70152 Research Tutorial - Report

Ruoyu Hu

1 Paper Summary

The paper, titled "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" was published to the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) by Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova of the Google AI Language team.

The paper primarily builds upon the architecture introduced in the preceding transformers paper (Vaswani et al., 2017), taking the transformer encoder architecture and performing *pre-training* on large, unlabelled corpora[1] in an unsupervised manner to learn language representations. The paper proposes two versions of the BERT model: $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$, containing 12 and 24 transformer blocks respectively. The two pre-training tasks introduced in this paper are *Masked Language Modelling* (MLM) and *Next Sentence Prediction* (NSP), both of which remain relevant to pre-training large language models as of the time of writing.

MLM is described in the paper as a pre-training task that "masks" 15% of token during training, these masked tokens are replaced 80% of the time with a special `[MASK]` token, 10% of the time with a random different word and 10% of the time is replaced with the original token. The intuition being that a model, in order to understand language, and thus produce high quality representations, should be able to distinguish whether a given word should be present in the sentence, and where it should occur. **NSP**, or binarised NSP introduced in this paper presents the model with a pair of sentences at training time separated by a special `[SEP]` token, and tasks the model with predicting whether the second sentence/segment is likely to be one that follows the first. This intuitively forces the model to learn the semantic information contained within the respective sentences to gain, to some degree, understanding of the semantic relation between the two sentences.

Classification in each pre-training task is performed by added task-specific parameters at training time that are not carried over to fine-tuning. These classifier layers take the computed token representations for each token and maps to the required output dimensions (i.e. n_{vocab} in MLM, 2 in NSP). In the case of sentence-level classification, the paper introduces a special `[CLS]` token, that is prepended to the beginning of the sentence, and whose token embedding forms a sentence-level representation as the nature of the scaled-dot-product attention mechanism introduced in (Vaswani et al., 2017) and used in this paper computes a token's representation with respect to the representation of every other token in the same sentence.

The paper utilises the pre-train then fine-tune paradigm that has become ubiquitous in NLP with regards to the uses of large, pre-trained language models (PLMs) in the years since. The model would be pre-trained in an unsupervised manner on vast amounts of data and then *fine-tuned* (Sun et al., 2019) with a much lower learning rate on a smaller amount of task-specific data. This approach has seen significant improvements in state-of-the-art across many tasks within the field, but have also seen a shift towards models of such size and trained on so much data as to be infeasible for individuals to attempt. For reference, the larger of the two BERT models proposed in this paper contained 350M parameters, whereas the largest GPT3 model contains 175B parameters Brown et al. (2020).

Nevertheless, the unsupervised pre-training of language representations has allowed for great improvements in the encoding of semantic information into sentence embeddings (Mikolov et al., 2013). An area in which this has been pivotal is multilingual NLP, where the alignment of high-quality sentence embeddings Ruder et al. (2019) has significant potential for transfer learning between high-resource to low-resource languages (Gritta and Iacobacci, 2021; Gritta et al., 2022), greatly reducing the digital divide in languages (Hu et al., 2020).

[1] BookCopus, EN Wikipedia, 3.3B words total

1.1 Benchmarks and evaluation

The paper evaluates its approach using 3 popular benchmarks around the time: GLUE (Wang et al., 2018), SQuAD (Rajpurkar et al., 2016) and SWAG (Zellers et al., 2018).

GLUE The *General Language Understanding Evaluation* benchmark is composed of 8 tasks across sentence classification, paraphrase detection and natural language inference. For the purpose of direct comparison with OpenAI GPT (Radford et al., 2018), this paper foregoes evaluation on a task (Winograd NLI) from GLUE that has historically been shown to be problematic. Both versions of the BERT model improves significantly over the SOTA set by OpenAI GPT at the time (+4.5% and +7.0% respectively). GLUE has since been replaced by SuperGLUE (Wang et al., 2019), a harder, more comprehensive benchmark since 2019.

SQuAD The *Stanford Question Answering Dataset* is composed of crowd-sourced question/answer pairs, whereby given a question and a passage, the model must derive the answer from the given passage. The best-performing BERT setup outperforms previous SOTA on SQuADv1.1 by +2.7F1, and the best performing BERT setup outperforms previous SOTA by +5.1F1 on the more difficult SQuADv2.0 dataset. It is noted in the paper, however, that the SQuADv1.1 dataset may not be a fair evaluation of QA capabilities, as SOTA figures from BERT has outperformed human-evaluation by +2.0F1.

SWAG *Situations With Adversarial Generations* is a benchmark for evaluating grounded common-sense inference. Given a sentence and a set of possible followups, the model is tasked with selecting the most likely sentence to follow. The large BERT model outperforms previous SOTA by +8.3%

Since the publication of BERT, the three reported benchmarks have all been replaced with more difficult tasks, as the reported SOTA figures have significantly closed the gap, and in some cases, surpassed human evaluation. It can be observed that almost all models at the top of the updated benchmarks are BERT-based models such as RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2020).

To summarise, the paper presented a transformer encoder model pre-trained in an unsupervised manner and demonstrated the effectiveness of representation learning in the pre-train-then-fine-tune setup that has become increasingly adopted in Natural Language Processing today. In the process significantly improving state-of-the-art performance in multiple benchmarks including Natural Language Inference, Question Answering and Named-Entity Recognition. The paper then provided extensive ablations over its setup, further justifying their design choices.

2 Discussion and Reflection

As part of the group discussion following the presentation, several points were discussed over the paper and the presentation.

2.1 Comparison to OpenAI GPT

The paper compared the BERT models' performance on many benchmarks against OpenAI GPT (Radford et al., 2018), which was the best-performing transformer-based pre-trained language model at that time, and had also outperformed many of the previous non-transformer SOTA scores on various benchmarks. Whilst both models shared similarities, it must be considered that GPT is a decoder-only model, which does not explicitly train to produce language presentation, but rather to minimise the *language modelling loss* or to produce quality textual responses. Additionally, it was not mentioned in the main body of the paper that GPT is pre-trained on significantly less data than BERT (800M vs. 3300M), it would have been a far more accurate comparison of a transformer encoder-only model (BERT) and decoder-only model (GPT) if the paper has shown ablation experiments using a BERT model that has only been pre-trained on BookCorpus, the same dataset as GPT.

2.2 Biased Data in Pre-training

It was pointed out during the discussion that due to pre-trained language models learning language structure from unlabelled data, biases present in the training data can be learned into the model. This can be unsuitable for healthcare applications, especially if the downstream task of the language model is sensitive in nature. We note that debiasing language models is an active area of research (Zhou et al., 2021; Meade et al., 2022), with many methods aimed at reducing bias in pre-trained encoder language models. For example: manually sanitising the pre-training data, enforcing de-biasing objectives concurrently to downstream tasks etc. Due to

the variety in active methods and the variety of metrics in measuring language model bias, it is difficult to report definitively the appropriate method that should be undertaken to ensure that bias inherent in pre-training has minimal impact in downstream tasks. This is nonetheless an incredibly exciting area of research, and it remains to be seen what directions it may take in the coming years.

2.3 Choice of Metrics

It was noted that both accuracy and F1 were used at different points in the paper, it was discussed whether this is symptomatic of cherry-picking results. We note that accuracy is commonly used across the field as the sentence classification metric and F1 likewise the metric for token-level tasks such as *named-entity recognition, part-of-speech-tagging* etc. The metrics used in the paper are also established benchmarks, as such it is very likely that the paper is simply using metrics that were most commonly used in previous works (Radford et al., 2018) to allow of direct comparisons.

2.4 Healthcare Applications

A question was raised during regarding the healthcare applications of BERT. Due to the pre-training tasks selected, BERT has a good level of performance on QA and paraphrase tasks that significantly improved upon previous SOTA, which can be used to aid clinical decision-making by fact-extraction and summarising large bodies of text. BERT's performance in aiding natural language understanding (NLU), also allows for applications in mental health support (Ji et al., 2021) and conversational agents for digital psychotherapy (Alazraki et al., 2021). Dialogue agents benefit from applications of BERT (or BERT-based models) in NLU by improving understanding of user utterances to provide more appropriate responses, thus improving user experience.

2.5 Bidirectionality of BERT

It was pointed out that the bidirectionality of BERT was not elaborated upon much in the presentation itself, despite being part of the name of the paper and its key contribution. It is noted that the transformer encoder architecture already computes attention over all tokens. The BERT paper however does extensive ablations over performance impact of bidirectionality, including comparing results to OpenAI GPT (Radford et al., 2018, 2019) and their own implementation of unidirectional models.

2.6 Presentation Quirks

Cohort feedback at the end of the presentation reported several points of improvements that should be considered in future presentations of this nature. It was reflected that the speaker was speaking at a pace that was considered by some to be difficult to keep up with, the speaker should be more aware of pacing of the presentation moving forward. Another issue raised was the abundance of text and tables in sections presenting the benchmark scores from the paper. The table captions were lifted verbatim from the paper in order to explain what each table was showing, it was reflected that this actually made the tables harder to interpret in the presentation, as the actual numbers were much smaller as a consequence, and a majority of cohort members were unfamiliar with the benchmarks used. It was suggested for the presenter to focus more on highlighting the important contributions from the paper within the tables, and to avoid over-cluttering slides of this nature in the future. Additionally, it was raised that the references show in most of the slides are in author-year format, instead of in full. We note that the references are show in full in the bibliography section, and the author-year format is a quirk of the slides being made in latex. Nonetheless this will be improved upon in future presentations.

3 Summary

The presented paper by (Devlin et al., 2018) introduced methods that have since become widely used in many application across Natural Language Processing today, significantly improving the state-of-the-art in many popular NLP benchmarks in areas including NLI, QA, NER etc. The paper provides extensive ablations over their methodology, contributing to the widespread adoption of many of their methods in current literature. The presentation largely succeeded in conveying the key contributions of the paper to an audience with varying amounts of prior NLP knowledge, as well as offering an insight into the impact of the paper on current research. Nevertheless, areas of improvement were identified by the audience that could be improved upon in future presentations.

4 Acknowledgements

The student was supported by UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1].

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-December:5999–6009, 2017. ISSN 10495258. URL https://arxiv.org/abs/1706.03762v5.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? *China National Conference on Chinese Computational Linguistics*, pages 194–206, 2019. URL http://arxiv.org/abs/1905.05583.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 5 2020. ISSN 10495258. doi: 10.48550/arxiv.2005.14165. URL https://arxiv.org/abs/2005.14165v4.

Tomáš Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013. URL https://aclanthology.org/N13-1090.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631, 2019. doi: 10.1613/jair.1.11640. URL http://arxiv.org/abs/1706.04902 http://dx.doi.org/10.1613/jair.1.11640.

Milan Gritta and Ignacio Iacobacci. Xeroalign: Zero-shot cross-lingual transformer alignment. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 371–381, 2021. doi: 10.18653/v1/2021.findings-acl.32. URL https://arxiv.org/abs/2105.02472v2.

Milan Gritta, Ruoyu Hu, and Ignacio Iacobacci. Crossaligner & co: Zero-shot transfer methods for task-oriented cross-lingual natural language understanding. *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4048–4061, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.319. URL https://aclanthology.org/2022.findings-acl.319.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *37th International Conference on Machine Learning, ICML 2020*, PartF168147-6:4361–4371, 2020. URL https://arxiv.org/abs/2003.11080v5.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Proceedings of the 1st Workshop*, pages 353–355, 4 2018. doi: 10.48550/arxiv.1804.07461. URL https://arxiv.org/abs/1804.07461v3.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 2383–2392, 6 2016. doi: 10.48550/arxiv.1606.05250. URL https://arxiv.org/abs/1606.05250v3.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 93–104, 2018. doi: 10.18653/V1/D18-1009. URL https://aclanthology.org/D18-1009.

Alec Radford, Karthik Narasimhan, Tim Slimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018. URL https://www.cs.ubc.ca/ amuham01/LING530/papers/radford2018improving.pdf.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32, 5 2019. ISSN 10495258. doi: 10.48550/arxiv.1905.00537. URL https://arxiv.org/abs/1905.00537v3.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, and Paul G Allen. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. ISSN 2331-8422. URL https://arxiv.org/abs/1907.11692v1.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Microsoft Dynamics. Deberta: Decoding-enhanced bert with disentangled attention. *International Conference on Learning Representations ICLR*, 6 2020. doi: 10.48550/arxiv.2006.03654. URL https://arxiv.org/abs/2006.03654v6.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. Challenges in automated debiasing for toxic language detection. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3143–3155, 2021. doi: 10.18653/v1/2021.eacl-main.274. URL https://arxiv.org/abs/2102.00086v1.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 1:1878–1898, 10 2022. doi: 10.48550/arxiv.2110.08527. URL https://arxiv.org/abs/2110.08527v3.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*, 2021. URL https://arxiv.org/abs/2110.15621v1.

Lisa Alazraki, Ali Ghachem, Neophytos Polydorou, Foaad Khosmood, and Abbas Edalat. An empathetic ai coach for self-attachment therapy. *Proceedings - 2021 IEEE 3rd International Conference on Cognitive Machine Intelligence, CogMI 2021*, pages 78–87, 2021. doi: 10.1109/COGMI52975.2021.00019. URL https://doi.org/10.1109/COGMI52975.2021.00019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1:9, 2019. URL https://github.com/codelucas/newspaper.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186, 2018. URL https://arxiv.org/abs/1810.04805v2.