BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA

Sultan Alrowili
University of Delaware
Newark, Delaware, USA
alrowili@udel.edu

K. Vijay-Shanker
University of Delaware
Newark, Delaware, USA
vijay@udel.edu

Abstract

The impact of design choices on the performance of biomedical language models recently has been a subject for investigation. In this paper, we empirically study biomedical domain adaptation with large transformer models using different design choices. We evaluate the performance of our pretrained models against other existing biomedical language models in the literature. Our results show that we achieve state-of-the-art results on several biomedical domain tasks despite using similar or less computational cost compared to other models in the literature. Our findings highlight the significant effect of design choices on improving the performance of biomedical language models.

1 Introduction

The amount of biomedical literature has grown substantially in recent years. This growth created a demand for powerful biomedical language models. Transformer-based language models, such as BERT (Devlin et al., 2019), have shown effectiveness in capturing the contextual representation of corpora at large volume. To address the lack of biomedical contextual representation, both BioBERT (Lee et al., 2019), and SciBERT (Beltagy et al., 2019) have adapted BERT to the biomedical domain.

Recently, several Transformer-based models have been introduced, including Megatron (Shoeybi et al., 2020), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020) and ELECTRA (Clark et al., 2020). These models show impressive performance gains over BERT in the general domain leading most NLP leader boards. However, these models have been evaluated with environmental design factors varying in several dimensions (e.g., vocabulary and corpora domain, loss function, training steps, batch size, and model's scale). Understanding the contribution of these factors to the performance of the language models is challenging, especially when our goal is to shift the contextual representations to the biomedical domain.

This challenge motivates us to investigate the impact of design choices on the performance of biomedical language models. Moreover, highlighting this impact is critical when evaluating new applications in BioNLP, where each application may evaluate its performance against other models that use different design setups. In this work, we pretrain and evaluate different variants of large biomedical Transformer-based models across different design factors.

Thus, our contributions in this paper includes :

(i) We pretrain four different variations of Transformer-based models including: $\text{ELECTRA}_{\text{Base}}$, $\text{ELECTRA}_{\text{Large}}$, $\text{BERT}_{\text{Large}}$ and $\text{ALBERT}_{\text{xxlarge}}$ on biomedical domain corpora using Tensor Processing Units TPUs.

(ii) We fine-tune and evaluate our pretrained models on several downstream biomedical tasks. We present a comprehensive evaluation that highlights the impact of design choices on the performance of biomedical language models.

(iii) We released our pretrained models along with our Github repository.[1]

2 Related Work

2.1 Transformer-based Language Models

The introduction of the BERT model (Devlin et al., 2019) has initiated the advancement of Transformer-based models. Consequently, the investigation of the architecture and design choices of BERT introduced new state-of-the-art models. By exploiting the advantage of using the large batch size and increasing the size of the corpus,

[1]Our pre-trained models and our Github repository are accessible at https://github.com/salrowili/BioM-Transformers.

Proceedings of the BioNLP 2021 workshop, pages 221–227

RoBERTa (Liu et al., 2019) has achieved significant performance gains on all downstream tasks.

The loss function and scalability of BERT were also a subject for investigation by ELECTRA (Clark et al., 2020) and ALBERT (Lan et al., 2020). ELECTRA reaches state-of-the-art results by introducing a binary loss function. This loss function uses generative and discriminative models to accelerate the learning curve. Furthermore, the ALBERT model introduces multiple ideas to the BERT model to improve performance and scalability, including parameter-sharing technique, LAMB optimizer, and factorization of embedding layers. Both ELECTRA and ALBERT are now leading most of NLP benchmarks, including SQuAD (Rajpurkar et al., 2016) and GLUE (Wang et al., 2018).

2.2 Biomedical Language Models

In this section, we will briefly summarize the current state-of-the-art biomedical language models. We should also note that there are other insightful models in literature such as ClinicalBERT (Alsentzer et al., 2019), BlueBERT (Peng et al., 2019), BioELECTRA (Ozyurt, 2020) and BioMedBERT (Chakraborty et al., 2020) .

BioBERT (Lee et al., 2019) is a $\text{BERT}_{\text{Base}}$ model that has been pretrained on biomedical corpora, including PubMed and PMC articles for 23 days on eight V100 GPUs. In our evaluation, we use $\text{BioBERT}_{\text{Base}}$v1.1, which extends the pre-training steps of $\text{BioBERT}_{\text{B}}$ to 1M steps and was trained on PubMed abstracts only.

SciBERT (Beltagy et al., 2019) is a $\text{BERT}_{\text{Base}}$ model that has been pretrained on 1.14M biomedical and computer science papers from Semantic Scholar Corpus .

PubMedBERT (Gu et al., 2021) follows a similar approach of BioBERT by pretraining the BERT model on large biomedical corpora, including PubMed abstracts and PMC articles. PubMedBERT, in contrast to BioBERT, is pretrained using a large batch size (8192) and studies various effects on domain adaptation. The paper also introduces the BLURB benchmark, which is a collection of downstream biomedical tasks.

$\text{BioMegaTron}_{\text{345m}}$ (Shin et al., 2020) is a large-scale model (345m parameters) by NVIDIA based on MegaTron architecture. (Shoeybi et al., 2020). BioMegaTron introduces a variety of large biomedical language models examining the choice of corpora and vocabulary domain.

BioRoBERTa (Lewis et al., 2020) extends the state-of-the-art results by testing different design choices. Similar to BioMegaTron's approach, BioRoBERTa models investigate the effect of vocabulary and corpora domain on the performance of biomedical language model.

3 Pretraining our Language Models

We pretrain all our models using the original implementation of BERT, ALBERT, and ELECTRA. We use TensorFlow 1.15 and TPUv3-512 units to pretrain our large models and TPUv3-32 to pretrain our $\text{BioM-ELECTRA}_{\text{B}}$ model.

3.1 BioM-ALBERT

Initially, we pretrain our model $\text{BioM-ALBERT}_{\text{xxlarge}}$ on PubMed abstracts only. $\text{BioM-ALBERT}_{\text{xxlarge}}$ is based on $\text{ALBERT}_{\text{xxlarge}}$ architecture which has larger hidden layer size (4096) than both BERT_{L} and $\text{ELECTRA}_{\text{L}}$ (1024). We build our specific domain vocabulary, which has a size of 30K words, using the sentence piece model (Kudo and Richardson, 2018). We maintain the same hyperparameters that (Lan et al., 2020) use, except that we increase the batch size to 8192, decrease the initializer range to 0.01. We pretrain $\text{BioM-ALBERT}_{\text{xxlarge}}$ with a learning rate of 1.76e-3 for 264K steps.

Table 1 show the details of our pretrained models compared to the existing model in the literature. The goal to pretrain $\text{BioM-ALBERT}_{\text{xxlarge}}$ is to understand the impact of using ALBERT's techniques on domain adaptation. Moreover, we introduce PMC articles at 264k step, to study the influence of adding PMC articles on the language model. $\text{BioM-ALBERT}_{\text{xxlarge}}$ is the first model that we pretrain and fine-tune among our large models.

3.2 BioM-ELECTRA

We build our $\text{BioM-ELECTRA}_{\text{Base}}$ and $\text{BioM-ELECTRA}_{\text{Large}}$ based on ELECTRA architecture (Clark et al., 2020). We pre-train $\text{BioM-ELECTRA}_{\text{L}}$ on PubMed abstracts only using specific domain vocabulary generated by PubMedBERT, which has a size of 28,895 words. Our evaluation of $\text{BioM-ALBERT}_{\text{xxlarge}}$ on downstream tasks, influences our decision to pretrain BioM-ELECTRA on PubMed abstracts only. We use

Model	Steps	Batch	C	Corpus	Vocabulary
$\text{RoBERTa}_{\text{Base}}$	500k	8192	4.00x	Web crawl	50K Web crawl
$\text{ELECTRA}_{\text{Base++}}$	4M	256	1.00x	XLNET Data	30K Wikipedia + Books
$\text{SciBERT}_{\text{Base}}$	-	-	-	Semantic Scholar	30K PMC+CS
$\text{BioBERT}_{\text{Base}}$	1M	256	0.25x	PubMed Abstracts	30K Wikipedia + Books
$\text{PubMedBERT}_{\text{Base}}$	64K	8192	0.50x	PubMed Abstracts	29K PubMed Abstracts
$\text{PubMedBERT}_{\text{Base+}}$	64K	8192	0.50x	PubMed+PMC	30K PubMed+PMC
$\text{BioM-ELECTRA}_{\text{Base}}$	500K	1024	0.50x	PubMed Abstracts	29K PubMedBERT
$\text{ELECTRA}_{\text{Large}}$	1.7M	2048	3.40x	XLNET Data	30K Wikipedia + Books
$\text{ALBERT}_{\text{xxlarge}}$	1.5M	4096	6.00x	Wikipedia + Books	30k Wikpedia + Books
$\text{BioRoBERTa}_{\text{Large}}$	500K	8192	4.00x	PubMed+PMC+M	50K PubMed+PMC+M
$\text{BioM-BERT}_{\text{Large}}$	690K	4096	2.76x	PubMed+PMC	30k Wikipedia + Books
$\text{BioM-ELECTRA}_{\text{Large}}$	434K	4096	1.73x	PubMed Abstracts	29K PubMedBERT
$\text{BioMegaTron}_{\text{345m}}$	800K	512	0.40x	PubMed+PMC-CC	50K PubMed Abstracts
$\text{BioM-ALBERT}_{\text{xxlarge}}$	264K	8192	2.11x	PubMed Abstracts	30k PubMed (ours)

Table 1: Design choices for our pretrained models and state-of-the-art models. The computational ratio (C) represents the ratio between the number of steps multiplied by the batch size where $\text{ELECTRA}_{\text{base++}}$ is the baseline. XLNet (Yang et al., 2020) data set consist of 33B tokens (130GB) of English corpora. We split the table based on the scale and the domain of language models. CC: Commercial use Collection.

similar pre-training hyperparameters setting described by (Clark et al., 2020) except that we use a larger batch size for $\text{BioM-ELECTRA}_{\text{base}}$ (1024) and $\text{BioM-ELECTRA}_{\text{large}}$ (4096). We pretrain our $\text{BioM-ELECTRA}_{\text{base}}$ for 500K steps and $\text{BioM-ELECTRA}_{\text{large}}$ model for 434K steps .

The main objective to pretrain $\text{BioM-ELECTRA}_{\text{Base}}$ is to study the effect of ELECTRA function by comparing its performance with $\text{PubMedBERT}_{\text{Base}}$ and $\text{RoBERTa}_{\text{Base}}$. Furthermore, we build our $\text{BioM-ELECTRA}_{\text{Large}}$ model to study the effect of model scale by comparing it with $\text{BioM-ELECTRA}_{\text{Base}}$ and $\text{PubMedBERT}_{\text{Base}}$ where other factors are similar. We should also note that we choose general domain model $\text{ELECTRA}_{\text{B++}}$ as a baseline model instead of $\text{ELECTRA}_{\text{B}}$ model. The difference between $\text{ELECTRA}_{\text{B}}$ and $\text{ELECTRA}_{\text{B++}}$ is that $\text{ELECTRA}_{\text{B}}$ is pretrained with less steps (1M) and on smaller corpora (Wikipedia+ Books) (Clark et al., 2020).

3.3 BioM-BERT

We pretrain $\text{BioM-BERT}_{\text{Large}}$ model on PubMed abstracts and PMC articles using the same vocabulary of $\text{BioBERT}_{\text{Base}}$. $\text{BioBERT}_{\text{Base}}$ uses a general domain vocabulary pretrained on English Wikipedia and Books Corpus. Our $\text{BioM-BERT}_{\text{Large}}$ model aims to study the effect of using general domain vocabulary and PubMed + PMC corpora on downstream biomedical tasks. We use a batch size of 4096, a learning rate of 2e-4, and we set the pretraining steps to 700K. However, since we use preemptible TPUs, our TPUs preempted at 690K. We use the ELECTRA implementation of BERT to pretrain our $\text{BERT}_{\text{Large}}$ model. This implementation uses a dynamic masking feature without using next-sentence prediction objective.

4 Fine-Tuning

4.1 Downstream Tasks

Our choices of downstream biomedical tasks are similar to (Shin et al., 2020). For Named Entity Recognition (NER) and Relation Extraction (RE), we generate our training, development, and test data using the same script that PubMedBERT uses (Gu et al., 2021).

Named Entity Recognition Our choices for NER tasks including: BC5CDR-Chemical, BC5CDR-Disease (Li et al., 2016) and NCBI-Disease task. (Doğan et al., 2014). These tasks aim to identify chemical and disease entities using IOB tagging format (Ramshaw and Marcus, 1995). For NER tasks, we use entity-Level F1 score, which is a common standard in the literature.

Relation Extraction is a text classification task where we classify each sequence from a list of labels (classes). For RE task, we choose the ChemProt task (Krallinger et al., 2015) , which is a task that classifies chemical-protein interactions. We use micro-level F1 score on the

five most common classes. We reproduce the results of BioRoBERTa$_L$ [2] on ChemProt task since BioRoBERTa uses a different pre-processing script than (Gu et al., 2021).

Question Answering We use the same BioASQ7B-factoid dataset that (Lee et al., 2019) use, which is in the format of SQuADv1.1. We use Mean Reciprocal Rank (MMR) as an evaluation metric for this task. Moreover, as it is a common practice, we fine-tune our models on BioASQ task using a checkpoint fine-tuned on SQuAD2.0 task (Rajpurkar et al., 2016).

4.2 Fine-Tuning Hyperparameters

We conduct a hyperparameters grid search using the development data set on TPUv3-8. We use TensorFlow 1.15 to fine-tune our model for all tasks, except that we use Transformers library (Wolf et al., 2020) to fine-tune our BioM-ALBERT on NER tasks. Since we are fine-tuning different architectures, we extend our grid search range to : learning rate (1e-4, 2e-4, 1e-5 - 7e-5), batch size (24, 32, 48, 64, 128) and (2-5) epochs . We fixed our choices of hyperparameters for each set of tasks, model's scale, and architecture. The details of our fine-tuning hyperparameters can be found in Appendix A.1.

5 Results and Discussion

Table 2 shows our evaluation results. We categorize models into four categories based on the domain and the scale of each model. We show the results of BioM-BERT$_L$ and BioM-ALBERT$_{xxlarge}$ at different steps. We report entity-level F1 for NER tasks, micro-level F1 for ChemProt, F1 for SQuAD2.0, and Mean Reciprocal Rank (MMR) for BioASQ. We add SQuAD results to track the direction of contextual representation between the general and biomedical domain.

5.1 ELECTRA Objective

The effect of the ELECTRA objective can be seen from comparing both PubMedBERT$_B$ and BioM-ELECTRA$_B$, where they both use similar design choices, vocabulary set, and C ratio. Our evaluation shows that the ELECTRA function improves the performance on ChemProt, SQuAD, and BioASQ tasks. On the SQuAD task, our BioM-ELECTRA$_B$ exceeds RoBERTa$_B$ despite using biomedical corpora and less C ratio. On NER tasks, BioM-ELECTRA$_B$ performs better on the NCBI-disease and worse on the BC5-CDR task. In contrast, BioM-ELECTRA$_{large}$ performs better than other large models on the BC5-CDR dataset, which excludes the assumption that ELECTRA function negatively affects BioM-ELECTRA$_B$ performance on BC5-CDR tasks

5.2 Named Entity Recognition

Specific domain vocabulary significantly improves the results on NER tasks. Results of BioM-ELECTRA$_L$ and BioRoBERTa$_L$ show that biomedical corpora choices have a marginal effect on NER tasks. Our results also show that the gap between base-scale and large-scale biomedical models on NER tasks is relatively smaller than RE and QA tasks, especially for NCBI-Disease task.

5.3 Relation Extraction

On ChemProt task, BioM-BERT$_{Large}$ achieve 78.8 F1 score at 100K step with a C ratio of 0.4x matching the performance of BioRoBERTa$_L$ which has a C ratio of 4.0x. At 1.6x C ratio (400K), it exceeds by a significant margin all large-scale biomedical models. BioM-BERT$_L$ is the only large model in Table 2 that has PP design choice, which highlights the critical impact of general domain vocabulary on some RE tasks such as ChemProt.

5.4 Question Answering

Our results highlight that question answering tasks are sensitive to out-of-domain corpora. This sensitivity can be clearly seen when we introduce (PP) design to BioM-ALBERT$_{xxlarge}$. The performance decreases significantly on the BioASQ challenge. In contrast, the performance on the SQuAD dataset increase to 88.0%. This increase is not caused by extending the training steps since SQuAD score remains stable at 215K and 264K steps.

Moreover, we can observe a gap of 3.9% in the SQuAD benchmark between BioM-ELECTRA$_{Large}$ and BioM-ELECTRA$_{Base}$. However, this gap is not reflected in the BioASQ benchmark since it is in the format of SQuADv1.1, highlighting the need to have a biomedical questing answering task in the format of SQuADv2.0.

Furthermore, our evaluation shows that ELECTRA$_{B++}$ model achieve state-of-the-art result on BioASQ for base-scale models. We attribute this performance to the fact that we use

[2] BioRoBERTA released their models at https://github.com/facebookresearch/bio-lm. We use following hyperparameters to reproduce results (lr: 2e-5 , batch size: 16, epochs : 10, seeds: 10, 42, 1234, 12345, 666).

Model	Design		BC5CDR-		NCBI-	Chem-	QA	
	C	Design	Chem.	Dise.	Dise.	Prot	SQuAD	BioASQ
$RoBERTa_B$	4.00x	G	89.4	80.7	86.6	73.0	83.7	-
$ELECTRA_{B++}$	1.00x	G	90.7	83.0	86.3	73.7	86.2	52.5
$SciBERT_B$	-	S V	92.5	84.7	88.3	75.0	-	-
$BioBERT_B$	0.25x	P	92.6	84.7	**89.1**	76.1	-	41.1
$PubMedBERT_B$	0.50x	P V	93.3	85.6	87.9	77.2	79.1	51.6
$PubMedBERT_{B+}$	0.50x	PP V	93.4	85.6	88.3	77.0	80.9	51.9
$BioM\text{-}ELECTRA_B$	0.50x	P V	93.1	85.2	88.4	77.6	84.4	52.3
$ELECTRA_L$	3.40x	G	91.6	84.4	87.6	75.3	90.7	53.0
$ALBERT_{xxlarge}$	6.00x	G	89.7	81.7	85.5	75.8	**90.2**	53.1
$BioRoBERTa_L$	4.00x	PPM V	93.7	85.2	89.0	78.8	-	-
$BioM\text{-}BERT_L$								
100K	0.40x	PP	-	-	87.8	78.8	84.0	-
400K	1.60x	PP	-	-	88.5	79.8	86.5	-
690K	2.76x	PP	92.4	84.5	88.6	**80.0**	87.3	53.4
$BioM\text{-}ELECTRA_L$	1.73x	P V	**93.8**	85.9	89.0	78.6	88.3	54.1
$BioMegaTron_{345m}$	0.40x	PP V	92.5	**88.5**	87.0	77.0	84.2	52.5
$BioM\text{-}ALBERT_{xxlarge}$								
215K	1.70x	P V	-	-	-	79.0	87.0	55.1
264K	2.11x	P V	93.5	85.2	88.7	79.3	87.0	**56.9**
+64K	2.60x	PP V	-	-	-	79.2	88.0	54.5

Table 2: Evaluation results of our pretrained models. For NER and ChemProt, we use reported results of $SciBERT_B$, $RoBERTa_B$, $BioBERT_B$, $PubMedBERT_B$, $PubMedBERT_{B++}$ (Gu et al., 2021), BioMegaTron (Shin et al., 2020), $BioRoBERTa_L$ (Lewis et al., 2020). We generate QA results for all models, except that we use reported results for BioMegaTron, BioBERT (Shin et al., 2020), $RoBERTa_B$ (Dai et al., 2020). BioMegaTron uses sub-tokens evaluation for NER tasks rather than whole-entity evaluation and uses different pre-processed data set for ChemProt task. Our results are the average scores of five different runs. **B**: Base, **L**: Large, **P**: PubMed, **PP**: PubMed+PMC, **PPM**: PubMed+PMC+MMIC, **V**: Specific domain vocabulary, **S**: Semantic Scholar, **G**: General domain model.

a SQuAD fine-tuned checkpoint to fine-tune our models on BioASQ task. In contrast, the gap between the general and biomedical domain is worse on NER and RE tasks since we are not using any general domain fine-tune checkpoints.

5.5 Fine-Tuning Time

Table 3 shows the fine-tuning efficiency. All base-scale models in Table 2 have similar fine-tuning time to $BioM\text{-}ELECTRA_B$ since they are built on $BERT_B$ architecture. Also all models that are based on $BERT_L$, such as $BioRoBERTa_L$ have similar fine-tuning time to $BioM\text{-}ELECTRA_L$. Our evaluation shows that hidden layer size (H) significantly influences the fine-tuning time.

Model	H	Time	Ratio
$BioM\text{-}ELECTRA_B$	768	03:01	0.35x
$BioM\text{-}ELECTRA_L$	1024	08:27	1.00x
$BioM\text{-}ALBERT_{xxlarge}$	4096	31:15	3.67x

Table 3: Fine-Tuning time of our pre-trained models. We fine-tune all models on ChemProt data set for 3 epochs with a batch size of 32 and max seq. length of 128 on 3090RTX GPU with PyTorch (FP16).

6 Conclusion

We introduce four biomedical Transformer-based language models. Our results show that language models with general domain vocabulary and PubMed+PMC corpora perform better on the ChemProt task. Language models with specific domain vocabulary and PubMed abstracts perform better on NER and QA tasks. In the future, we are planning to extend our evaluation to additional biomedical tasks and investigate implementing early existing (Zhou et al., 2020) to reduce the fine-tuning time. Also, we are planning to build an End-to-End ensemble QA system with our large models and Sentence-BERT (Reimers and Gurevych, 2019) to address pandemic issues such as COVID-19.

Acknowledgment

We would like to acknowledge the support we have from Tensorflow Research Cloud (TFRC) team to grant us access to TPUv3 units. The authors also would like to thank Professor Dr. Li Liao for his insightful discussion and comments on BioM-ALBERT model, anonymous reviewers from the BioNLP2021 workshop for their constructive feedback on our initial manuscript, Hoo Chang Shin for clarifying the experimental design of BioMegaTron.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. 2020. BioMedBERT: A pre-trained biomedical language model for QA and IR. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 669–679, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

Zihang Dai, Guokun Lai, Yiming Yang, and Quoc V. Le. 2020. Funnel-transformer: Filtering out sequential redundancy for efficient language processing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10. 24393765[pmid].

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M. Lowe, Roger A. Sayle, Riza Theresa Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, Tsendsuren Munkhdalai, Keun Ho Ryu, S. V. Ramanan, Senthil Nathan, Slavko Žitnik, Marko Bajec, Lutz Weber, Matthias Irmer, Saber A. Akhondi, Jan A. Kors, Shuo Xu, Xin An, Utpal Kumar Sikdar, Asif Ekbal, Masaharu Yoshioka, Thaer M. Dieb, Miji Choi, Karin Verspoor, Madian Khabsa, C. Lee Giles, Hongfang Liu, Komandur Elayavilli Ravikumar, Andre Lamurias, Francisco M. Couto, Hong-Jie Dai, Richard Tzong-Han Tsai, Caglar Ata, Tolga Can, Anabel Usié, Rui Alves, Isabel Segura-Bedmar, Paloma Martínez, Julen Oyarzabal, and Alfonso Valencia. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7(1):S2.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Jinhyuk 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.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016. Baw068.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ibrahim Burak Ozyurt. 2020. On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 104–112, Online. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

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, Austin, Texas. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. BioMegatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706, Online. Association for Computational Linguistics.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-lm: Training multi-billion parameter language models using model parallelism.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. 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, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.

Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. 2020. Bert loses patience: Fast and robust inference with early exit.

A Appendix

A.1 Fine-Tuning Hyperparameters

Task	Model	E	LR	B
NER	$\text{ELECTRA}_{\text{B}}$	5	2e-4	48
NER	$\text{BioM-ELECTRA}_{\text{B}}$	5	2e-4	48
NER	$\text{BioM-ELECTRA}_{\text{L}}$	5	7e-5	32
NER	$\text{ELECTRA}_{\text{L}}$	5	7e-5	32
NER	$\text{BioM-BERT}_{\text{L}}$	5	7e-5	32
NER	$\text{BioM-ALBERT}_{\text{xxl}}$	4	3e-5	16
NER	$\text{ALBERT}_{\text{xxl}}$	4	3e-5	16
RE	$\text{ELECTRA}_{\text{B}}$	4	1e-4	32
RE	$\text{BioM-ELECTRA}_{\text{B}}$	4	1e-4	32
RE	$\text{BioM-ELECTRA}_{\text{L}}$	4	7e-5	32
RE	$\text{ELECTRA}_{\text{L}}$	4	7e-5	32
RE	$\text{BioM-BERT}_{\text{L}}$	4	7e-5	32
RE	$\text{BioM-ALBERT}_{\text{xxl}}$	5	3e-5	128
RE	$\text{ALBERT}_{\text{xxl}}$	5	3e-5	128
SQ.	PubMedBERT	2	5e-5	32
SQ.	$\text{BioM-ELECTRA}_{\text{B}}$	3	1e-4	32
SQ.	$\text{BioM-ELECTRA}_{\text{L}}$	3	5e-5	32
SQ.	$\text{BioM-BERT}_{\text{L}}$	5	5e-5	48
SQ.	$\text{BioM-ALBERT}_{\text{xxl}}$	2	3e-5	128
Bio.	$\text{BioM-ELECTRA}_{\text{B}}$	4	2e-5	24
Bio.	$\text{ELECTRA}_{\text{B}}$	4	2e-5	24
Bio.	$\text{BioM-ELECTRA}_{\text{L}}$	4	2e-5	24
Bio.	$\text{ELECTRA}_{\text{L}}$	4	2e-5	24
Bio.	PubMedBERT	3	1e-5	128
Bio.	$\text{BioM-ALBERT}_{\text{xxl}}$	3	1e-5	128
Bio.	$\text{ALBERT}_{\text{xxl}}$	3	1e-5	128

Table 4: Fine-Tuning hyperparameters of our pretrained models and base-line general models. We fine-tune all listed models with TensorFlow 1.15 on TPUv3-8 unit. (SQ.: SQuAD2.0, Bio.: BioASQ7B-Factoid, E: Epochs, LR: learning rate, B: Batch size).