Survey of BERT (Bidirectional Encoder Representation Transformer) types

Athar Hussein Mohammed, Ali H. Ali

Department of Computer Science, Faculty of Computer Science and Maths, University of Kufa

Department of Electronic and Communications Engineering, Faculty of Engineering, University of Kufa

email: alih.alathari@uokufa.edu.iq

Abstract

There are many algorithms used in Natural Language Processing (NLP) to achieve good results, such as Machine Learning (ML), Deep Learning (DL) and many other algorithms. In Natural Language Processing, the first challenge is to convert text to numbers for using by any algorithm that a researcher choose. So how can convert text to numbers? This is happen by using Word Embedding algorithms such as skip gram, bags of words, BERT and etc. Representing words as numerical vectors by relying on the contents has become one of the effective methods for analyzing texts in machine learning, so that each word is represented by a vector to determine its meaning or to know how close or distant this word from the rest of the other word.

BERT (Bidirectional Encoder Representation Transformer) is one of the embedding methods. It is designed to pre-trained form left and right in all layer deep training. It is a deep language model that is used for various tasks in natural language processing. In this paper we will review the different versions and types of BERT.

Keywords

BERT, transformer, natural language processing (NLP), model, bidirectional encoder.

1- Introduction

Natural Language Processing (NLP) is a sub-science of artificial intelligence. This science enables us to create software that can analyze and simulate an understanding of natural languages [1][2]. Main areas of natural language processing such as: machine reading of texts, speech discrimination automatic text or speech generation, machine translation, understand and answer questions need a model. Instead of building a model from scratch to solve a similar problem, you use the model trained on other problem as a beginning point. There are many types of pre-trained models, some types look to text from left to right such as ELMo [3], GPT [4], while BERT is a bidirectional look to text from the two sides from left and from right sides. BERT is a pre-trained model is used for tasks feature based and fine tuning. BERT depend on transformer, where the transformer consist of two part 1- encoder 2- decoder. BERT take only the part encoder and neglected the decoder. The encoder consist of two parts the first part is self attention, the second part is feed forward neural network. There are two sizes of BERT (BERT base and BERT large). The input to BERT are tokens that must convert to vectors of numbers, this is done by using embedding algorithms. In BERT using word to vector to convert tokens into vector, because the Transformer can not deal with string just with numerical vectors. Before enter texts must tokenize by word pieces method, that deal with words that not found in vocabulary by divide it. BERT using the first token is [cls] and the final token is [sep] to refer to the end of sentence. Each of embedding tokens are add to segment and position embedding. The final result will enter to transformer. The output of BERT is series of vectors. BERT model consist of two steps, pre-trained and fine tuning. The pre-training of...
BERT is done by two unsupervised tasks: masked LM and The next sentence prediction. The point of power of BERT is ability to self-attention from left and right from both sides. This property allows BERT to predict the masked word. Therefore to train BERT model must mask some input tokens randomly and predict these masked tokens, this method is called Masked LM. In this pre-training method using 15% of input word pieces of the token randomly. The result of this operation is a bidirectional pre-trained model. The next sentence prediction, where there are many downstream tasks like Question Answering (QA) and natural language inference (NLI) depend on understanding relation found between two sentences. Therefore to capture a model that understands the relations between sentences must be pre-trained model by using predict the next sentence from the corpus by using 50% sentence A and B where B follow A in corpus, and 50% random sentence from corpus. Fine tuning step is straightforward, where self-attention in Transformer will allow BERT to apply to many tasks either involve single text or pair of text [5]. In this paper, listed types of BERT that either used other type of data or enhanced the method used in BERT.

2- BERT types

After the original BERT, there are many types appear that explain in the following text:

2.1 BioBERT (Biomedical Bidirectional Encoding Representation Transformer)

Appear in 2019, as a result of increasing biomedical documents vastly, this lead to need to bio medical text mining. The need to specific model in biomedical field rather than general model like BERT that trained on general texts such as Wikipedia and general books. Therefore when use bert with biomedical will not give good result because the training data is general not specific. therefore appear the BioBERT. The data used in training process take from PubMed and PMC. Biobert use the bert base size as in left side table (1). The fine tuning step contain the three tasks, Named Entity Recognition, Relation Extraction and Question Answering [7][8][9][10].

2.2 Clinical BERT (Biomedical Bidirectional Encoding Representation Transformer)

Clinical BERT appear as a result of applying bert on clinical text. This is useful in disease prediction, relationship between treatments and outcomes or summarize to huge volume of text. The clinical text came from recording states patents during 30 days in hospital. It used the same architecture of original bert, where also use the Transformer and self-attention mechanism [6]. Clinical texts contain jargons, abbreviations and notes that may contain different syntax and grammar errors, comparing to common language in books that not contain errors. Therefore it hard to understand these notes without professional training. It use the same pre-trained tasks in bert that contain masked token and next sentence prediction [11][12][13][14][15].

2.3 AraBERT (Arabic Biomedical Bidirectional Encoding Representation Transformer)

The morphology of Arabic language is rich but less resources and less explored of syntax compared to English language, by using natural language processing, Question Answering (QA), Named Entity Recognition (NER) and Sentiment Analysis (SA) that are tasks of NLP formed a challenging tackle. These limitation lead to use bert model in Arabic language to encourage researcher to apply NLP to Arabic texts. So using bert with huge corpus and apply pre-trained. The result is very success when compare with the result of original bert. Arabert use the same architecture of bert. The fine tuning step contain Sequence Classification, Named Entity Recognition and Question Answering [16][17][18][19][20].

2.4 SCIBERT (Biomedical Bidirectional Encoding Representation Transformer)

SciBERT is pre-trained model has the same architecture of BERT [5] to handle a large database in science field. It is a new model to enhance the performance of many NLP tasks in scientific field. It trained on large corpus that collect from semantic scholar. when training the scibert use the full text for the collected paper rather than just abstracts. SCIBERT apply on many NLP tasks such as: Relation Classification (REL), Text Classification (CLS), Named Entity Recognition (NER), Dependency Parsing (DEP), and PICO Extraction (PICO). Where PICO is like NER where it is extracting task the spans. REL special case of classification of text, so the model able to predict type of relation between entities [21][22][23][24].
2.5 RoBERTa: A Robustly Biomedical Bidirectional Encoding Representation Transformer Approach

In 2019 Yinhan al. added an enhancement on BERT large. These enhancements include: the training data used longer than original BERT [5], the batches is bigger on most data, also omit the next sentence prediction, the training is done on longer sequences. The masking task changed dynamically when applied on training data. The corpus used is a new dataset that taken from (CC-News) [25].

2.6 DeBERTa: Decoding enhanced Biomedical Bidirectional Encoding Representation Transformer with Disentangled Attention

In 2020 Pengcheng He. suggest a new architecture to enhance BERT and RoBERTa models by using two innovative approaches. The first approach is disentangled attention mechanism, using two vectors to represent every word that encode its content and position. Disentangled matrices used to compute attention weight on their contents and positions. The second approach the output softmax layer were change by an enhanced mask decoder to find the masked tokens in the pre-training step. These two approaches enhanced the efficiency of this model and also enhanced the downstream tasks. When compare DeBERTa model with RoBERTa large, DeBERTa used half of training data produce better result on many types of NLP tasks [26][27][28][29].

2.7 AlphaBERT

In 2020, Yen-Pin Chen et al. Improved BERT summarization model by using the information of hospital systems built on character-level tokens. This model use as token unit the English alphabet (characters level). The alphaBERT used the same architecture of bert where use the transformer just encoder part. The size of corpus used in alphaBERT was smaller than the corpus of bert that used in pre-training step. By using the character level can reduce the model size without loss in performance, where the case of out of corpus not appear because depending on character level [30][31][32][33][34].

2.8 BERTje: A Dutch BERT model

In 2019, Wietse de Vries, et al. successfully concluded BERTje that also consist of pre-trained and fine tuning, but by using Dutch language, and applied successfully on NLP tasks for this language. This model is applied by using the same architecture of BERT [5][35].

2.9 DistILBERT: a Distilled version pf Biomedical Bidirectional Encoding Representation Transformer

In 2020 Victor et al. proposed an approach that are faster, smaller, lighter and cheaper than original bert model. This approach is called DistILBERT, where can obtain good performance on large numbers of tasks. The size of bert was reduced by 40% and become 60% faster when applying DistILBERT model that depend on distilled compression technique. This compressed models are small and can execute on the edge such as mobile devices. By using triple loss, this lead that the transformer become 40% smaller [36].

2.10 ALBERT (A Lite Biomedical Bidirectional Encoding Representation Transformer)

When increase the size of model this lead to increase performance on downstream tasks. But this increasing lead to become harder because the limitations of memory (GPU/TPU) and longer training times. To solve these problems, ALBERT introduce two parameter reduction techniques to reduce memory consumption and increase the speed of training for BERT [5]. ALBERT use two parameter reduction techniques lead to lift the main obstacles in scaling pre-trained models. The first technique is a factorized embedding parameterization. By dividing the big vocabulary embedding matrix into two small matrices, isolated the size of the hidden layers from the size of vocabulary embedding. This isolation makes it easier to grow the hidden size without increasing the size of parameters for the vocabulary embedding. The second technique is sharing the cross-layer parameter. This technique stops the parameter from rising with the depth of the network. Both techniques reduce the total of parameters for BERT without effect on performance, this lead to enhance parameter efficiency. An ALBERT architecture similar to BERT-large has 18x less parameters and can be trained about 1.7x quicker.

To more enhanced the performance of ALBERT use a self-supervised loss for sentence order prediction (SOP) [37].
2.11 BioALBERT: (Biomedical A Lite Biomedical Bidirectional Encoding Representation Transformer)

In 2020 usman etal. Proposed a pre-trained language model that it is simple and effective for biomedical Named Entity Recognition (NER). This model used a lite bert and apply with biomedical text, by using the same techniques used in 2-4-10 part that lead to increase the speed of training and decrease memory conception [38].

2.12 Mobile BERT (Mobile Bidirectional Encoding Representation Transformer)

NLP models have large model sizes and high latency cannot apply in mobile devices that have limited resource. Therefore MobileBERT is compressing and accelerating the original BERT model. Same the original BERT, MobileBERT is task-agnostic, it can use to many NLP tasks by simple fine-tuning. MobileBERT is a thin style of BERT large. In the train step of MobileBERT, must firstly train teacher model. After the training step, conduct knowledge transfer from the teacher model to MobileBERT. The studies display that MobileBERT is 5.5 faster and 4.3 smaller than BERT base [39].

2.13 FlauBERT

FlauBERT is a model that trained on a very large French corpus. The training phase used the CNRS news (French National Centre for Scientific Research) Jean Zay supercomputer. FlauBERT train to use in NLP tasks such as word sense disambiguation, text classification, natural language inference, paraphrasing and parsing. There are different versions of FlauBERT [40].

2.14 SqueezeBERT

The computation by BERT and RoBERTa are very costly, when using BERT-base will take 1.7 seconds to classify a text snippet on a Pixel 3 smartphone. So by using SqueezeBERT, many operation in self attention step replace with grouped convolutions. This model run x faster than bert on pixel3 with high accuracy [41].

2.15 Camem BERT (Biomedical Bidirectional Encoding Representation Transformer)

Most offered models trained by using English corpus. This lead to a limitation of using bert model in other languages. Therefore Camem BERT solve this problem, where training the transformer monolingual that represent based for other languages. CamemBERT enhanced the state of the art in all four downstream tasks [42].

The following table show the previous models:

| Model name      | characteristics                                      |
|-----------------|------------------------------------------------------|
| BIOBERT         | Used biomedical texts                                |
| Clinical BERT   | Used clinical notes                                  |
| AraBERT         | Used Arabic texts                                    |
| SCIBERT         | Used science texts                                   |
| RoBERTa         | Train on common text more than original BERT         |
| DeBERTa         | used half of training data of RoBERTa                |
| AlphaBERT       | The size of corpus used in alphaBERT was smaller than the corpus of bert that used in pre-training step. By using the character level |
| BERTje          | using Dutch language                                |
| DistlBERT       | depend on distilled compression technique             |
| AlBERT          | introduce two parameter reduction techniques to reduce memory consumption and increase the speed of training for BERT. |
| BioALBERT       | used a lite bert and apply with biomedical text       |
| MobileBERT      | compressing and accelerating the original BERT model  |
| FlauBERT        | trained on a very large French corpus                |
| SqueezeBERT     | many operation in self attention step replace with grouped convolutions |
| Camem BERT      | training the transformer monolingual that             |
Table (1) the characteristic of models.

3. Conclusions
This paper offers approximately all types of BERT model, that is used to perform word embedding. This work will easy to the researcher to know the types of bert and the enhancement done in the training time or in memory consumption, also using bert model with other types of data, more specific fields, rather than with general data that used in original bert, where bert consist of two steps: the first step is pre-training where the data enter to the layers of transformer and the result of this step are vectors for words. The second step is fine-tuning. The pre-training step consist of two steps: the masked LM and Next Sentence Prediction (NSP).

4. References
[1] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and cross lingual focused evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 1–14, Vancouver, Canada. Association for Computational Linguistics.
[2] Wei Wang, Ming Yan, and Chen Wu. 2018b. Multi granularity 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.
[3] Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embedding: Architecture and representation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1499–1509.
[4] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.
[5] Devlin J. et al. (2019) “Bert: pre-training of deep bidirectional transformers for language understanding”. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), MN, USA, pp. 4171–4186. Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1423
[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin,” Attention is all you need”. In Advances in Neural Information Processing Systems, pages 6000–6010, 2017.
[7] Jinyu Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. (2019)” BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. Bioinformatics, 36(4):1234–1240.
[8] Alsentzer , E. et.al. “Publicly available clinical bert embeddings”. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, pp. 72–78. Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-1909,2019.
[9] Bhasuran, B. and Natarajan, I. “Automatic extraction of gene-disease associations from literature using joint ensemble learning”, PLoS One, 13, e0200699, 2018.
[10] Bravo, A. et al. “Extraction of relations between genes and diseases from text and large-scale data analysis”, implications for translational research. BMC Bioinformatics, 16, 55, 2015.
[11] Kexin Huang , Jaan Altosaarand rahjesh Ranganath “Clinical BERT: Modeling Clinical Notes and Predicting Hospital Readmission” arXiv:1904.05342v3 [cs.CL] 29 Nov 2020.
[12] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann , and M.B.A. McDermott. “Publicly Available ClinicalBERTEmbeddings”. In:arXiv:1904.03323,2019.(13) S. Basu Roy, A. Teredesai, K. Zolfaghari, R. Liu, D. Hazel, S. Newman, and A. Marinez. “Dynamic Hierarchical Classification for Patient Risk-of-Readmission”. In: Knowledge DiscoveryandDataMining,2015.
[14] W.Boag,D.Doss,T.Naumann,andP.Szolovits.“What’sinNote?UnpackingPredictiveValueinClinical Note Representations ”. In :AMIA Joint Summitton Translational Science,2018.
[15] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information”, In: Transactions of the Association for Computational Linguistics, 2017.

[16] Wissam Antoun, Fady Baly, Hazem Hajj, “ArabBERT: Transformer-based Model for Arabic Language Understanding” arXiv:2003.0104v3 [cs.CL] 28 Jun 2020.

[17] Abdelali, A., Darwish, K., Durrani, N., and Mubarak, H., “Farasa: A fast and furious segmenter for Arabic”. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pages 11–16, 2016.

[18] Abdul-Mageed, M., Alhuzali, H., and Elaraby, M., “You tweet what you speak: A city-level dataset of arabic dialects” In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[19] Abu Farha, I. and Magdy, W., “Mazajak: An online Arabic sentiment analyser”. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 192–198, Florence, Italy, August. Association for Computational Linguistics.

[20] Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., and Le, Q. V., ”Towards a human-like open-domain chatbot”, 2020.

[21] Iz Beltagy Kyle Lo Arman Cohan “SCIBERT: A Pretrained Language Model for Scientific Text” arXiv:1903.10676v3 [cs.CL] 10 Sep 2019.

[22] Emily Alsentzer, John R. Murphy, Willie Boag, WeiHung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott, “Publicly available clinical bert embeddings” In Clinical NLP workshop at NAACL, 2019.

[23] Tim Dettmers”TPUs vs GPUs for Transformers (BERT)”. http://timdettmers.com/2018/10/17/tpus-vs-gpus-for-transformers-bert/. Accessed: 2019-02-22, 2019.

[24] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady,”Structural scaffolds for citation intent classification in scientific publications”. In NAACL-HLT, pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics, 2019.

[25] Yinhan Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach” arXiv:1907.11692v1 [cs.CL] 26 Jul 2019.

[26] Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, “deberta: decoding-enhanced bert with disentangled attention” arXiv:2006.03654v2 [cs.CL] 3 Jan 2021.

[27] Iz Beltagy, Matthew E Peters, and Arman Cohan, “Longformer: The long-document transformer”, arXiv preprint arXiv:2004.05150, 2020.

[28] Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini,”The fifth pascal recognizing text entailment challenge”, In Proc Text Analysis Conference (TAC’09), 2009.

[29] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., “Language models are few-shot learners”, arXiv preprint arXiv:2005.14165, 2020.

[30] Yen-Pin Chen et al., “Modified Bidirectional Encoder Representations From Transformers Extractive Summarization Model for Hospital Information Systems Based on Character-Level Tokens (AlphaBERT): Development and Performance Evaluation” JMIR Med Inform 2020 | vol. 8 | iss. 4 | e17787 | p. 2. http://medinform.jmir.org/2020/4/e17787/.

[31] Gigioli P, Sagar N, Rao A, Voyles J. “Domain-Aware Abstractive Text Summarization for Medical Documents”. Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 3-6, 2018, Madrid, Spain URL: https://ieeexplore.ieee.org/document/8621539 [doi: 10.1109/BIBM.2018.8621539].

[32] Nallapati R, Zhou B, Guilcher C., “Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond”. In: Association for Computational Linguistics.; Association for Computational Linguistics; Presented at: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning; August 2016; Berlin, Germany URL: https://www.aclweb.org/anthology/K16-1028/ [doi: 10.18653/v1/K16-1028], 2016.

[33] See A, Liu P. “Get To The Point: Summarization with Pointer-Generator Networks. In: Association for Computational Linguistics”, Presented at: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Vancouver,
Canada p. 1073-1083 URL: https://www.aclweb.org/anthology/P17-1099 [doi: 10.18653/v1/P17-1099], 2017.

[34] Zhou Q, Yang N, Wei F, Huang S, Zhou M. “Neural Document Summarization by Jointly Learning to Score and Select Sentences”, Association for Computational Linguistics; Presented at: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2018/July; Melbourne, Australia p. 654-663 URL: https://www.aclweb.org/anthology/P18-1061 [doi: 10.18653/v1/p18-1061], 2018.

[35] Wietse de Vries et al. "BERTje: A Dutch BERT Model" , arXiv:1912.09582v1 [cs.CL] 19 Dec 2019.

[36] Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", Hugging Face, EMC²: 5th Edition Co-located with NeurIPS’19, arXiv:1910.01108v4 [cs.CL] 1 Mar 2020.

[37] Zhenzhong Lan et al. "albert: a lite bert for self-supervised learning of language representations" arXiv:1909.11942v6 [cs.CL] 9 Feb 2020, Published as a conference paper at ICLR 2020.

[38] Usman Naseem, Matloob Khushi, Vinay Reddy, Sakthivel Rajendran, Imran Razzak and Jinman Kim "BioALBERT: A Simple and Effective Pre-trained Language Model for Biomedical Named Entity Recognition" arXiv:2009.09223v1 [cs.CL] 19 Sep 2020.

[39] Zhiqing Sun et al. ”MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices” arXiv:2004.02984v2 [cs.CL] 14 Apr 2020.

[40] Hang Le et al. "FlauBERT: Unsupervised Language Model Pre-training for French" arXiv:1912.05372v4 [cs.CL] 12 Mar 2020.

[41] Forrest N. Iandola et al. "SqueezeBERT: What can computer vision teach NLP about efficient neural networks?“ arXiv:2006.11316v1 [cs.CL] 19 Jun 2020.

[42] Louis Martin "CamemBERT: a Tasty French Language Model" Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7203–7219 July 5 - 10, 2020. c 2020 Association for Computational Linguistics.