This base model has a total of 110 million parameters. The large model has 24 layers with a hidden size of 1024 and 16 self-attention heads. The large model has 340 million total parameters. The model sizes chosen for the paper were selected based on the ability to compare them with existed published data.

The BERT model is capable of handling inputs in two formats. It can handle either a single sentence or a pair of sentences. For description purposes a BERT model’s inputs are generally described as sequences which can refer to either input option and their tokenized representation. Each token sequence is represented using the vocabulary of the WordPiece embeddings (Wu et al., 2016). Special tokens are used to distinguish the start and end of a sentence as well as when two sentences are contained within the same sequence.

Traditional language models can only be trained as left-to-right or right-to-left. This limitation is due to the ability of a word to "see itself" in a multi-layered approach. This results in the model being easily able to predict words in training data but ultimately failing in a real application. The bi-directional approach is only possible because of the MLM. The masking of random words prevents their easy prediction in subsequent layers and allows for the bi-0directional approach. The standard approach, as described in relevant literature, is to mask 15% of the tokens in any sequence (Devlin et al., 2019). The model is constrained to only predict the masked inputs instead of all tokens in the input sequence. The actual token that is used to mask the input token varies by set percentages and is used to account for the fact that the general masking token is not present in the fine-tuning step.

Next sentence prediction is an important feature in models because it relies on and demonstrates the ability of the model to understand the relationship between sentences. Traditional language models do not capture these relationships. Pre-training a model for next sentence prediction is easy because it simply relies on feeding the model sequences of sentences that are marked as either legitimate sequences or illegitimate ones. It is critical when pre-training a model that the sequences fed into it come from a real textual body rather than random combinations. Fine-tuning a model that is been pre-trained adequately is as simple as replacing the inputs with the actual inputs and outputs.

The BERT model has been tested against a number of different standards and metrics. The results of tests are discussed in (Devlin et al., 2019) and will be briefly summarized. BERT was tested against the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018a). The GLUE benchmark is a large collection of NLP tasks and is used as a benchmark for test state of the art NLP models. For GLUE, the results of BERT were compared to other state of the art models such as OpenAI GPT, Pre-OpenAI SOTA, and BiLSTM+ELMo+Attn. BERT outperformed all of these models.

BERT was also tested against both Stanford Question Answering Dataset (SQuAD) v1.1 and v2.0. The SQuAD tasks are to predict the answer of a question based on the text of the question if given a paragraph from Wikipedia. The answer is limited to once sentence in v1.1 and can encompass multiple in v2.0. When tested in both versions the BERT model was able to outperform the previous best scores by appreciable margins.

The last test described in the paper (Devlin et al., 2019) was Situations With Adversarial Generations (SWAG). The goal of SWAG is to predict the next term in a sequence based from a list of possible terms. BERT outperforms the comparison models.

Through the various tests performed against industry benchmarks the BERT model has been shown to be an effective, robust, and versatile model with state-of-the-art performance across a variety of NLP tasks. The major contribution of the BERT model is its bidirectional architecture that allows it to address a wide range of NLP problems. In the future new models can build on the approach laid out be BERT to further improve the performance of NLP models and use these improved models to address a wide variety of problems facing modern language models.

Cody Webster
October 26, 2020

References (all references included from original paper, paper highlighted in grey)

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1638–1649.

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. 2018. Character-level language modeling with deeper self-attention. arXiv preprint arXiv:1808.04444.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research, 6(Nov):1817–1853.

Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2009. The fifth PASCAL recognizing textual entailment challenge. In TAC. NIST.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In Proceedings of the 2006 conference on empirical methods in natural language processing, pages 120–128. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In EMNLP. Association for Computational Linguistics.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. Computational linguistics, 18(4):467–479.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo LopezGazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. arXiv preprint arXiv:1312.3005.

Z. Chen, H. Zhang, X. Zhang, and L. Zhao. 2018. Quora question pairs. Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In ACL.

Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1914– 1925.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning, pages 160–167. ACM.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Lo¨ıc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In Advances in neural information processing systems, pages 3079–3087.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In Proceedings of the Third International Workshop on Paraphrasing (IWP2005).

William Fedus, Ian Goodfellow, and Andrew M Dai. 2018. Maskgan: Better text generation via filling in the . arXiv preprint arXiv:1801.07736.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. CoRR, abs/1606.08415.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics. Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In ACL. Association for Computational Linguistics.

Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2018. Reinforced mnemonic reader for machine reading comprehension. In IJCAI.

Yacine Jernite, Samuel R. Bowman, and David Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. CoRR, abs/1705.00557.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In ACL.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In Advances in neural information processing systems, pages 3294–3302.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In International Conference on Machine Learning, pages 1188–1196.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In Aaai spring symposium: Logical formalizations of commonsense reasoning, volume 46, page 47.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In International Conference on Learning Representations.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In NIPS.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In CoNLL.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems 26, pages 3111–3119. Curran Associates, Inc.

Andriy Mnih and Geoffrey E Hinton. 2009. A scalable hierarchical distributed language model. In D.

Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, Advances in Neural Information Processing Systems 21, pages 1081–1088. Curran Associates, Inc.

Ankur P Parikh, Oscar Tackstr ¨ om, Dipanjan Das, and ¨ Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In EMNLP.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532– 1543.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In ACL.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In NAACL.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1499–1509.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In ICLR.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642.

Fu Sun, Linyang Li, Xipeng Qiu, and Yang Liu. 2018. U-net: Machine reading comprehension with unanswerable questions. arXiv preprint arXiv:1810.06638. Wilson L Taylor. 1953. Cloze procedure: A new tool for measuring readability. Journalism Bulletin, 30(4):415–433.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In CoNLL.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL ’10, pages 384–394.

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.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning, pages 1096–1103. ACM.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018a. Glue: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355.

Wei Wang, Ming Yan, and Chen Wu. 2018b. Multigranularity hierarchical attention fusion networks for reading comprehension and question answering. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. arXiv preprint arXiv:1805.12471.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In NAACL.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In Advances in neural information processing systems, pages 3320–3328.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining local convolution with global self-attention for reading comprehension. In ICLR.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision, pages 19–27.