BioELECTRA:Pretrained Biomedical text Encoder using Discriminators

Kamal Raj Kanakarajan and **Bhuvana Kundumani** and **Malaikannan Sankarasubbu**
SAAMA AI Research Lab, Chennai, India
{kamal.raj, bhuvana.kundumani, malaikannan.sankarasubbu}@saama.com

Abstract

Recent advancements in pretraining strategies in NLP have shown a significant improvement in the performance of models on various text mining tasks. In this paper, we introduce BioELECTRA, a biomedical domain-specific language encoder model that adapts ELECTRA (Clark et al., 2020) for the Biomedical domain. BioELECTRA outperforms the previous models and achieves state of the art (SOTA) on all the 13 datasets in BLURB benchmark and on all the 4 Clinical datasets from BLUE Benchmark across 7 NLP tasks. BioELECTRA pretrained on PubMed and PMC full text articles performs very well on Clinical datasets as well. BioELECTRA achieves new SOTA 86.34%(1.39% accuracy improvement) on MedNLI and 64% (2.98% accuracy improvement) on PubMedQA dataset.

1 Introduction

Following the success of BERT (Devlin et al., 2018) (Bidirectional Encoder Representations from Transformers) in the general domain, the pretrain-and-finetune approach has been used in the Biomedical domain. With large scale free text available from PubMed and PubMed central (millions of articles), biomedical domain has large unlabelled domain-specific corpus. However, the biomedical domain has labelled datasets that are very small compared to the general domain. Thus the transfer learning approach is well suited for Biomedical domain.

In the biomedical domain, BioBERT (Lee et al., 2020), BlueBERT (Peng et al., 2019) and ClinicalBERT (Alsentzer et al., 2019) are the initial models based on BERT. These models follow continual pretraining approach where the model weights are initialised with weights from BERT trained on Wikipedia and Book Corpus and uses the same vocabulary. Recent models SciBERT (Beltagy et al., 2019), PubMedBERT (Gu et al., 2020) and Biolm (Lewis et al., 2020) have shown that pretraining from scratch using domain specific corpora along with domain specific vocabulary improves the model performance significantly.

In this work, we adapt ELECTRA (Clark et al., 2020), a recent and powerful general domain model for the biomedical domain and we release BioELECTRA - a biomedical domain specific language encoder model. We follow the domain specific pretraining approach where the ELECTRA model is pretrained on PubMed and PubMed Central (PMC) full text articles. ELECTRA outperforms BERT, ALBERT (Lan et al., 2019), XLNet (Yang et al., 2020) and RoBERTa (Liu et al., 2019) on the GLUE (Wang et al., 2019) Benchmark and SQuAD (Rajpurkar et al., 2016a).

In particular, we make the following contributions.

1. We release BioELECTRA(P), BioELECTRA(P + F), BioELECTRA(P + F) LT(Longer Training of additional 1M steps) and BioELECTRA(W + P) pretrained from scratch using Biomedical domain text. Pretrained weights for all these models are publicly released through huggingface transformers(Wolf et al., 2020) model hub.

2. We evaluate our BioELECTRA models on all the 13 datasets in the BLURB (Gu et al., 2020) benchmark and on all the 4 clinical datasets from BLUE (Peng et al., 2019) benchmark across 7 NLP tasks.

3. BioELECTRA model achieves state-of-the-art (SOTA) results on all the 13 datasets in BLURB benchmark and achieves SOTA on all the Clinical datasets from BLUE Benchmark.

4. We publicly release the code[1] and parameters to reproduce our research results.

[1]The code and models are available at https://github.com/kamalkraj/BioELECTRA

Proceedings of the BioNLP 2021 workshop, pages 143–154

2 Related work

Pretrained word embeddings (Mikolov et al., 2013), (Pennington et al., 2014) and contextualised word embeddings (Peters et al., 2018) have helped the deep learning algorithms to improve their performance in NLP tasks. ULMFiT (Howard and Ruder, 2018), introduces the transfer learning approach to Natural language processing and OpenAI GPT (Radford et al., 2018), pretrains a transformer (Vaswani et al., 2017) for learning general language representations. Similar to ULMFiT and OpenAI GPT, BERT (Devlin et al., 2018) follows this fine tuning approach and introduces a powerful bidirectional language representation model using the transformer based model architecture. BERT achieves SOTA on most NLP tasks without any heavily-engineered task specific architectures. Following the success of BERT, XLNet (Yang et al., 2020) with generalized autoregressive pretraining and RoBERTa (Liu et al., 2019) with robust pretraining techniques experiment with different pretraining objectives. ALBERT (Lan et al., 2019) uses weight sharing and embedding factorisation to reduce memory consumption and increase the training speed. ELECTRA (Clark et al., 2020) introduces sample-efficient 'replaced token detection' pretraining technique. ELECTRA-small, trained with very little compute outperforms GPT and performs comparably with larger models like RoBERTa and XLNet.

Recent works adapt BERT to scientific, biomedical and clinical domains. BioBERT (Lee et al., 2020) pretrains BERT with data from PubMed and PubMed Central (PMC) articles. BlueBERT (Peng et al., 2019) pretrains BERT on PubMed, PMC and MIMIC III (Johnson et al., 2016) data. ClinicalBERT (Alsentzer et al., 2019) initialises with BioBERT weights and pretrains on data from MIMIC III. SciBERT (Beltagy et al., 2019), PubMedBERT (Gu et al., 2020) and Bio-lm (Lewis et al., 2020) pretrain BERT based models from scratch with domain specific data. SciBERT pretrains on 1.14M papers from Semantic Scholar (Ammar et al., 2018), PubMedBERT on PubMed and PMC data and Bio-lm (Lewis et al., 2020) on data from PubMed, PMC and MIMIC III. Benchmarks in biomedical NLP - BLUE (Biomedical Language Understanding Evaluation) and BLURB (Biomedical Language Understanding & Reasoning Benchmark) are released by BlueBERT and PubMedBERT respectively.

3 Methods

3.1 Pretraining from scratch using domain specific corpora

The pioneers in applying transfer learning to NLP, pretrain Language Model(LM) on unlabelled large corpora in the general domain like Wikipedia articles, Web Text, Books corpus, Gigaword, web crawl etc. Biomedical literature has specific concepts and terms that are not part of the general domain. To enable the models to learn these features very specific to the biomedical domain, BioNLP models, BioBERT (Lee et al., 2020) and BlueBERT (Peng et al., 2019) use the mixed-domain pretraining approach (Gu et al., 2020). In mixed-domain approach, the model initialises with BERT weights and vocabulary trained on general domain text and the model is pretrained on the biomedical text.

Biomedical domain with its publicly available literature which is growing exponentially by the year makes it well suited for domain specific pretraining from scratch. Using a general domain vocabulary for biomedical text results in complex and specific terms being split into numerous subwords, as they do not exist in the general domain vocabulary. Hence a model trained on these word pieces might not generalise well for the domain specific downstream tasks. Recent work PubMedBERT (Gu et al., 2020) and Bio-lm (Lewis et al., 2020) pretrain a language model from scratch on PubMed abstracts and use the vocabulary that is generated from PubMed abstracts. These models outperform the BioBERT and BlueBERT models on biomedical and clinical NLP tasks.

3.2 Data

We use data very similar to PubMedBERT for fair comparison.

PubMed Abstracts We use text from 22 million PubMed abstracts downloaded as of January 2021. 27 GB of cleaned text with approximately 4.2 billion words are used.

PubMed Central (PMC) We obtained full text from 3.2 million PubMed Central (PMC) [2] articles as of January 2021. After cleaning the data, we use 57GB of text with approximately 9.6 billion words.

Preprocessing We used pubmed_parser parser[3] for

[2] https://www.ncbi.nlm.nih.gov/pmc/

[3] https://github.com/titipata/pubmed_parser

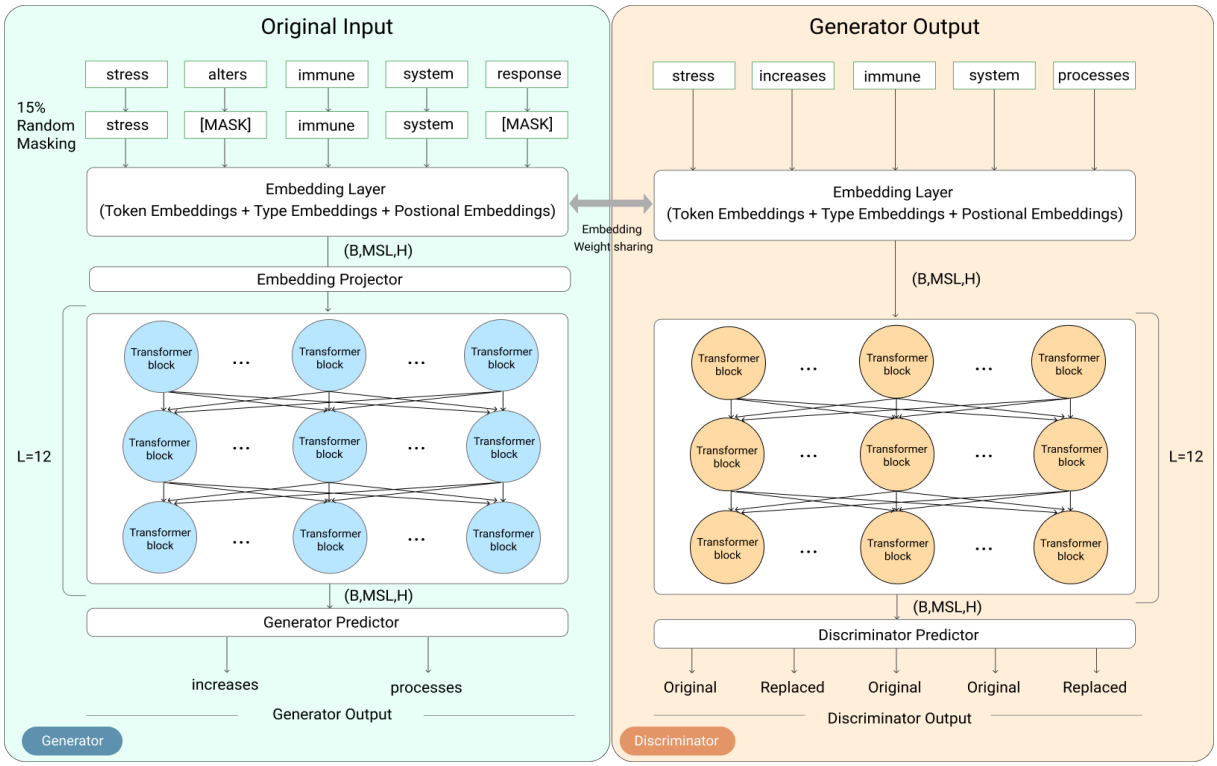

Figure 1: Overview of ELECTRA-Base model Pretraining. Output shapes are mentioned in parenthesis after each block.(B=Batch Size, MSL=Maximum Sequence Length, H=Hidden size)

extracting the abstracts and full text articles. We used SciSpacy(Neumann et al., 2019) for sentence tokenization.

3.3 ELECTRA

Architecture ELECTRA (Clark et al., 2020) pretraining architecture consists of a Generator and a Discriminator network. Each of them consists of Encoder blocks of the transformer (Vaswani et al., 2017) architecture. The generator size is chosen smaller than the Discriminator to make ELECTRA computationally efficient. The size of the Hidden dimension (H) of the transformer encoder in Generator is reduced to $1/3$ the size of the Discriminator. The Generator and Discriminator share the weights of the Embedding layer, which is composed of token embeddings, position embeddings and type embeddings. An embedding projector is added to Generator after the Embedding layer to project the embedding dimension H to H/3. Figure 1 shows pretraining configuration of ELECTRA-Base model. The Generator is trained with maximum likelihood as in ELECTRA paper and Generator is not given a noise input vector as in General Adversarial Networks (GANs). The Discriminator is trained very similar to a classifier with cross entropy loss. After pretraining only the Discriminator is used for all the finetuning.

Input/Output representations ELECTRA follows the Input/Output representations of BERT (Devlin et al., 2018). The first token is always the [CLS] token whose final hidden state is used for finetuning sentence level tasks. For single sentence tasks, the tokenized input sequence should follow the [CLS] token and end with [SEP]. For sentence pair tasks, the tokenized input sentences should be separated by a [SEP] token. Type and Position embeddings which indicate the sentence that it belongs to (sentence1/sentence2) are added to the input token embeddings. Final input representation of a given token is the summation of its token, position and type embeddings which are learnt during the training.

Pretraining Task ELECTRA introduces ***replaced token prediction*** pretraining task where the model is trained to distinguish real input tokens from synthetically generated tokens. Random words are selected in the input text and replaced with tokens generated by a small Generator network. The Discriminator network then predicts whether the input token is original or replaced. This novel approach ensures that the model learns from all the input tokens and not just from 15% of the

Dataset	Task	Train	Dev	Test	Evaluation Metrics
BC5-chem (Li et al., 2016)	NER	5203	5347	5385	F1 entity-level
BC5-disease (Li et al., 2016)	NER	4182	4244	4424	F1 entity-level
NCBI-disease (Doğan et al., 2014)	NER	5134	787	960	F1 entity-level
BC2GM (Smith et al., 2008)	NER	15197	3061	6325	F1 entity-level
JNLPBA (Collier and Kim, 2004)	NER	46750	4551	8662	F1 entity-level
ShARe/CLEFE* (Suominen et al., 2013)	NER	4628	1075	5195	F1 entity-level
EBM PICO(Nye et al., 2018)	PICO	339167	85321	16364	Macro F1 word-level
ChemProt (Krallinger et al., 2017)	Relation Extraction	18035	11268	15745	Micro F1
DDI (Herrero-Zazo et al., 2013)	Relation Extraction	25296	2496	5716	Micro F1
GAD (Bravo et al., 2015)	Relation Extraction	4261	535	534	Micro F1
i2b2-2010* (Uzuner et al., 2011)	Relation Extraction	3110	11	6293	Micro F1
BIOSSES (Soğancıoğlu et al., 2017)	Sentence Similarity	64	16	20	Pearson
ClinicalSTS** (Wang et al., 2020)	Sentence Similarity	1312	329	412	Pearson
HoC (Baker et al., 2015)	Document Classification	1295	186	371	Micro F1
MedNLI* (Romanov and Shivade, 2018)	Inference	11232	1395	1422	Accuracy
PubMedQA (Jin et al., 2019)	Question Answering	450	50	500	Accuracy
BioASQ (Nentidis et al., 2019)	Question Answering	670	75	140	Accuracy

Table 1: Datasets from BLURB and BLUE benchmark. Number of instances in train, dev, and test set along with the evaluation metrics used for each of the datasets is listed. * Clinical domain dataset from BLUE. ** Instead of MedSTS from BLUE we used ClinicalSTS released by (Wang et al., 2020)

tokens in the input text as in BERT. This makes the pretraining task computationally effective. As recent work (Liu et al., 2019) (Yang et al., 2020) suggests that using 'next sentence prediction' does not show consistent improvement in the scores, ELECTRA does not use any such 'contrastive learning' techniques for pretraining. Since ELECTRA does not have a contrastive learning technique, there is no pooling projection layer in ELECTRA.

4 Experiments

4.1 BioELECTRA pretraining

We pretrain ELECTRA from scratch with PubMed abstracts and PMC full text articles mentioned in Section 3.2. PubMedBERT (Gu et al., 2020) and BioBERT (Lee et al., 2020) pretrained BERT-Base models with biomedical domain specific corpus. In this paper, we experiment only with ELECTRA-Base architecture to ensure a fair comparison with these models. Four ELECTRA-Base models are trained - BioELECTRA (P) on PubMed abstracts, BioELECTRA (P+F) on PubMed abstracts and PMC full text articles, BioELECTRA (P+F) with longer training (2M steps) and BioELECTRA (W+P) on Wikipedia and PubMed abstracts. BioELECTRA(P) and BioELECTRA(P+F) models are trained with 1M steps with a batch size of 512. The number of training steps are chosen to make our work comparable with BioBERT[4] and PubMedBERT.[5] BioELECTRA(P+F) LT is trained like BioELECTRA(P+F) with an additional 1M steps. For BioELECTRA(W+F), a continual training approach is adopted where the model is initialised with ELECTRA-BASE general domain weights. It is pretrained further with PubMed abstracts for 100k, 200k and 400k steps. We publish our results of BioELECTRA(W+F) pretrained with 200k steps as these results were comparable with PubMedBERT BLURB (Gu et al., 2020) score.

SciBERT (Beltagy et al., 2019) shows that models trained on uncased vocabularies perform slightly better than the cased models in biomedical domain even for NER tasks. Hence we use the uncased biomedical domain-specific vocabularies from PubMedBERT for all our experiments. The optimization techniques and parameters from ELECTRA paper are followed. All our models are trained on Tensor Processing Unit(TPU) v3-8 instances. Refer Appendix A for complete model and optimizer details.

4.2 Datasets

We finetune our ELECTRA-Base models on 17 NLP datasets - 13 biomedical datasets from the

[4]BioBERT was trained with a batch size of 256 with 1M steps in pretraining and 1M steps in continual pretraining.

[5]PubMedBERT was trained with a batch size of 8,192 for 62,500 steps.

BLURB (Gu et al., 2020) benchmark and 4 clinical datasets from the BLUE (Peng et al., 2019) benchmark. We group our datasets based on the NLP tasks. We do not discuss the datasets in detail due to space constraints. Details on train, dev, test split, benchmark they belong to, evaluation metric used can be found in Table 1. Detailed description of the datasets are available in the BLURB(Gu et al., 2020) and BLUE(Peng et al., 2019) paper.

4.2.1 Named Entity Recognition (NER)

NER task aims at recognizing and predicting the entities e.g (chemicals, diseases, genes, proteins) in the given text. We use *BC5-Chemical, BC5-Disease, NCBI-Disease, BC2GM, JNLPBA* biomedical datasets from the BLURB benchmark. These datasets have the same train, dev and test split as released by (Crichton et al., 2017). In addition to these, *ShARe/CLEFE* clinical dataset used by BLUE benchmark which uses the train, dev and test split released by (Suominen et al., 2013) is used for NER task.

4.2.2 PICO extraction (PICO)

PICO task is very similar to NER, where the model aims to predict the Participants, Interventions, Comparisons and Outcomes entities in the given text. *EBM PICO* (Nye et al., 2020) dataset from the BLURB benchmark which has the same train, test and dev split as the original dataset is used for this task.

4.2.3 Relation Extraction (RE)

Relation Extraction task predicts relations and their types between the two entities mentioned in the given sentences (e.g, gene–disease relations, protein–chemical relations). We use *DDI, ChemProt and GAD* datasets from the BLURB benchmark and *i2b2-2010* clinical dataset in the BLUE benchmark. GAD dataset in BLURB benchmark uses train, dev and test split created by (Lee et al., 2020). For DDI, BLURB uses the original dataset by (Herrero-Zazo et al., 2013) and release their own train, dev and test datasets. BLURB uses the train, dev and test split from the original dataset (Krallinger et al., 2017) for ChemProt. BLUE uses the train, dev and test split released by (Uzuner et al., 2011)

4.2.4 Sentence Similarity

Sentence Similarity task predicts the similarity score based on how similar are the given pair of sentences. *BIOSSES* dataset from BLURB benchmark and *ClinicalSTS* dataset instead of the MedSTS dataset is chosen from BLUE benchmark. BLURB uses the train, dev and split created by (Peng et al., 2019). *ClinicalSTS* dataset is chosen as that is the latest version provided by n2c2 2019 challenge(Wang et al., 2020). It has added 574 more samples for training and a new test set of 412 samples. As this dataset doesn't have a public train and dev split, we have split it into 80% train and 20% dev set and we use the original test set for evaluation.

4.2.5 Document classification

Document classification task aims to predict the multiple labels for the given text. Evaluation for Document classification task is done at the document level where we aggregate the labels over all the sentences in a document. We use *HoC* dataset from BLURB benchmark which uses the original dataset by (Baker et al., 2015) to create their own train, dev and test split.

4.2.6 Natural Language Inference (NLI)

Natural Language Inference task predicts whether the relation between two sentences are entailment, contradiction or neutrality. *MedNLI* (Romanov and Shivade, 2018) dataset from the BLUE benchmark which uses the original train, dev and test split is used for this task.

4.2.7 Question Answering (QA)

Question Answering task aims to predict the answers in the context when a question text is given as the first sentence. The answers are either two-way (yes/ no) or three-way (yes/ maybe/ no). *PubMedQA and BioASQ* datasets from BLURB benchmark are used for our experiments. For both PubMedQA (Jin et al., 2019) and BioASQ (Nentidis et al., 2019), BLURB uses the original train, dev and test split.

4.3 Fine tuning

ELECTRA (Clark et al., 2020) applies very minimal architectural changes for finetuning downstream tasks. We follow the same approach as ELECTRA for finetuning BioELECTRA on the various downstream tasks. BIO encoding scheme is adopted for the NER tasks where B stands for Beginning, I stands for Inside and O stands for Outside. All the NER datasets in BLURB benchmark and *ShARe/CLEFE* in BLUE benchmark have

	BioBERT cased (P)	SciBERT uncased (CS+F)	ClinicalBERT cased (W+P+M)	BlueBERT cased (W+P+M)	PubMedBERT uncased (P)	BioELECTRA uncased (P)
BC5-chem	92.85	92.49	90.80	91.19	93.33	**93.60**
BC5-disease.	84.70	84.54	83.04	83.69	85.62	**85.84**
NCBI-disease	89.13	88.10	86.32	88.04	87.82	**89.38**
BC2GM	83.82	83.36	81.71	81.87	84.52	**84.69**
JNLPBA	79.35	79.45	78.59	78.68	80.06	**80.17**
EBM PICO	73.18	73.12	72.06	72.54	73.38	**74.26**
ChemProt	76.14	75.24	72.04	71.46	77.24	**78.20**
DDI	80.88	81.06	78.20	77.78	82.36	**82.76**
GAD	80.94	80.90	78.40	77.24	82.34	**83.70**
BIOSSES	89.52	86.25	91.23	85.38	92.30	**92.49**
HoC	81.54	80.66	80.74	80.48	82.32	**83.50**
PubMedQA	60.24	57.38	49.08	48.44	55.84	**64.02**
BioASQ	84.14	78.86	68.50	68.71	87.56	**88.57**
BLURB score	80.29	78.80	77.19	76.19	81.10	**82.47**

Table 2: Comparison of pretrained BioNLP models on the BLURB (Gu et al., 2020) benchmark. The BLURB score is the macro average of mean test results for each of the six tasks (NER, PICO, Relation Extraction, Sentence Similarity, Document Classification, Question Answering). Refer Table 1 for the evaluation metric used for each task. (P - PubMed abstracts, CS - Computer Science, F - PubMed Central full text articles, W - Wikipedia, M - MIMIC III (Johnson et al., 2016))

a single entity. (e.g. Disease in BC5-disease). PICO, a sequential tagging task is solved using the NER task approach and Document classification task for *HoC* dataset is solved as multi label classification task. The datasets in NER, PICO and Document classification tasks follow the single sentence representation. As mentioned in section 3.3, each tokenized input sequence follows the [CLS] token and ends with the [SEP] token. Sentence Similarity, Question Answering and Natural Language Inference tasks all have sentence pairs in their inputs. We process the sentence pairs as [CLS]sentence1[SEP]sentence2[SEP] very similar to BERT. In the Question Answering task, 'question' is treated as sentence1 and 'context' is treated as sentence2.

ELECTRA (Clark et al., 2020) uses the vector representation of the [CLS] token to generate the output for all the given NLP tasks except NER and PICO. For NER and PICO, representations for each token is used to classify the entities. A simple linear layer is added to the output of ELECTRA for finetuning. ELECTRA does not use LSTM (Hochreiter and Schmidhuber, 1997), CRF (Lafferty et al., 2001) layers for NER tasks. Figure 2 in appendix B illustrates the finetuning architecture for the NLP tasks. Mean-square error is used for regression tasks and cross entropy loss is used for classification tasks. Similar to BERT finetuning, all the layers are fine-tuned together along with task specific prediction layer. We use 'discriminative finetuning' similar to ELECTRA, where only the final layer is trained with the original learning rate and all other layers use a learning rate with a decay factor. For finetuning, Adam (Kingma and Ba, 2017) optimizer with a slanted triangular learning rate scheduler which linearly warms up (10% of steps) followed by linear decay (90% of steps) is used. We also use a dropout probability of 10%. We experiment with the following hyper parameters: learning rate [3e-5, 5e-5, 1e-4, 1.5e-4, 2e-4], batch size [16, 32], layer-wise learning-rate decay out of [0.9, 0.8, 0.7] and epochs [3,5]. BIOSSES (Soğancıoğlu et al., 2017), PubMedQA (Jin et al., 2019), BioASQ (Nentidis et al., 2019) and ClinicalSTS (Wang et al., 2020) are finetuned for longer epochs. For more details on the hyper parameters, refer Appendix B. We ran 10 fine tuning runs on BIOSSES, BioASQ and PubMedQA since the datasets are relatively smaller and 5 runs on all the other datasets. The average score is reported as the final score for the evaluation metric.

	BioBERT cased (P)	ClinicalBERT cased (W+P+M)	BlueBERT cased (P)	BlueBERT cased (P+M)	PubMedBERT uncased (P)	PubMedBERT uncased (P+F)	BioELECTRA uncased (P)	BioELECTRA uncased (P+F)
MedNLI	82.63	82.70	82.2	84	83.82	84.17	86.27	**86.34**
i2b2-2010	72.81	74.82	74.4	76.4	75.14	73.93	**76.50**	75.73
ShARe/CLEFE	80.73	82.15	75.4	77.1	74.45	74.77	**83.71**	83.15
ClinicalSTS	85.91	85.63	86.03	84.57	86.72	86.16	**89.07**	88.34

Table 3: Comparison of pretrained language models on the BLUE (Peng et al., 2019) benchmark. (P - PubMed abstracts, F - PubMed Central full text articles, W - Wikipedia, M - MIMIC III (Johnson et al., 2016))

5 Results

We finetune all of the four BioELECTRA models mentioned in 4.1 for seven biomedical text mining tasks (NER, PICO, Relation Extraction, Sentence Similarity, Document Classification, Question Answering and Natural Language Inference) that are part of the BLURB (Gu et al., 2020) and BLUE (Peng et al., 2019) benchmark.

BLURB benchmark Out of the four BioELECTRA models, BioELECTRA (P) model pretrained from scratch on PubMed abstracts alone along with biomedical domain specific vocabulary (from PubMedBERT (Gu et al., 2020)) achieves new State-of-the-Art (SOTA) results on all of the datasets in BLURB benchmark. Our results on BioELECTRA (P) along with the scores for BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), ClinicalBERT (Alsentzer et al., 2019) , BlueBERT (Peng et al., 2019) and PubMedBERT (Gu et al., 2020) for all the tasks in the BLURB benchmark are shown in table 2. The scores on these datasets for all these models are taken from the BLURB benchmark. As we do not have details on train, test and dev split of datasets used by Bio-lm (Lewis et al., 2020) paper, we are not able to compare our results with their results. For NCBI-Disease, where the train, test and dev split is publicly available, our model (89.38%) performs better than the Bio-lm Base (PM + Voc) model (88.2%). ELECTRA performs significantly better than all other BERT based models on the SQuAD (Rajpurkar et al., 2016b) benchmark in the general domain. Similarly, BioELECTRA (P) model has significantly higher scores on the Question Answering tasks. It achieves new SOTA of 64.02% (3.78% increase over the previous SOTA) on PubMedQA and with a new SOTA of 88.57% (1.01 % increase over the previous SOTA) on BioASQ. Our overall BLURB score (macro average of the average metric for each of the six tasks) is 82.40% which is 1.3% higher than PubMedBERT BLURB score of 81.10%.

BLUE benchmark We present results of BioELECTRA (P) pretrained on PubMed abstracts alone and BioELECTRA (P+F) pretrained on both PubMed abstracts and PubMed full text articles on four of the clinical datasets in the BLUE benchmark in table3. We compare the performance of our models with the results of BioBERT, ClinicalBERT, BlueBERT and PubMedBERT. Since the scores on the train, dev and test split of these clinical datasets by BioBERT, ClinicalBERT, BlueBERT and PubMedBERT are not available, we used their pretrained weights on these datasets and documented the results. We do not have the results of SciBERT model as it was trained on mixed domain data. Out of the four datasets in the BLUE benchmark, we have results of Biolm for i2b2-2010 and MedNLI. Since we do not have the train, dev and test split used by Biolm for i2b2-2010, we compare our results only for the MedNLI dataset. Score of our BioELECTRA (P+F) model 86.34% is significantly higher than Biolm Base model (PM + Voc) score of 83.2%. We also note that BioELECTRA performs better than BERT based models trained on MIMIC data. BioELECTRA (P) achieves new SOTA on three of the datasets - i2b2-2010, ShARe/CLEFE and ClinicalSTS. BioELECTRA (P+F)'s score of 86.34% on MedNLI task is marginally (0.07%) higher than the score of BioELECTRA (P)'s score of 86.27% and this is the new SOTA for MedNLI dataset for models trained on PubMed abstracts and PubMed Central full text articles.

Our models pretrained on domain specific text along with domain specific vocabulary have consistently shown that the pretraining from scratch with domain specific data enables the model to capture the contextual representations of the language better.

Vocab	BioELECTRA P PubMed	BioELECTRA P+F PubMed	BioELECTRA P+F (LT) PubMed	BioELECTRA W+P General
BC5-chem	93.60	93.51	**93.75**	93.03
BC5-disease	**85.84**	85.55	85.32	84.66
NCBI-disease	**89.38**	88.43	88.73	88.45
BC2GM	**84.69**	84.61	84.68	83.90
JNLPBA	**80.17**	79.98	80.10	79.63
EBM PICO	**74.26**	73.88	73.86	73.33
ChemProt	**78.20**	77.76	76.76	77.06
DDI	82.76	**83.53**	82.34	79.68
GAD	83.70	84.18	**85.67**	83.16
BIOSSES	92.49	**93.80**	91.45	88.65
HoC	**83.50**	82.79	83.20	82.30
PubMedQA	**64.02**	63.80	62.21	61.20
BioASQ	88.57	91.42	**91.50**	90.01
BLURB Score	82.47	**82.72**	82.24	80.96
MedNLI	86.27	**86.34**	85.36	83.53
i2b2-2010	**76.50**	75.73	76.17	75.48
ShARe/CLEFE	**83.71**	83.15	83.52	83.02
ClinicalSTS	**89.07**	88.34	89.02	88.46

Table 4: Comparison of BioELECTRA models on BLURB (Gu et al., 2020) and BLUE (Peng et al., 2019) benchmark. (P - PubMed abstracts, F - PubMed Central full text articles, W - Wikipedia, LT - Longer Training)

Comparison of BioELECTRA models Table 4 shows the comparison of results of our models BioELECTRA(P), BioELECTRA (P+F) and BioELECTRA (P+F) LT with longer training of additional 1 million steps and BioELECTRA (W+P). BioELECTRA (W+P) is pretrained from scratch on Wikipedia and PubMed abstracts along with a general domain vocabulary (BERT (Devlin et al., 2018) uncased vocabulary). We observe that BioELECTRA (P+F) LT with longer training of 2 million steps does not give substantial improvements on all of the tasks. BioELECTRA (P+F) LT model's result is slightly better than BioELECTRA (P) on BC5-chem dataset. BioELECTRA (P+F) LT model's result on GAD and BioASQ datasets are marginally better than BioELECTRA (P+F). BioELECTRA (P+F) performs slightly better than BioELECTRA (P) on DDI and BIOSSES datasets.

The results clearly show that all BioELECTRA models pretrained from scratch with biomedical domain text and domain specific vocabulary perform better than the model pretrained on both general and biomedical domain text with general domain vocabulary. However it is interesting to note that BioELECTRA (W+P) model has significantly better results for i2b2-2010, ShARe/CLEFE and ClinicalSTS datasets than PubMedBERT. BioELECTRA (W+P)'s score for MedNLI is comparable to that of PubMedBERT (Gu et al., 2020).

6 Conclusion and Future Work

We release BioELECTRA-base models pretrained from scratch on biomedical domain specific text and evaluate the performance on seven different biomedical NLP tasks with 17 datasets. We achieve SOTA on all the datasets in the BLURB (Gu et al., 2020) benchmark and all four clinical datasets in the BLUE (Peng et al., 2019) benchmark. Our results show that pretraining from scratch with biomedical domain text helps the model to learn better contextual representations. We release the pretrained weights for all our models and the code for reproducibility.

We plan to explore and experiment with our domain specific pretraining approach on ELECTRA-LARGE models. We also intend to train ELECTRA-BASE and ELECTRA-LARGE mod-

els on MIMIC III (Johnson et al., 2016) clinical notes and evaluate the performance of the models on biomedical NLP tasks. As ELECTRA shows a significant improvement on SQuAD (Rajpurkar et al., 2016b), we want to focus on Biomedical QA tasks (span prediction) and evaluate domain specific pretrained ELECTRA models performance.

Acknowledgements

This research was supported by Google's TensorFlow Research Cloud (TFRC). We also extend our thanks to Samuel Gurudas for his assistance in creating the diagrams in this research paper.

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.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2015. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440.

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.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics*, 16(1):55.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Nigel Collier and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics*, 18.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, GP Rodríguez, et al. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. 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.

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. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.

Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. 2019. Results of the seventh edition of the bioasq challenge. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 553–568. Springer.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access.

Benjamin Nye, Ani Nenkova, Iain Marshall, and Byron C. Wallace. 2020. Trialstreamer: Mapping and browsing medical evidence in real-time. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 63–69, 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 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016a. 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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016b. Squad: 100,000+ questions for machine comprehension of text.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(S2):S2.

Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. 2017. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231, Berlin, Heidelberg. Springer Berlin Heidelberg.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3266–3280.

Y. Wang, S. Fu, F. Shen, S. Henry, O. Uzuner, and H. Liu. 2020. The 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity: Overview. *JMIR Med Inform*, 8(11):e23375.

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.

A Pretraining

Hyperparameter	Discriminator/Generator
Number of layers	12
Hidden Size	768/256
FFN inner hidden size	3072/1024
Attention heads	12/4
Attention head size	64
Embedding Size	768
Mask percent	15
Learning Rate Decay	Linear
Warmup steps	10000
Learning Rate	2e-4
Adam ϵ	1e-6
Adam β_1	0.9
Adam β_2	0.999
Attention Dropout	0.1
Dropout	0.1
Weight Decay	0.01
Batch Size	512
Train Steps	1M

Table 5: Pre-train hyperparameters.

All the BioELECTRA models are trained on TPU v3-8 instances. Adopting *bfloat16*[6] training helped us in improving the training speed. Very similar to BERT, we train the model in 2 phases, 90% of steps with sequence length of 128 (phase1) and 10% of steps with sequence length of 512 (phase2) to learn the positional embeddings. Model training reached 1M steps in 5 days (phase1 - 4 days and phase2 - 1day). For pretraining, we use the original ELECTRA code[7] released by authors. Refer table 5 for details regarding all the parameters.

B Finetuning

Figure 2 shows different architecture schema of different models.

- Single Sentence Classification : ChemProt, DDI, GAD, i2b2-2010, HoC
- Entity Classification: BC5-chem, BC5-disease, NCBI-Disease, BC2GM, JNLPBA, ShARe/CLEFE, EBM PICO

[6]https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus

[7]https://github.com/google-research/electra

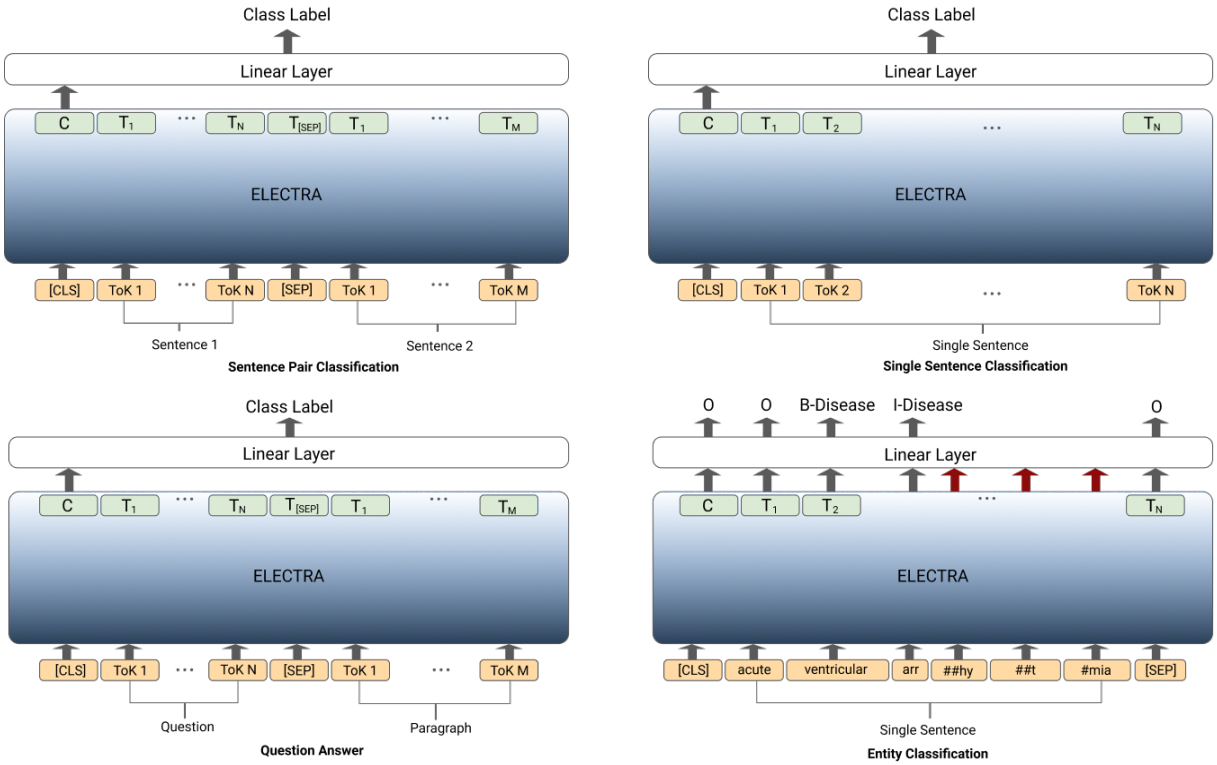

Figure 2: Overview of BioELECTRA model finetuning.

Hyperparameter	Value
Adam ϵ	1e-6
Adam β_1	0.9
Adam β_2	0.999
Layerwise LR decay	0.8
Learning rate decay	Linear
Warmup fraction	0.1
Attention Dropout	0.1
Dropout	0.1
Weight Decay	0

Table 6: Common hyperparamters across tasks

- Sentence Pair Classification: BIOSSES, ClinicalSTS
- Question Answering: PubMedQA, BioASQ

'Discriminative finetuning' is adopted where the learning rate varies across the layers. The learning rate decays across the layers from top to bottom with a factor of 0.8 for all the NLP tasks. The colour gradient in figure 2 represents this . For a learning rate of 1e-4 , only the task specific prediction layer (final layer) is finetuned at this rate. With a decay factor of 0.8, the embedding layer for that particular task is finetuned at a learning rate of 5.5e-6. Table 6 shows the common hyperparameters used across tasks, and table 7 shows task specific hyperparameters.

Dataset	LR	BS	MSL	EPOCHS
BC5-chem	2e-4	16	256	5
BC5-disease	2e-4	16	256	5
NCBI-disease	2e-4	32	128	5
BC2GM	2e-4	32	256	5
JNLPBA	2e-4	16	256	3
ShARe/CLEFE	2e-4	32	512	5
EBM PICO	2e-4	32	256	3
ChemProt	1e-4	32	256	5
DDI	2e-4	32	256	3
GAD	2e-4	32	128	5
i2b2-2010	2e-4	32	128	5
BIOSSES	1.5e-4	16	128	60
ClinicalSTS	5e-5	32	128	10
HoC	2e-4	32	128	5
MedNLI	1e-4	32	128	5
PubMedQA	2e-4	32	512	20
BioASQ	2e-4	32	512	20

Table 7: LR : Learning Rate, BS : Batch Size, MSL : Maximum Sequence Length