Fine-tuning Pretrained Language Models for Biomedical Tasks

Monash University

Jirarote Jirasirikul, Ehsan Shareghi, Reza Haffari

Abstract

Contextualized representation using **pre-trained language models was able to encode high-quality semantics** that offers good performance on downstream NLP tasks. The well-known **BERT architecture achieved state-of-the-art** results on various tasks and benchmarks.

Biomedical BERT Language Model:

- BioBERT – Continual pre-training
- PubMedBERT – From scratch pre-training

Biomedical NLP Benchmark Tasks:

- Hallmark of cancer (HoC)
- PubMedQA

We fine-tune these language models via various strategies such as **input preprocessing**, **ensembling**, and **adapter layers**.

Executive Summary:

- The necessity of using domain-specific language models
- The ensembling technique builds a more robust model that enhance precision in exchange for recall.
- The adapter requires extensive used of GPU to fine-tuning downstream tasks together with the language model.
- Adding more context could confuse models and create inadequate contextualized representation

BLURB Leaderboard

Model	BLURB Score	HoC	PubMedQA
PubMedBERT (uncased)	81.50	82.62	55.84*
BioBERT (cased)	80.34	81.54	60.24*
BERT base (uncased)	76.11	80.20	51.62*

Acknowledgement

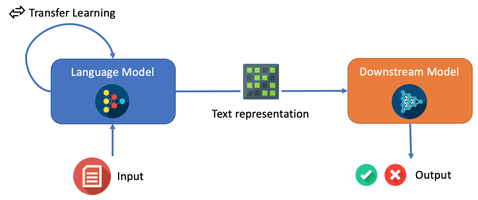
We appreciate huggingface.co and adapterhub.ml for contribution toward the NLP community

The code of our experiment are available at:
https://github.com/blizrys/BioMed-BERT-Eval

Methodology

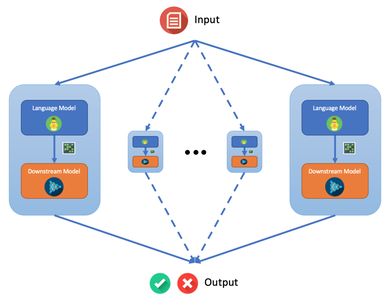
1. Input Preprocessing

In this research, we believe that **each dataset have its characteristics**. We observe the dataset from a statistical and practical point of view. We **preprocess the data before feeding it into our corresponding model**, freeze the language model and fine-tuned the downstream task layer.



2. Ensemble Model

The ensemble method is a machine learning technique that **combines several base models to produce one optimal predictive model**. We set up this ensemble model by generating best configuration checkpoints with a mixture of BioBERT and PubMedBERT language models.



3. BERT Adapter

BERT Adapter is a **new lightweight architecture** as an alternative to full fine-tuning of a pre-trained language model on a downstream task. It consists of **adding only a small set of newly introduced parameters** in between the transformer layers. We instantiate adapter-based tuning models using the proposed architecture and compare them against the traditional top layer fine-tuning approach on Biomedical NLP tasks

Experiments and Results

1. Input Preprocessing

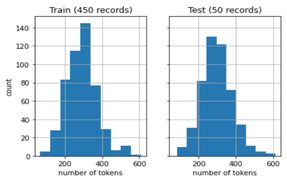
Hall of Cancer

We discover that adding multiple sentences into language model could confuse the contextualize representation. The best performance is demonstrated when n=0

Hallmark of Cancer			Previous n-sentences (F1 score)			
batch size	lr.	Language Model	n = 0	n = 1	n = 2	n = 3
16	5e-5	PubMedBERT (uncased)	**81.26**	79.54	79.11	77.93
		BioBERT (cased)	**82.33**	79.70	77.37	77.97
		BERT base (uncased)	**71.44**	67.32	67.55	65.72
32	5e-5	PubMedBERT (uncased)	**80.87**	79.30	79.32	77.90
		BioBERT (cased)	**82.47**	79.40	78.09	77.12
		BERT base (uncased)	**70.61**	66.28	66.87	64.75

PubMedQA

Without an approach to select abstract sentences, a portion of 1.55% of the training dataset and 16% of the test dataset exceed BERT maximum token processable.



2. Ensemble Model

Hall of Cancer

- 'Majority Vote' ensembling can improves precision.
- 'At least one model' ensembling can improves F1 BUT also add false negatives.

Both types of ensembling is useful depending on the objectives.

Mix BioBERT and PubMedBERT	Label count			Majority Vote			Atleast 1 model		
batch size = 32, lr = 5e-5	Train	Dev	Test	Pres.	Recall	F1score	Pres.	Recall	F1score
activating invasion and metastasis	448	64	62	91.89	54.84	68.69	83.87	83.87	83.87
avoiding immune destruction	185	20	14	**58.33**	50.00	**53.85**	**42.86**	64.29	**51.43**
cellular energetics	136	44	19	90.91	52.63	66.67	88.24	78.95	83.33
enabling replicative immortality	219	11	24	100.00	54.17	70.27	76.92	83.33	80.00
evading growth suppressors	268	23	53	71.43	**18.87**	29.85	65.22	**56.60**	60.61
genomic instability and mutation	523	103	65	80.00	36.92	50.53	68.57	73.85	71.11
inducing angiogenesis	238	60	23	88.24	65.22	75.00	67.86	82.61	74.51
resisting cell death	602	72	79	95.35	51.90	67.21	79.01	81.01	80.00
sustaining proliferative signaling	674	85	105	76.60	34.29	47.37	71.28	63.81	67.34
tumor promoting inflammation	375	45	40	80.00	30.00	43.64	71.79	70.00	70.89

PubMedQA

The scores are more consistent when ensembling more models. Mixing language models only show a minimal drop in the score while maintaining overall performance. Ensembling is a powerful approach to construct a robust model.

PubMedQA	batch size = 32, lr = 3e-3	Weight Average		
Language Model	Ensemble Model	Precision	Recall	F1 score
All PubMedBERT	No ensemble (previous score)	56.56	59.20	57.12
	10 models	**56.55**	**59.80**	**57.57**
	100 models	55.68	59.00	56.57
	1000 models	55.48	58.40	56.14
All BioBERT	No ensemble (previous score)	50.04	52.20	50.94
	10 models	53.11	56.20	54.21
	100 models	53.26	57.00	54.59
	1000 models	53.60	57.00	54.66
Mix PubMedBERT and BioBERT	10 models	56.36	59.60	56.88
	100 models	55.66	59.40	55.93
	1000 models	55.59	59.60	56.13

3. BERT Adapter

BERT adapters can achieve the comprehensive performance with a traditional fine-tuned downstream model. However, the adapter fine-tunes the downstream parameters within the language model. This causes a higher GPU usage is needed than the traditional fine-tuned downstream model.

			HoC (F1 Score)			PubMedQA (Accuracy)		
batch size	lr.	Language Model	BLURB	Tradition	Adapter	BLURB	Tradition	Adapter
8	1e-5	PubMedBERT (uncased)	82.62	**85.93**	83.04	55.84	**67.20**	55.20
	3e-3			53.68	53.68		55.20	**62.58**
	5e-5			56.28	**88.64**		55.28	**55.32**
8	1e-5	BioBERT (cased)	81.54	**87.14**	82.95	60.24	**53.52**	55.20
	3e-3			53.68	53.68		55.20	**58.06**
	5e-5			61.23	**88.84**		**56.74**	55.22
8	1e-5	BERT (uncased)	80.20	**84.55**	75.85	51.62	53.76	**55.20**
	3e-3			53.68	53.68		55.20	55.20
	5e-5			59.79	**86.64**		54.02	**55.16**

Conclusion

We recognize that **biomedical tasks can be sensitive** and might need to be **evaluated from various aspects** rather than just an accuracy or F1 score on the leaderboard. As such research that **focuses on improving performance scores might need to be aware of this delicate matter**. Some of our findings have pointed out some basic but necessary limitations of BERT models that should be addressed. Being more task-oriented when handling the biomedical domain might be crucial to adopt the model on to practical tasks. We hope that further research will be more mindful of these findings.