Supplementary Material: Pseudo-Masked Language Models for Unified Language Model Pre-Training

Hangbo Bao [1] **Li Dong** [2] **Furu Wei** [2] **Wenhui Wang** [2] **Nan Yang** [2] **Xiaodong Liu** [2] **Yu Wang** [2] **Songhao Piao** [1] **Jianfeng Gao** [2] **Ming Zhou** [2] **Hsiao-Wuen Hon** [2]

1. Hyperparameters for Pre-Training

As shown in Table 1, we present the hyperparameters used for pre-training UniLMv2$_{\text{BASE}}$. We use the same WordPiece (Wu et al., 2016) vocabulary and model size as BERT$_{\text{BASE}}$ (Devlin et al., 2018). We follow the optimization hyperparameters of RoBERTa$_{\text{BASE}}$ (Liu et al., 2019) for comparisons.

Layers	12
Hidden size	768
FFN inner hidden size	3072
Attention heads	12
Attention head size	64
Max relative position	128
Training steps	0.5M
Batch size	7680
Adam ϵ	1e-6
Adam β	(0.9, 0.98)
Learning rate	6e-4
Learning rate schedule	Linear
Warmup ratio	0.048
Gradient clipping	0.0
Dropout	0.1
Weight decay	0.01

Table 1. Hyperparameters for pre-training UniLMv2$_{\text{BASE}}$.

2. GLUE Benchmark

Table 2 summarizes the datasets used for the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019).

Dataset	#Train/#Dev/#Test
Single-Sentence Classification	
CoLA (Acceptability)	8.5k/1k/1k
SST-2 (Sentiment)	67k/872/1.8k
Pairwise Text Classification	
MNLI (NLI)	393k/20k/20k
RTE (NLI)	2.5k/276/3k
QNLI (NLI)	105k/5.5k/5.5k
WNLI (NLI)	634/71/146
QQP (Paraphrase)	364k/40k/391k
MRPC (Paraphrase)	3.7k/408/1.7k
Text Similarity	
STS-B (Similarity)	7k/1.5k/1.4k

Table 2. Summary of the GLUE benchmark.

3. Hyperparameters for NLU Fine-Tuning

Table 3 reports the hyperparameters used for fine-tuning UniLMv2$_{\text{BASE}}$ over SQuAD v1.10 (Rajpurkar et al., 2016) / v2.0 (Rajpurkar et al., 2018), and the GLUE benchmark (Wang et al., 2019). The hyperparameters are searched on the development sets according to the average performance of five runs. We use the same hyperparameters for both SQuAD question answering datasets. Moreover, we list the hyperparameters used for the GLUE datasets in Table 3.

	SQuAD v1.1/v2.0	GLUE
Batch size	32	{16, 32}
Learning rate	2e-5	{5e-6, 1e-5, 1.5e-5, 2e-5, 3e-5}
LR schedule	Linear	
Warmup ratio	0.1	{0.1, 0.2}
Weight decay	0.01	{0.01, 0.1}
Epochs	4	{10, 15}

Table 3. Hyperparameters used for fine-tuning on SQuAD, and GLUE.

[1]Harbin Institute of Technology [2]Microsoft Research. Correspondence to: Li Dong <lidong1@microsoft.com>, Furu Wei <fuwei@microsoft.com>.

Proceedings of the 37^{th} *International Conference on Machine Learning*, Vienna, Austria, PMLR 119, 2020.

4. Hyperparameters for NLG Fine-Tuning

As shown in Table 4, we present the hyperparameters used for the natural language generation datasets, i.e., CNN/DailyMail (See et al., 2017), XSum (Narayan et al., 2018), and SQuAD question generation (QG; Du & Cardie 2018; Zhao et al. 2018). The total length is set to 512 for QG, and 768 for CNN/DailyMail and XSum. The maximum output length is set to 160 for CNN/DailyMail, and 48 for XSum and QG. The label smoothing (Szegedy et al., 2016) rate is 0.1. During decoding, we use beam search to generate the outputs. Length penalty (Wu et al., 2016) is also used to score candidates.

	CNN/DailyMail	XSum	QG
Fine-Tuning			
Learning rate	7e-5	7e-5	2e-5
Batch size	64	64	48
Weight decay		0.01	
Epochs	14	14	16
Learning rate schedule		Linear	
Warmup ratio	0.02	0.02	0.1
Label smoothing		0.1	
Max input length	608	720	464
Max output length	160	48	48
Decoding			
Length penalty	0.7	0.6	1.3
Beam size	5	5	8

Table 4. Hyperparameters used for fine-tuning and decoding on CNN/DailyMail, XSum, and question generation (QG).

References

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

Du, X. and Cardie, C. Harvesting paragraph-level question-answer pairs from wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 1907–1917, 2018.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Narayan, S., Cohen, S. B., and Lapata, M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, Brussels, Belgium, 2018.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, 2016.

Rajpurkar, P., Jia, R., and Liang, P. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pp. 784–789, 2018.

See, A., Liu, P. J., and Manning, C. D. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1073–1083, Vancouver, Canada, 2017.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

Zhao, Y., Ni, X., Ding, Y., and Ke, Q. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3901–3910, Brussels, Belgium, 2018.