Comparing Self-Supervised Learning Pretext Tasks for Language Modeling

Dean Huang, Michael Tang, Betty Wu
Master's in Interdisciplinary Data Science, Duke University
December 2021

Abstract

Self-supervised learning is rapidly gaining popularity. It enables the model to learn from a large quantity of unlabeled dataset, and then the model can transfer the knowledge learned to perform specific downstream tasks.

Self-supervised learning generally has two main stages.

- *Pre-training*: train the model on a large general-purpose unlabeled dataset with various pretext tasks.
- *Fine-tuning*: use the weights learned in the pre-training stage as initial weights and train the model on a smaller labeled dataset to perform specific tasks.

In this project, we are interested in comparing the impact of different pretext tasks on a downstream task in language modeling. We found that Masked Language Modeling (MLM) performs the best, achieving 0.86 test accuracy; Sentence Order Prediction (SOP) produced the lowest accuracy at 0.77 on the IMDB dataset.

Objectives

The pretext tasks we explored were:

- **Masked Language Modeling** (MLM) (Devlin et al., 2018)
- **Next Sentence Prediction** (NSP) (Devlin et al., 2018)
- **Sentence Order Prediction** (SOP) (Lan et al., 2019)
- **Sentence Permutation** (SP) (Lewis et al., 2019)

The downstream task is **sentiment classification** on the IMDB dataset.

The model architecture is **3-layer Transformer encoders**.

Methods

Task #1 MLM

Randomly masked	A quick [MASK] fox jumps over the [MASK] dog
	↓ ↓
Predict	A quick brown fox jumps over the lazy dog

Task #2 NSP

Sentence 1	Sentence 2	Next Sentence
I am going outside	I will be back in the evening	yes
I am going outside	You know nothing John Snow	no

Task #3 SOP

Sentence 1	Sentence 2	Correct Order
I completed high school	Then I did my undergrad	yes
Then I did my undergrad	I completed high school	no

Task #4 SP

Original	I did X. Then I did Y. Finally I did Z.
Shuffle	Then I did Y. Finally I did Z. I did X.
Recover	I did X. Then I did Y. Finally I did Z.

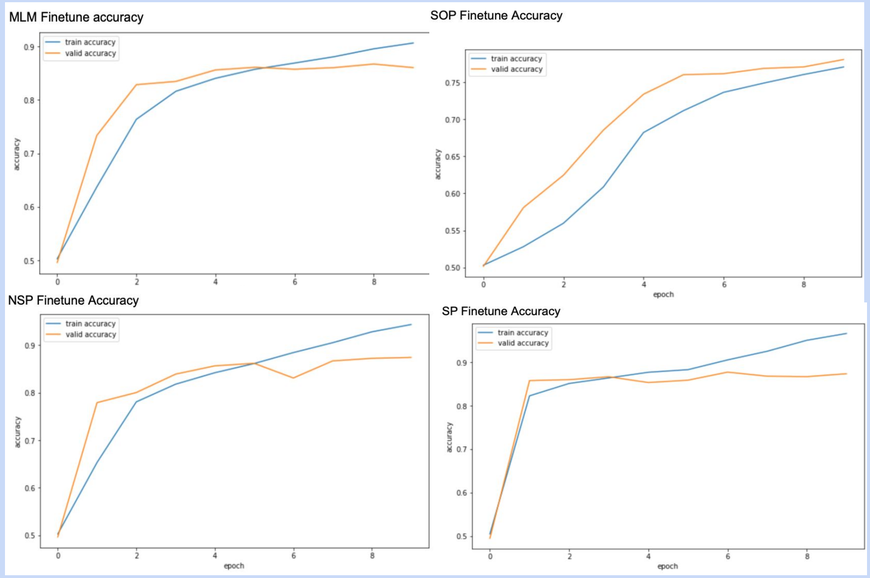
Results

In order for the results to be comparable among different pretext tasks, we used the same hyperparameters for all pre-training and fine-tuning tasks. All pre-training tasks were trained for 20 epochs, and fine-tuning tasks were trained for 10 epochs. In addition to each individual task, we also conducted ablation studies that combined different pretext tasks to compare the performance.

Hyperparameters:

- Tokenizer: BERT Uncased Tokenizer
- Batch Size: 16
- Embedding Size: 256
- Learning Rate: 3e-4
- Dropout Rate: 0.1
- Adaptive Learning Rate Scheduler: Cosine Schedule with Warm-up

	Best Pretrain Accuracy	Finetuning Test Accuracy
Baseline		0.861
MLM	0.035	0.860
NSP	0.578	0.858
SOP	0.510	0.773
SP	0.250	0.851
MLM + NSP	0.297	0.849
MLM + SOP	0.273	0.861
MLM + SP	0.249	0.851



Conclusion

Pre-training Stage:

Across all pretext tasks including hybrid pretext tasks,

- **NSP** and **SOP** perform the best.
- **MLM** has the lowest accuracy.

Fine-tuning Stage:

Across all pretext tasks including hybrid pretext tasks,

- **MLM** and **MLM + SOP** produced the best performances with test accuracy at 0.86.
- **SOP** was the least accurate task with accuracy at 0.77.
- Other tasks have accuracy at around 0.85.

Overfitting/Underfitting:

- All but SOP show signs of overfitting in both pre-training and fine-tuning stages.
- SOP shows signs of underfitting.

Take-home Points:

- Compared to other tasks, SOP converges slowly. More epochs are needed to achieve better results.
- SP and SP-hybrid tasks converge faster than others. Stronger regularizations may be needed to mitigate overfitting.
- Different tasks require different sets of hyperparameters in order to achieve better performance.
- The performance on the pre-training phases is *not* an indicator of the performance on the specific downstream task, as shown by MLM.

Future Work:

- Hyperparameter tuning to further boost performance.
- Regularizations to mitigate overfitting.
- Different combinations of pretext tasks to further investigate the marginal effect of each pretext task and how different tasks interact with each other.
- Pre-training to be performed on a larger dataset.

References

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." (https://arxiv.org/pdf/1810.04805.pdf)

Lan, Zhenzhong, et al. "Albert: A lite bert for self-supervised learning of language representations." (https://arxiv.org/pdf/1909.11942.pdf)

Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." (https://arxiv.org/pdf/1910.13461.pdf)

* The images in the Methods section were adapted from Chaudhary's blog in 2020 (https://amitness.com/2020/05/self-supervised-learning-nlp/)