BERT VISION : *Background Reading*

A summary list of background reading necessary for the BERTVision research project.

Siduo Jiang
siduojiang@berkeley.edu

Cristopher Benge
cris.benge@berkeley.edu

Casey King
caseyking@berkeley.edu

Andrew Fogarty
apfogarty@berkeley.edu

Submitted in partial fulfillment of the requirements
for the final project of MIDS W210 Capstone
to the
Faculty of Graduate Studies
of the School of Information
at the University of California, Berkeley

Project GitHub Repository: BERTVision

© Siduo Jiang, Cristopher Benge, William Casey King, Andrew Fogarty : January, 2021

Publication submitted to the graduate studies faculty of the University of California, Berkeley for evaluation and consideration: *Albert Todeschini PhD, Puya H. Vahabi PhD*

This document was typeset using LaTeX, with a mixture of `classicthesis` developed by André Miede. The bibliography was processed by Biblatex. Robert Slimbach's Minion Pro acts as both the text and display typeface. Sans-serif text is typeset in Slimbach and Carol Twombly's Myriad Pro; monospaced text uses Jim Lyle's `Bitstream Vera Mono`.

© BERT VISION : Improving span annotation and classification task performance using parameter-efficient model architectures trained on BERT's hidden state activations.

All rights reserved. No part of this publication may be reproduced or transmitted, in any form or by any means, without permission from the authors.

PROJECT ABSTRACT

We present BERT Vision, a highly parameter-efficient approach for NLP tasks that significantly reduces the need for extended BERT fine-tuning. Our compression method uses information from the hidden state activations of each BERT transformer layer, which is discarded during typical BERT inference. Our method aims to maintain maximal BERT performance at a fraction of the training time and GPU/TPU expense. Furthermore, we extend the utility of our compressed model architecture by evaluating the performance on transfer learning to a wider range of NLP tasks post-compression.[1] [2]

READING LIST

1. Kim, S., Gholami, A., Yao, Z., Mahoney, M. W. & Keutzer, K. *I-BERT: Integer-only BERT Quantization* 2021. arXiv: 2101.01321 [cs.CL].

2. Rajpurkar, P., Jia, R. & Liang, P. Know What You Don't Know: Unanswerable Questions for SQuAD. *CoRR* **abs/1806.03822.** arXiv: 1806.03822. http://arxiv.org/abs/1806.03822 (2018).

3. Van Aken, B., Winter, B., Löser, A. & Gers, F. A. How Does BERT Answer Questions? *Proceedings of the 28th ACM International Conference on Information and Knowledge Management.* http://dx.doi.org/10.1145/3357384.3358028 (Nov. 2019).

4. Lin, M., Chen, Q. & Yan, S. *Network in network* in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings* (2014).

5. Tenney, I., Das, D. & Pavlick, E. *BERT Rediscovers the Classical NLP Pipeline* in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Florence, Italy, July 2019), 4593–4601. https://www.aclweb.org/anthology/P19-1452.

6. Kim, Y., Jernite, Y., Sontag, D. & Rush, A. M. *Character-Aware neural language models* in *30th AAAI Conference on Artificial Intelligence, AAAI 2016* (2016). ISBN: 9781577357605. arXiv: 1508.06615.

7. Rajpurkar, P., Zhang, J., Lopyrev, K. & Liang, P. *SQuad: 100,000+ questions for machine comprehension of text* in *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings* (2016). ISBN: 9781945626258.

8. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. *BERT: Pre-training of deep bidirectional transformers for language understanding* in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (2019). ISBN: 9781950737130. arXiv: 1810.04805.

9. Vaswani, A. *et al. Attention is all you need* in *Advances in Neural Information Processing Systems* (2017).

[1] *BERTVision* - so named for our method of peering within BERT for the signal hidden therein.

[2] See GitHub repository: BERTVision

10. Conneau, A., Schwenk, H., Cun, Y. L. & Barrault, L. *Very deep convolutional networks for text classification* in *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference* (2017). ISBN: 9781510838604.

11. Kuefler, A. R. *Merging Recurrence and Inception-Like Convolution for Sentiment Analysis* https://cs224d.stanford.edu/reports/akuefler.pdf (2020).

12. Limaye, G., Pandit, M. & Vinay, S. *BertNet: Combining BERT language representation with Attention and CNN for Reading Comprehension* https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/default/15783457.pdf (2020).

13. Takeuchi, D. & Tran, K. *Improving SQUAD 2.0 Performance using BERT + X* https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/default/15737384.pdf (2020).

14. Ramachandran, P. *et al. Stand-Alone Self-Attention in Vision Models* 2019. arXiv: 1906.05909 `[cs.CV]`.

15. Peters, M. E. *et al. Deep contextualized word representations* 2018. arXiv: 1802.05365 `[cs.CL]`.

16. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. *CoRR* **abs/1610.02357.** arXiv: 1610.02357. http://arxiv.org/abs/1610.02357 (2016).

17. Ma, X., Wang, Z., Ng, P., Nallapati, R. & Xiang, B. *Universal Text Representation from BERT: An Empirical Study* 2019. arXiv: 1910.07973 `[cs.CL]`.

18. Houlsby, N. *et al.* Parameter-Efficient Transfer Learning for NLP. *CoRR* **abs/1902.00751.** arXiv: 1902.00751. http://arxiv.org/abs/1902.00751 (2019).

19. Liu, X. *et al.* Stochastic Answer Networks for SQuAD 2.0. *CoRR* **abs/1809.09194.** arXiv: 1809.09194. http://arxiv.org/abs/1809.09194 (2018).

20. Szegedy, C. *et al.* Going Deeper with Convolutions. *CoRR* **abs/1409.4842.** arXiv: 1409.4842. http://arxiv.org/abs/1409.4842 (2014).

21. Tenney, I. *et al.* What do you learn from context? Probing for sentence structure in contextualized word representations. *CoRR* **abs/1905.06316.** arXiv: 1905.06316. http://arxiv.org/abs/1905.06316 (2019).

22. Zhu, J. *et al.* Incorporating BERT into Neural Machine Translation. *ArXiv* **abs/2002.06823** (2020).

23. Chen, Z., Trabelsi, M., Heflin, J., Xu, Y. & Davison, B. D. Table Search Using a Deep Contextualized Language Model. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* http://dx.doi.org/10.1145/3397271.3401044 (July 2020).

24. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter* 2019. arXiv: 1910.01108 `[cs.CL]`.

25. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-j. *BLEU: a Method for Automatic Evaluation of Machine Translation* in (2002), 311–318.

26. Wu, Y. *et al.* Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR* **abs/1609.08144.** arXiv: 1609.08144. http://arxiv.org/abs/1609.08144 (2016).