| **SciDeBERTa: Learning DeBERTa for Science**  **Technology Documents and Fine-Tuning**  **Information Extraction Tasks** | Deep learning-based language models (LMs) have transcended the gold standard (human  baseline) of SQuAD 1.1 and GLUE benchmarks in April and July 2019, respectively. As of 2022, the  top five LMs on the SuperGLUE benchmark leaderboard have exceeded the gold standard. Even people  with good general knowledge will struggle to solve problems in specialized fields such as medicine and  artificial intelligence. Just as humans learn specialized knowledge through bachelor’s, master’s, and doctoral  courses, LMs also require a process to develop the ability to understand domain specific knowledge.  Thus, this study proposes SciDeBERTa and SciDeBERTa(CS) as a pre-trained LM (PLM) specialized  in the science technology domain. We further pretrain the DeBERTa, which was trained with a general  corpus, with the science technology domain corpus. Experiments verified that SciDeBERTa(CS) continually  pre-trained in the computer science domain achieved 3.53% and 2.17% higher accuracies than SciBERT  and S2ORC-SciBERT, respectively, which are science technology domain specialized PLMs, in the task  of recognizing entity names in SciERC dataset. In the JRE task of the SciERC dataset, SciDeBERTa(CS)  demonstrated a 6.7% higher performance than baseline SCIIE. In the Genia dataset, SciDeBERTa achieved  the best performance compared to S2ORC-SciBERT, SciBERT, BERT, DeBERTa and SciDeBERTa(CS).  Furthermore, re-initialization technology and optimizers after Adam were explored during fine-tuning to  verify the language understanding of PLMs. | In this section, we review the previous works on  domain-specific PLMs and efficient fine-tuning techniques  for natural language understanding.  A. RECENT TREND OF TRANSFORMER MODEL  We summarize the previous works on transformer-based [18]  advanced LMs and pre-training corpus for PLMs.  There are two mainstreams of transformer-based LMs:  BERT [13], which consists of encoder blocks, and GPT [19],  which consists of decoder blocks. BERT improves perfor-  mance specialized in NLU such as sentence and word clas-  sifications, whereas GPT improves performance specialized  in natural language generation. BERT uses two methods in  pretraining: masked language modeling (MLM), which pre-  dicts the randomly masked tokens in input sentences and  next sentence prediction (NSP), which matches the order of  sentences.  BERT-based PLMs have been enhanced into various  models with improved pre-training tasks. SpanBERT [20]  improved MLM to predict spans instead of tokens that  are relatively easy to solve. StructBERT [21] and ERNIE  2.0 [22] have changed the pre-training task to predict the  sentence order of several sentences instead of two sen-  tences. RoBERTa [19] separates each document and uses  doc-sentence to sample inputs from only one document to  improve context representation understanding by training  consecutive sentences in the same document.  B. LANGUAGE MODELING ON SCIENTIFIC TECHNOLOGY  DOCUMENTS  General LMs based on transformers solve downstream tasks  using two kinds of datasets in each training process. First,  a large amount of unlabeled text data is pretrained by self-  supervised learning, and the model acquires a universal lan-  guage representation. The trained knowledge is transferred  by fine-tuning a PLM to the target data. The target data is  labeled task data, and its size is relatively small compared to  the pre-training data.  In the pretraining step, a general domain corpus is used  for extracting knowledge that can be generally useful in NLP  tasks. BERT has been pretrained with 13 GB of plain texts in  total consisting of 800 M and 2,500 M words collected from  the BooksCorpus [23] and English Wikipedia, respectively.  XLNet [24] and RoBERTa [19] optimized BERT based on  the observation that BERT was underfitted. They trained a  model much longer with a larger batch size on more data.  The PLM shows an improved task performance when the  gap is small between the corpuses used in pretraining and  fine-tuning. The BooksCorpus and English Wikipedia data  used in pretraining the BERT have few noises (for example,  few spelling mistakes) and use formal writing style. There-  fore, the PLM trained on these data shows good performance  in most NLP task benchmarks and leaderboards that have  similar characteristics. However, these models find it difficult  to achieve good performance in social media conversations,  product reviews, and community posts, which have many  noises and are informal. This is particularly true if the target  domain include technical terms that do not belong to the  general language domain, such as financial, legal, biomedi-  cal, and scientific texts. Thus, TweetBERT, FinBERT, Legal-  BERT, BioBERT, PubMedBERT, and SciBERT have been  researched as specialized LMs that pretrain the BERT with a specific domain corpus instead of a general domain cor-  pus [11], [25]–[29].  The pretraining data of SciBERT, the representative PLM  in the science and technology field is composed of 82%  biomedical domain and 18% computer science domain with  3.2B tokens. The S2ORC [12] dataset, which was released  later, collected data in a more balanced manner in more  diverse fields of science and technology. Among the models  specializing in the science domains, S2ORC-SciBERT, which  has been pretrained with 16.4 B tokens, shows better perfor-  mance in processing tasks in the science and technology field  than SciBERT.  BioBERT [28] and PubMedBERT [29], which are state-  of-the-art models for biomedical NLP tasks, were both pre-  trained with biomedical domain text data collected from  PubMed and PubMed Central (PMC). PubMed and PMC  are databases developed and maintained by the National  Library of Medicine. PubMed provides citations of biomed-  ical journals together with abstracts, and PMC archives full-  text articles. BioBERT uses a continual pretraining method  that additionally pretrains the general BERT with biomedi-  cal articles from the general standpoint that the knowledge  provided by a general domain LM would still be useful in  biomedicine. In practice, continual pretraining works effec-  tively when there is a small amount of domain-specific data  for pretraining. PubMedBERT considers that PuMed and  PMC provide biomedical unlabeled text data sufficient to  pretrain a general LM: 33 million abstracts in PubMed and  7.6 million articles in PMC. Hence, PubMedBERT randomly  initializes all parameters of BERT and performs pretraining  completely with biomedical in-domain texts only. PubMed-  BERT pretrained from scratch showed a better performance  in some biomedical NLP tasks than BioBERT. |
| --- | --- | --- |

[11] I. Beltagy, K. Lo, and A. Cohan, ‘‘SciBERT: A pretrained language model

for scientific text,’’ in Proc. Conf. Empirical Methods Natural Lang. Pro-

cess., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), 2019,

pp. 3615–3620.

[12] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. Weld, ‘‘S2ORC:

The semantic scholar open research corpus,’’ in Proc. 58th Annu. Meeting

Assoc. Comput. Linguistics, 2020, pp. 4969–4983.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘‘BERT: Pre-training of

deep bidirectional transformers for language understanding,’’ in NAACL-

HLT, 2019, pp. 4171–4186

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,

L. Kaiser, and I. Polosukhin, ‘‘Attention is all you need,’’ in Proc. Adv.

Neural Inf. Process. Syst., 2017, pp. 5998–6008.

[19] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. (2018). Improv-

ing Language Understanding by Generative Pre-Training. [Online].

Available: https://s3-us-west-2.amazonaws.com/openaiassets/research-

covers/languageunsupervised/languageunderstandingpaper.pdf

[20] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, ‘‘Span-

BERT: Improving pre-training by representing and predicting spans,’’

Trans. Assoc. Comput. Linguistics, vol. 8, pp. 64–77, Jan. 2020.

[21] W. Wang, B. Bi, M. Yan, C. Wu, J. Xia, Z. Bao, L. Peng, and L.

Si, ‘‘StructBERT: Incorporating language structures into pre-training for

deep language understanding,’’ in Proc. 7th Int. Conf. Lear. Represen.

(ICLR), 2019.

[22] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, ‘‘ERNIE

2.0: A continual pre-training framework for language understanding,’’ in

Proc. AAAI Conf. Artif. Intell., 2020, vol. 34, no. 5, pp. 8968–8975.

[23] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba,

and S. Fidler, ‘‘Aligning books and movies: Towards story-like visual

explanations by watching movies and reading books,’’ in Proc. IEEE Int.

Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 19–27.

[24] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V.

Le, ‘‘XLNet: Generalized autoregressive pretraining for language under-

standing,’’ in Proc. 33rd Int. Conf. Neural Inf. Process. Syst., 2019,

pp. 5753–5763.

[25] M. M. A. Qudar and V. Mago, ‘‘TweetBERT: A pretrained language

representation model for Twitter text analysis,’’ 2020, arXiv:2010.11091.

[26] D. Araci, ‘‘FinBERT: Financial sentiment analysis with pre-trained lan-

guage models,’’ 2019, arXiv:1908.10063.

[27] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and

I. Androutsopoulos, ‘‘LEGAL-BERT: The muppets straight out of

law school,’’ in Proc. Findings Assoc. Comput. Linguistics, 2020,

pp. 2898–2904.

[28] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang,

‘‘BioBERT: A pre-trained biomedical language representation model for

biomedical text mining,’’ Bioinformatics, vol. 36, no. 4, p. 1234, Sep. 2019.

[29] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann,

J. Gao, and H. Poon, ‘‘Domain-specific language model pretraining for

biomedical natural language processing,’’ ACM Trans. Comput. Health-

care, vol. 3, no. 1, pp. 1–23, Jan. 2022.