IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

MediBioDeBERTa: Biomedical Language Model with Continuous Learning and Intermediate Fine-Tuning

EUNHUI KIM[1](Member, IEEE), YUNA JEONG[2], and MYUNG-SEOK CHOI[3]
[1]Korea Institute of Science and Technology Information, 245 Daehak-ro, 34131, Korea (e-mail: ehkim@kisti.re.kr)
[2]Korea Institute of Science and Technology Information, 245 Daehak-ro, 34131, Korea (e-mail: jeongyuna@kisti.re.kr)
[3]Korea Institute of Science and Technology Information, 245 Daehak-ro, 34131, Korea (e-mail: mschoi@kisti.re.kr)

Corresponding author: Yuna Jeong

This research was supported by the Korea Institute of Science and Technology Information (KISTI) (KISTI: K-23-L04-C05, NTIS:1711198593) and a government-wide R&D fund project for Infectious Disease Research (GFID), Republic of Korea (KISTI: N-23-NT-CU04, HG23C1624) and helped by the National Supercomputing Center with their resources and technical supports

ABSTRACT The emergence of large language models (LLMs) has marked a significant milestone in the evolution of natural language processing. With the expanded use of LLMs in multiple fields, the development of domain-specific pre-trained language models (PLMs) has become a natural progression and requirement. Developing domain-specific PLMs requires careful design, considering not only differences in training methods but also various factors such as the type of training data and hyperparameters. This paper proposes MediBioDeBERTa, a specialized language model (LM) for biomedical applications. First, we present several practical analyses and methods for improving the performance of LMs in specialized domains. As the initial step, we developed SciDeBERTa v2, an LM specialized in the scientific domain. In the SciERC dataset evaluation, SciDeBERTa v2 achieves the state-of-the-art model performance in the named entity recognition (NER) task. We then provide an in-depth analysis of the datasets and training methods used in the biomedical field. Based on these analyses, MediBioDeBERTa, was continually trained on SciDeBERTa v2 to specialize in the biomedical domain. Utilizing the biomedical language understanding and reasoning benchmark (BLURB), we analyzed factors that degrade task performance and proposed additional improvement methods based on intermediate fine-tuning. The results demonstrate improved performance in three categories: named entity recognition (NER), semantic similarity (SS), and question-answering (QnA), as well as in the ChemProt relation extraction (RE) task on BLURB, compared with existing state-of-the-art LMs.

INDEX TERMS Language model, fine-tuning, domain-specific modeling, natural language processing.

I. INTRODUCTION

The emergence of large language models (LLMs) marked a significant milestone in the evolution of natural language processing (NLP) [1], [2], [3]. LLMs have been utilized to build a universal knowledge of a language by pre-training on large amounts of data and then being fine-tuned for specific downstream tasks. Typical transformer encoder-based language models (LMs) have been pre-trained to acquire general knowledge and contextual characteristics from large corpora, such as Wikipedia, news articles, and books [2], [4], [5], [6], [7]. This knowledge transfer to downstream tasks is the cornerstone of the successful application of a pre-trained language model (PLM) to many problems. The advantage of transfer learning has been robustly validated in many NLP applications.

However, it is premised on the assumption that the language features of the downstream task are similar to those used in pretraining. Language features vary in terms of language type, style, terminology, etc.; the larger the gap between the pretraining information and the downstream problem, the more difficult it is to take advantage of them [8], [9], [10], [11], [12]. This culminated in the development of domain-specific PLMs that relied on domain-specific data for pretraining. This trend also manifests in generative LMs employing transformer decoders. OpenAI's GPT model [13], as a representative example, utilizes in-context learning to acquire general task knowledge through prompts. This allows it to demonstrate proficiency across a wide range of tasks without the need for fine-tuning. However, its performance on specialized domains for which it hasn't been trained is comparatively less effective. Even when fine-tuned, it cannot surpass the performance of domain-specific pre-trained models [14].

Regarding the development of PLMs for specific domains, two primary approaches have been identified. One is continuous learning, which involves learning specialized datasets in addition to PLMs that have already learned general knowledge, and the other involves learning models from scratch using only specialized data [8], [9], [10], [11], [12], [15]. In the first method, only a relatively small specialized dataset is utilized for training, considering that general knowledge has already been acquired. This approach is particularly suited to scenarios in which computational resources are limited or specialized datasets are rare. Conversely, the second approach requires training models entirely from scratch, using only domain-specific data. This technique requires a significantly larger dataset and an extended training period, increasing computing resource requirements. However, the benefit of the latter strategy lies in its potential to yield highly domain-specialized models that often surpass the performance of the former, which is refined using a continuous learning approach [9], [11].

As Large Language Models (LLMs) have found increased use across multiple fields, the development of domain-specific Pre-trained Language Models (PLMs) has emerged as both a natural progression and a necessity.

The main contribution point of this paper can be summarized as follows:

- The proposed PLM, SciDeBERTa v2, which is trained from scratch on the S2ORC dataset [9], encompasses full text and outperforms the previous version, SciDeBERTa [12]. SciDeBERTa v2 achieves the state-of-the-art model on the NER task in the SciERC dataset [16].
- This study provides guideline for the development a biomedical domain-specific LM using the BLURB benchmark, which consists of 6 categories and 13 tasks. The process consists of pre-training, domain-specific pre-training, inter-mediate fine-tuning according to the task,and fine-tuning.
- The proposed model, MediBioDeBERTa-IFT, surpassed the existing state-of-the-art models of the same size in three kinds of categories (NER by 0.28%, SS by 0.2%, QnA by 1.29%) and one task (ChemProt RE by 14.96%) in BLRUB.

The remainder of the paper is structured as follows: In Section II we present a brief review of previous studies. Section III outlines the procedure of pretraining to develop an optimized PLM for a specialized domain. Section IV details further improvement techniques. We show experimental results and discussion in Section V, and finally, we conclude in Section VI.

II. RELATED WORK

Transformer-based LMs have recently been used to improve language comprehension and generation. The representative transformer-based LM, BERT [2], outperformed humans in the 2019 GLUE benchmark [17]. RoBERTa [4] optimized BERT by including as many documents as possible in a batch by excluding sentence-relationship matching. DeBERTa [7], an extension of RoBERTa, introduces disentangled attention, enhanced masked decoder, and accounts for the relative positions of tokens, achieving superior results in the SuperGLUE benchmark [18], despite its comparatively smaller size.

Specialized domain adaptation further enhances language comprehension. SciBERT [8] and S2ORC-SciBERT [9], trained from scratch using science and technology datasets exemplify this. SciDeBERTa [12], a DeBERTa extension tailored for computer science through continuous learning, demonstrates leading performance in specific tasks. In the biomedical field, BioBERT [10] leverages transfer learning from BERT pre-trained in the general domain, while PubMedBERT [11], trained from scratch using biomedical literature, excels in the BLURB benchmark. BioLinkBERT [19], with its focus on linked documents, currently leads in the BLURB leaderboard [11].

The emergence of generative LMs like BioGPT (1.5B) [15], and MedPalm (540B) [20] in the biomedical domain has shown promising results, particularly in the PubMedQA task. A comprehensive survey of medical domain-specific LLMs [3] analyzed the overall process of the LLMs as pre-training, medical-domain fine-tuning, and prompting. The prompting process is the case when LLMs(over 10B) are used with in-context-learning characteristics.

Our study leverages DeBERTa-v2 (100M), in conjunction with the SentencePiece tokenizer [21]. Our approach advances beyond the continuous learning methodology of the SciDeBERTa [12] to yield an optimized LM within the biomedical domain. This study's methodology reveals that a biomedical LM based on DeBERTa, trained progressively, mirrors the specialized trajectory typical in biomedical education, resulting in optimal performance.

III. PRETRAINING PROCEDURE OF MEDIBIODEBERTA

Domain-specific PLM development requires a careful design that considers the above differences in training methods and various factors such as the type of training data and hyper-parameters. This study aims to provide guidelines for the development of domain-specific PLMs. We analyze factors critical for optimizing PLMs to specific domains and assess the impact of each factor on performance. We specifically demonstrate the development process of MediBioDeBERTa, a biomedical domain-specific PLM.

The development process of MediBioDeBERTa, as illustrated in Fig. 1, encapsulates the following three key steps:

- Choosing DeBERTa as the base large language model (LLM), balancing computational cost and performance.
- Designing a domain-specific pretraining strategy by analyzing PLM performance changes based on pretraining

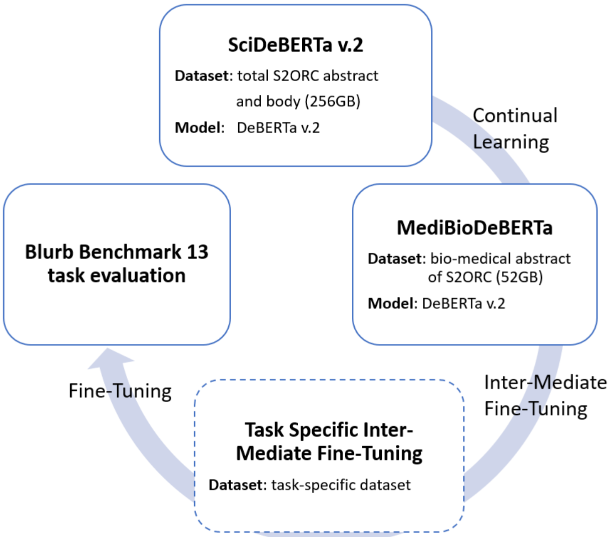
SciDeBERTa v.2
Dataset: total S2ORC abstract and body (256GB)
Model: DeBERTa v.2
Continual Learning
Blurb Benchmark 13 task evaluation
MediBioDeBERTa
Dataset: bio-medical abstract of S2ORC (52GB)
Model: DeBERTa v.2
Fine-Tuning
Inter-Mediate Fine-Tuning
Task Specific Inter-Mediate Fine-Tuning
Dataset: task-specific dataset

FIGURE 1: Summary of the MediBioDeBERTa training process.

methods and datasets.
- Optimizing performance for each task, leveraging results from the domain-specific PLM.

A survey in medical domain-specific LLMs [22] supports our process, which includes pre-training, medical-domain fine-tuning, and prompting.

A. BASE MODEL AND PRETRAINING MECHANISM

Pretraining approaches typically fall into two categories. The first involves training a randomly initialized model from scratch. The second method is to continue training a PLM that has already been pre-trained with other knowledge. The latter is often considered when developing a specialized LM because of the difficulty in securing data. In terms of performance, a PLM trained from scratch is generally expected to yield better performance due to its high domain-specific knowledge. Analyzing representative results in the biomedical field, PubMedBERT [11], which pretrains BERT from scratch, exhibited better overall performance than BioBERT [10], which pretrains BERT continuously. However, BioLinkBERT [19], which also pretrains BERT continuously, demonstrated that enhanced performance can be achieved by refining the pretraining algorithm.

Our study compares PLMs across different pretraining methodologies to establish an optimized process for biomedical PLMs. The comparative analysis, detailed in Section V, reveals that scratch-based training using domain-specific data typically outperforms continuous learning, as shown by PubMedBERT and BioBERT. Nonetheless, the most effective approach was training from scratch with general science data, followed by continuous learning in the biomedical domain (see Fig. 1).

In developing MediBioDeBERTa, we initially evaluated existing LMs, selecting DeBERTa as the foundation due to its balance of performance and computational efficiency. Our analysis, detailed in Table 5 in Section V, indicated that SciDeBERTa was less suitable as a base model. Thus, we refined SciDeBERTa's pretraining process to create SciDeBERTa v2, which, when further trained with a biomedical subset from the S2ORC dataset, led to the development of MediBioDeBERTa, a PLM tailored for the biomedical field.

Our findings suggest that choosing a domain-specific base model enhances performance compared to a general domain model. MediBioDeBERTa, for instance, benefits from a graduated training approach, mirroring the progressive specialization typical in biomedical studies. For an in-depth exploration of these experiments, see Section V-C.

B. ANALYSIS OF SCHOLARLY CORRELATION IN DATASET

It is crucial to perform analysis while constructing a dataset corresponding to a specific domain. In the experiment, we constructed a biomedical domain dataset by selectively extracting papers included in the 'biology' and 'medicine' categories from the S2ORC dataset [9] based on human intuition. However, the experimental results showed that the biomedical knowledge from this set was insufficient to solve the BLURB tasks. Therefore, we performed data analysis to construct an optimally refined dataset.

Our study utilized the S2ORC dataset, which consists of 81.1M scientific papers accompanied by detailed metadata tags. This dataset encompasses 19 science and technology disciplines, with each paper often classified under multiple disciplines. In our analysis, we aimed to identify disciplines that frequently overlap with the 'medicine' category.

The results of the correlation analysis across the 19 categories are shown in Fig. 2. We defined the disciplinary correlation C_{ij} between i row and j column as:

$$C_{ij} = \frac{N_{ij}}{N_i} = \frac{N_i \bigcap N_j}{\bigcup_j N_{ij}} \qquad (1)$$

where C_{ij} quantifies the proportion of papers in discipline i that are associated with discipline j. Here, N_i denotes the total count of papers in discipline i, and N_{ij} represents the number of papers that span both disciplines i and j within the S2ORC dataset's 'mag field of study'. N_{ii},for $i == j$, represents the proportion of research conducted within a single discipline, as depicted on the diagonal of Fig. 2. The mathematical notation C_{ij} also indicates the degree of interdisciplinary collaboration research between the field of discipline in region i and the field of discipline in region j.

In summary, the categories most frequently associated with 'medicine' were 'biology,' 'chemistry,' and 'psychology.' The correlations between these four fields are shown in Fig. 3. We therefore extracted the subset of data related to these four disciplines (a combined volume of 52 GB post-deduplication) as the dataset for continuous learning, as illustrated in Fig. 1. That is, we developed MediBioDeBERTa from SciDeBERTa v.2 through continual learning on

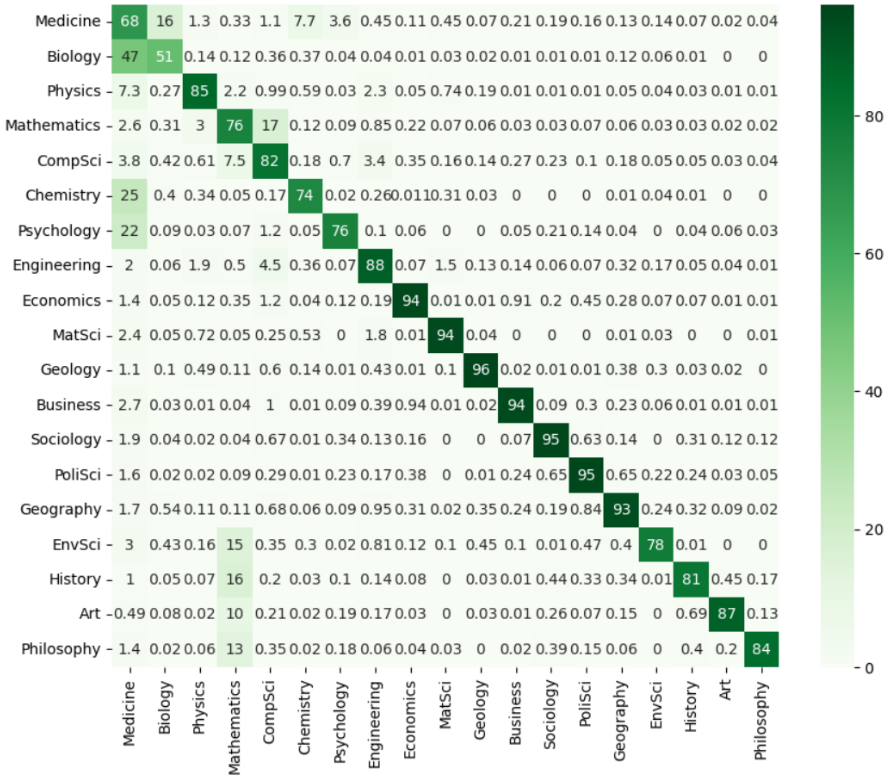

FIGURE 2: Confusion matrix of scholarly correlation for the 19 categories in the Semantic Scholar Open Research Corpus(S2ORC) dataset.

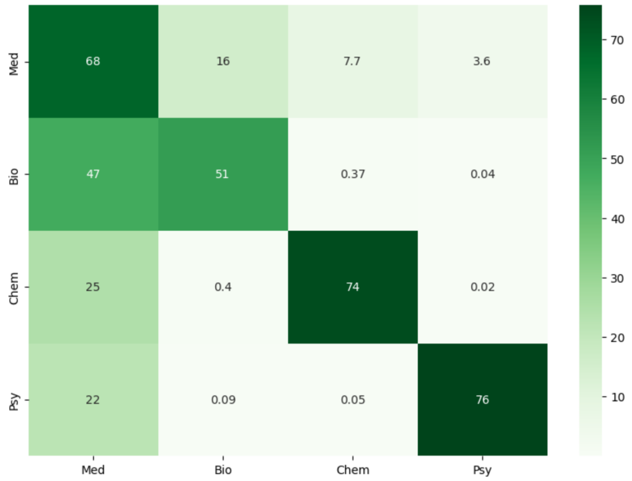

FIGURE 3: Confusion matrix of scholarly correlation for medicine-related categories in S2ORC.

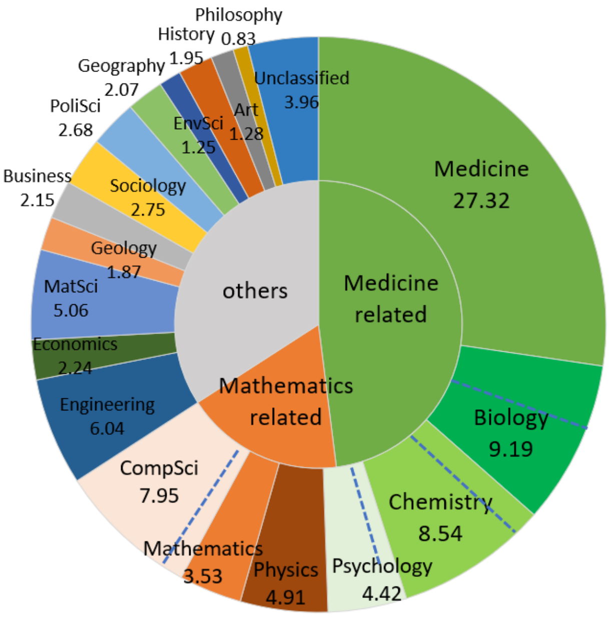

FIGURE 4: Category distribution of the S2ORC dataset and classification results.

an extracted subset of data, i.e., a Medicine-related dataset. Fig. 4 demonstrates the distribution of the 19 categories in the S2ORC dataset and the medicine-related, mathematics-related, and other categories classified in our study. When dividing the three types of major academic disciplines in Fig. 4, the fields of Medicine and Mathematics, which have a high proportion of interdisciplinary research, were extracted. The proportion of each major discipline in the total academic fields was calculated using the union operation for the interdisciplinary research areas. Mathematically, the major academic discipline B_i is defined as $B_i = \bigcup_j C_{ij}$, representing the aggregate interdisciplinary research areas related to discipline i.

IV. INTERMEDIATE FINE-TUNING FOR TASK KNOWLEDGE TRANSFER

Pretrained language models (PLMs), which learn general language representations, are typically fine-tuned with task-specific datasets to solve the downstream tasks. A key challenge in transfer learning is aligning the domain of the pre-training dataset with the target task. Previous studies have focused primarily on linguistic domains. This section discusses transfer learning improvements, focusing on enhancing the task knowledge of PLMs through intermediate fine-tuning (IFT).

A. TASK KNOWLEDGE IN PLMS

PLMs acquire task and linguistic knowledge during pre-training, which subsequently impacts the performance of downstream tasks. Therefore, it's essential to account for the task knowledge of the base model during this phase. DeBERTa [7] embraces the RoBERTa [4] structure but excludes the next sentence prediction (NSP). Additionally, it refines the masked language model (MLM) mechanism to consider the relative positions between tokens. Such a modification improves performance in tasks emphasizing token-to-token relations, closely mirroring MLM tasks. However, it has demonstrated limited efficacy for tasks that emphasize sentence-to-sentence or document-to-document correlations. DeBERTa suggests that for sequence pair tasks in the GLUE benchmark [17], better performance is achievable through IFT, particularly in tasks like multi-genre natural language inference (MNLI) for similarity prediction (STS-B), paraphrase identification (MRPC), and NLI (RTE).

B. INTERMEDIATE FINE-TUNING

Intermediate fine-tuning (IFT) has proven effective in enhancing downstream task performance, as demonstrated by DeBERTa [7]. Our study concentrates on intermediate tasks that leverage the task knowledge embedded in the base PLM. Initially, we evaluated task knowledge at the sentence level within the PLM. Table 1 compares the performance of each PLM on BIOSSES, the sentence similarity (SS) task of the BLURB [11]. Both BERT[2] and LinkBERT[19], pre-trained with NSP and its improved method, exhibited higher performance than DeBERTa. This difference is attributed to the varied task knowledge related to sentence relationship identification developed during pretraining. DeBERTa, when fine-tuned with MNLI, achieved comparable performance to BERT. Consequently, we employed IFT to augment the task knowledge of our model, MediBioDeBERTa.

TABLE 1: Performance of the BIOSSES task (micro F1).

BERT	LinkBERT	DeBERTa	DeBERTa$_{MNLI}$
83.21	87.19	76.14	83.15

† Our experimental results are average values of five runs.

BIOSSES, a semantic sentence similarity estimation task in the biomedical domain, and MNLI, an intermediate task from the general domain, demonstrate that IFT using general domain datasets can be effectively enhance performance in specialized domains. This approach is particularly useful in fields where acquiring domain-specific datasets is challenging. Table 2 summarizes the tasks and datasets for the IFT used in the experiments. Experimental results are presented in Section V-E.

C. INTERMEDIATE FINE-TUNING BY MULTITASK AND MULTI-FORMAT

Different perspectives of a task can reveal various aspects. Employing different metrics for the same dataset elucidates different classifications and similarities. According to SciRepEval [23], integrating varied task types during the training process enhances LM performance. Additionally, combining multitask learning with multi-format learning further improves results. In other words, enhancing sentence comprehension through tasks like NER, RE, and co-reference resolution (Coref), as well as incorporating regression tasks that use different metrics, such as report references frequency, can lead to performance gains. We attempted to apply these principles in our IFT approach. Table 2 outlines the target tasks for performance enhancement via IFT, along with their respective datasets and evaluation metrics. For the hallmarks of cancer (HoC) task involving document classification, the IFT was conducted using both regression with Kendall's τ metric and classification with the macro F1 metric. For the PubMedQA task, which involves QA, IFT combined proximity and search tasks. For the detailed experimental results, please refer to Section V-E and Table 7.

V. EXPERIMENT AND RESULTS

We utilized the DeBERTa-v2 (12-layer base model) [7] with 128K SentencePiece [21] tokens. Unlike DeBERTa-v1 which adopts RoBERTa's byte pair encoding (BPE), it employs SentencePiece due to memory constraints associated with byte-level tokenization. In this experiment, SentencePiece was used to train both SciDeBERTa v2 and MediBioDeBERTa. It took approximately 40 days to train SciDeBERTa v2 from scratch using a 256GB S2ORC scholar dataset on an A100 2-node connected by 40GB 8 NVLinks. Approximately, 67 hours were required to train MediBioDeBERTa using a 52GB medibio dataset based on SciDeBERTa v2 in a continuous-learning manner using an A100 3 node connected by 80GB of 8 NVLinks for 10K steps. See Section III for details on the dataset selection.

TABLE 2: Summary of the tasks and dataset for IFT.

Task	Dataset Name	Format†	Train Used/Total	Dev Used/Total	Eval Metric
BIOSSES	Clinical semantic textual similarity	CLF	2392/2392	730/730	Pearson
BIOSSES	MNLI semantic textual similarity	CLF	5750/5750	1501/1501	Pearson
HOC	citation count publication year	RGN	175K/175K 198K/198K	26K/26K 19K/19K	kendall's τ
	mesh descriptors field of studies	CLF	600K/2069K 500K/541K	40K/258K 40K/67K	Macro F1
PubMedQA	citation prediction	PRX	600K/676K	50K/143K	MAP
	search same author detection highly influential citation	SRCH	453K/453K 67K/67K 58K/58K	75K/75K 8.9K/8.9K 7K/7K	nDGC

† Format abbreviation: classification (CLF), regression (RGN), proximity (PRX), adhoc search (SRCH)

TABLE 3: Hyperparameters for pretraining of SciDeBERTa v2 and MediBioDeBERTa.

Hyperparameter	Assignment	
	SciDeBERTa v.2	MediBioDeBERTa
max training steps	500K	10K
warmup steps	50K	1K
batch size	8,192	49,152
learning rate	0.0001	0.0005
optimizer	AdamW	AdamW
weight decay	0.01	0.01
learning rate decay	linear	linear

TABLE 4: Comparison of the test performances (F1-score) of SciDeBERTa and SciDeBERTa v2.

Model	SciERC		
	NER	JRE	Coref
SciDeBERTa [12]	71.1 ± 0.6	46.0 ± 0.8	57.4 ± 0.6
SciDeBERTa v2	**72.4 ± 0.4**	**47.4 ± 1.2**	56.9 ± 0.8

† Our experimental results are average values of five runs.

A. HYPERPARAMETERS OF SCIDEBERTA V2 AND MEDIBIODEBERTA

The hyperparameters used for pretraining follow the training conditions of DeBERTa [7]. Table 3 provides further details. For MediBioDeBERTa, the training batch size per device was 4,096; 3 nodes were used, and the accumulated updates were performed 4 times. Thus, the total batch size is 49,152 $(4,096 \times 3 \times 4)$. The warmup was performed 1,000 times, which was 10% of the total steps.

B. PERFORMANCE COMPARISON OF SCIDEBERTA AND SCIDEBERTA V2 IN SCIERC DATASET

SciDeBERTa [12] is a model trained through continual learning on the S2ORC abstract dataset based on the DeBERTa architecture. In contrast, SciDeBERTa v2 is a domain-specific knowledge model trained from scratch on the S2ORC full dataset. We evaluated and compared the performances of SciDeBERTa [12] and SciDeBERTa v2 on the SciERC dataset. As shown in Table 4, SciDeBERTa v2 outperforms SciDeBERTa on the SciERC NER and JRE tasks.

C. PERFORMANCE COMPARISON OF DOMAIN-SPECIFIC PLMS BASED ON TRAINING METHODS AND DATASETS

In this section, we compared the model performance according to the pretraining algorithm and corpus type of medicine-related data, as detailed in Section III-B and illustrated in Fig. 4. We examined the influence of the pretraining algorithm, the base model, and the type of training data, given their pivotal roles in the development of domain-specific LMs. Table 5 summarizes each model's configuration and average performance on the BLURB benchmark. The detailed results for each task are presented in Table 6. We selected the #1 model as the MediBioDeBERTa from the experimental results.

The performance results of models #2 and #3 align with those of the existing PubMedBERT [11] and BioBERT [10], indicating that training specialized domain data from scratch yields better results than fine-tuning on top of a general domain model. The most favorable outcome was observed in model #1, in which the base LM for continuous learning was also specialized in the language domain. Based on these findings, we utilized MediBioDeBERTa with SciDeBERTa v2 as the base model for continuous learning.

In models #3 and #4, we observed variations in performance depending on the type of corpus. When dealing with papers predominantly used in specialized fields, it is crucial to decide whether to rely solely on relatively simpler abstracts or to integrate comprehensive full-text encompassing various formats and contents. Previous study [11] showed that full-text data yield better performance when trained sufficiently to acquire complex knowledge.

However, our empirical observations did not corroborate a significant enhancement in performance. Moreover, adjusting the corpus type to abstracts and pursuing additional continuous learning diminished performance. This suggests that the efficacy of the study [11] is contingent on the adequacy of the training data volume.

Conclusively, the comparative experiments presented in Table 5 substantiate that biomedical language models manifest optimal performance when trained in a graduated manner, similar to the progressive specialization in biological or

TABLE 5: PLM configuration and its average performance of the BLURB benchmark.

	Pretraining Algorithm	Base model	Corpus Type	Training Steps	BLURB Avg.
#1	continual learning	SciDeBERTa v2	abstract	90,000	**78.03**
#2	continual learning	DeBERTa$_{base}$v3	abstract	100,000	76.65
#3	from scratch	DeBERTa-v2	abstract	125,000	77.38
#4	from scratch	DeBERTa-v2	fulltext	125,000	76.46
#5	continual learning	#4	abstract	50,000	76.38

† Learning data is the medicine-related subset of S2ORC dataset described in III-B.
† Our experimental results are average values of five runs.

TABLE 6: Comparison of the test performances of the BLURB benchmark for PLM for Table 5.

Task	Dataset	#1	#2	#3	#4	#5
NER	BC5-chem	93.04	93.11	92.39	91.79	92.05
	BC5-disease	85.13	82.18	84.51	84.21	84.68
	NCBI-disease	89.08	89.19	89.05	88.78	88.23
	BC2GM	83.94	84.10	83.27	83.20	83.46
	JNLPBA	80.23	80.13	79.76	79.80	79.79
PICO	EBM PICO	73.73	74.12	73.59	73.67	73.75
RE	Chem Prot	77.80	74.27	75.53	74.02	74.95
	DDI	80.47	80.75	79.35	79.81	80.07
	GAD	82.23	81.79	78.32	80.17	79.93
SS	BIOSSES	57.68	66.56	62.78	48.39	42.52
DC	HOC	61.09	68.74	70.28	69.13	66.93
QnA	PubmedQA	56.84	49.84	52.00	59.36	58.12
	BioASQ	93.14	71.71	85.14	81.71	85.57
BLURB Avg.		**78.03**	76.65	77.38	76.46	76.38

† Our experimental results are average values of five runs.

TABLE 7: Comparison of the test performances of MediBioDeBERTa with other models in the BLURB benchmark.

Task (metric)	Dataset	BioLinkBERT (base)	SciDeBERTa v2 (full-FS)	MediBio DeBERTa	MediBio DeBERTa-IFT
NER(F1)	BC5-chem	93.38	92.75	**93.04**	
	BC5-disease	**85.45**	84.27	85.13	
	NCBI-disease	88.12	89.89	**89.08**	
	BC2GM	**84.39**	83.97	83.94	
	JNLPBA	78.78	66.2	**80.23**	
NER Avg.		86.02	83.42	**86.28**	
PICO(Macro F1)	EBM PICO	**74.2**	73.69	73.73	
RE(Micro F1)	Chem Prot	**78.1**	76.86	**93.04**	
	DDI	**81.12**	78.75	80.47	
	GAD	**82.51**	80.2	82.23	
RE Avg.		**80.81**	78.60	80.17	
SS(Micro F1)	BIOSSES	92.5	59.54	57.68	**92.7**
DC(Micro F1)	HOC	**84.73**	61.42	61.09	71.49
QnA(Accuracy)	PubmedQA	58.32	51.99	56.84	**59.33**
	BioASQ	91.57	67.86	**93.14**	
QnA Avg.		74.95	59.93	74.99	**76.24**
BLURB Avg.		**82.61**	74.41	78.03	81.72

† Our experimental results are average values of five runs.

medical studies built upon a broad science education.

D. PERFORMANCE COMPARISON OF MEDIBIODEBERTA IN THE BLURB BENCHMARK

As described in Table 7, the MediBioDeBERTa achieved the best performance in three tasks, named entity recognition (NER), sentence similarity (SS), and question & answering (Q&A), with average scores of 86.28%, 92.7%, and 76.32%, respectively. However, BioLinkBERT, which accommodates extensive document cross-references, outperformed the patient intervention comparison outcomes (PICO), relation extraction (RE), and document classification (DC) tasks. MediBioDeBERTa, encompassing not only 'medicine' and 'biology,' but also 'chemistry' category articles, showed a significant improvement of 14.94% over BioLinkBERT [19] in the ChemProt RE task, with an F1 score of 93.04%. Furthermore, the IFT of MediBioDeBERTa enhanced the performance by 35.02%, 10.4%, and 1.49% in the BIOSSES, HOC, and PubmedQA tasks, respectively.

E. EXPERIMENTS FOR IFT OF MEDIBIODEBERTA

We demonstrated the results of the IFT to improve the performance of sequence pair tasks. We aimed to transfer task knowledge through the IFT and tested both general and biomedical domain datasets to investigate dependencies based on the domain of the dataset used for the IFT. We used the Semantic Textual Similarity Benchmark (STSB) task of GLUE benchmark and ClinicalSTS datasets as the general and biomedical domain datasets, respectively. ClinicalSTS used both the ClinicalSTS2018 [24] and ClinicalSTS2019 datasets [25]. The IFT using both datasets improved the performance of the BIOSSES task, which is a sentence similarity task of the BLURB leaderboard, as shown in Table 2. An interesting observation from the experimental results was that using a general domain dataset led to better performance than using a biomedical domain dataset. This suggests that using a target domain dataset for task knowledge transfer through the IFT is not always necessary. This is predicted because domain knowledge has already been sufficiently learned during pretraining.

As suggested in SciREpEval models[23], we utilized both regression and classification formats for the IFT of the HOC task, which is a document classification task. The regression format consisted of the citation count and year of publication and the classification format included mesh descriptors and fields of study as described in Table 2.

Comparing the HOC performance of model #1 in Table 6 before applying IFT and MediBioDeBERTa in Table 7 after applying IFT, the performance increased by 10.4%. Similarly, four tasks in two different formats, prediction, and search, were employed to enhance the performance of PubMedQA. The prediction format utilized a citation prediction dataset and the search format involved searching for the same author and high-influence citations. This approach led to a performance improvement of 1.49%, as shown in Table 7.

VI. CONCLUSION

This study first presented SciDeBERTa v2, an LM trained from scratch on a scientific domain-specific S2ORC dataset using DeBERTa. SciDeBERTa v2 achieved superior performance compared to its predecessor, SciDeBERTa. To adapt the model to the bio-medical domain, we extracted biomedical data from S2ORC using correlation analysis and trained MediBioDeBERTa. Applying IFT enabled us to improve domain-specific task performance simply and effectively. Our model, MediBioDeBERTa, outperformed the state-of-the-art models in categories such as NER, SS, and Q&A and the ChemProt RE task, ranking 11th in the BLURB leaderboard with an average score of 81.72. Recently, generative language models like BioGPT [15] have been employed to directly pose questions and evaluate responses. In contrast, this study leverages an MLM-based NLU model, evaluating the QA task performance as a sequence classification task. Nevertheless, MediBioDeBERTa remains valuable for medical information processing applications, such as NER and sequence classification-based QA. These applications are crucial for extracting features in paragraphs through language understanding. Future work will involve scaling up MediBioDeBERTa to a 24-layer model based on the recent deberta v3 architecture [26] and integrating specialized biomedical news data. We anticipate improvements in applications related to infectious diseases.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in NAACL-HLT (1), 2019.

[3] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong et al., "A survey of large language models," arXiv preprint arXiv:2303.18223, 2023.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[5] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," arXiv preprint arXiv:1909.11942, 2019.

[6] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: Improving pre-training by representing and predicting spans," Transactions of the Association for Computational Linguistics, vol. 8, pp. 64–77, 2020.

[7] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," in International Conference on Learning Representations, 2021. [Online]. Available: https://openreview.net/forum?id=XPZIaotutsD

[8] I. Beltagy, K. Lo, and A. Cohan, "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), 2019, pp. 3615–3620.

[9] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. S. Weld, "S2orc: The semantic scholar open research corpus," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4969–4983.

[10] 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, pp. 1234–1240, 2020.

[11] 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 Transactions on Computing for Healthcare (HEALTH), vol. 3, no. 1, pp. 1–23, 2021.

[12] Y. Jeong and E. Kim, "Scideberta: Learning deberta for science technology documents and fine-tuning information extraction tasks," IEEE Access, vol. 10, pp. 60 805–60 813, 2022.

[13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners," Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.

[14] E. Lehman, E. Hernandez, D. Mahajan, J. Wulff, M. J. Smith, Z. Ziegler, D. Nadler, P. Szolovits, A. Johnson, and E. Alsentzer, "Do we still need clinical language models?" Proceedings of Machine Learning Research, Conference on Health, Inference, and Learning, vol. 209, pp. 578–597, 2023.

[15] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "Biogpt: generative pre-trained transformer for biomedical text generation and mining," Briefings in Bioinformatics, vol. 23, no. 6, 2022.

[16] Y. Luan, L. He, M. Ostendorf, and H. Hajishirzi, "Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3219–3232.

[17] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," in International Conference on Learning Representations, 2018.

[18] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 4938–4947.

[19] M. Yasunaga, J. Leskovec, and P. Liang, "Linkbert: Pretraining language models with document links," in ICML 2022 2nd AI for Science Workshop, 2022.

[20] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl et al., "Large language models encode clinical knowledge," Nature, vol. 620, no. 7972, pp. 172–180, 2023.

[21] T. Kudo and J. Richardson, "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, 2018, pp. 66–71.

[22] H. Zhou, B. Gu, X. Zou, Y. Li, S. S. Chen, P. Zhou, J. Liu, Y. Hua, C. Mao, X. Wu, Z. Li, and F. Liu, "A survey of large language models in medicine:progress, application, and challenge," arXiv preprint arXiv:2311.05112v1, 2023.

[23] A. Singh, M. D'Arcy, A. Cohan, D. Downey, and S. Feldman, "Scirepeval: A multi-format benchmark for scientific document representations," arXiv preprint arXiv:2211.13308, 2022.

[24] Y. Wang, N. Afzal, S. Liu, M. Rastegar-Mojarad, L. Wang, F. Shen, S. Fu, and H. Liu, "Overview of the biocreative/ohnlp challenge 2018 task 2: clinical semantic textual similarity," Proceedings of the BioCreative/OHNLP Challenge, vol. 2018, 2018.

[25] Y. Wang, S. Fu, F. Shen, S. Henry, O. Uzuner, H. Liu et al., "The 2019 n2c2/ohnlp track on clinical semantic textual similarity: overview," JMIR medical informatics, vol. 8, no. 11, p. e23375, 2020.

[26] P. He, J. Gao, and W. Chen, "Deberta v3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing," International Conference on Learning Representations(ICLR), 2023.

EUNHUI KIM received the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Korea, in 2009 and 2015, respectively, and received the B.S. degree in information communication engineering from Chungnam National University, Korea, in 2000. From 2000 to 2007, she was a researcher with Samsung Electronics in Seoul, Korea. From 2015 to 2018, she was a postdoctoral researcher at KAIST, Korea. In 2018, she was an invited professor with the National Center of Excellence in Software, at Chungnam National University, Korea. Since 2019, she has worked in the Data AI Center of the Korea Institute of Science and Technology Information as a senior researcher. Dr. Kim's research interests include machine learning, recommendation systems, lightweight deep neural network modeling in vision and language processing, and language modeling and its applications.

YUNA JEONG received a B.S. degree in computer engineering at Korea Polytechnic University (2012) and a Ph.D. degrees in computer engineering at Sungkyunkwan University (2019). She is a senior researcher in the Open XR Platform Research Center at the Korea Institute of Science and Technology Information (KISTI). Her main research interests include computer graphics, deep learning, and natural language processing.

MYUNG-SEOK CHOI received his B.S., M.S., and Ph.D. degrees from the Department of Computer Science at the Korea Advanced Institute of Science and Technology (KAIST), Korea, in 1996, 1998, and 2005, respectively. Since joining the Korea Institute of Science and Technology Information (KISTI) in 2005, he has held various roles and is currently the director of the AI Data Research Center at KISTI. His research interests are in the fields of machine learning, natural language processing, and open science.