

LREC-COLING 2024

AdaKron: an Adapter-based Parameter Efficient Model Tuning with Kronecker Product

Authors: **Braga Marco**, Raganato Alessandro, Pasi Gabriella



Overview

1. Introduction
2. Parameter Efficient Fine-Tuning
3. AdaKron
4. Experimental Setup
5. Results
6. Conclusion and Future Works



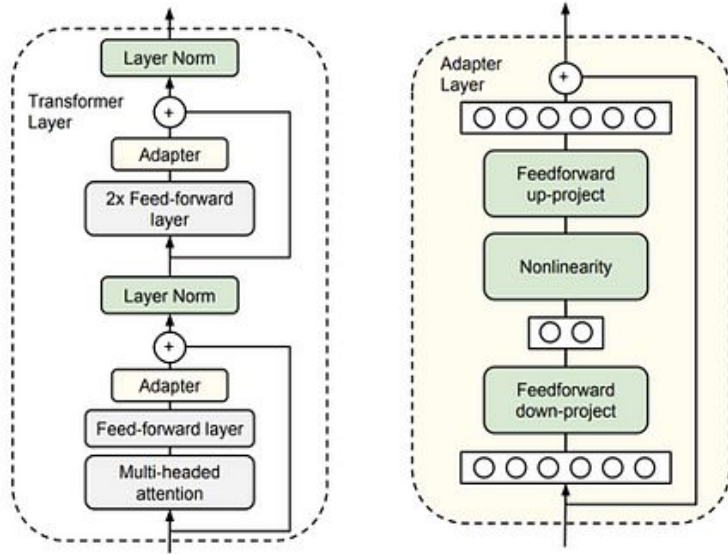
Introduction

The problem of **Fine-Tuning** Large Language Models (LLMs)

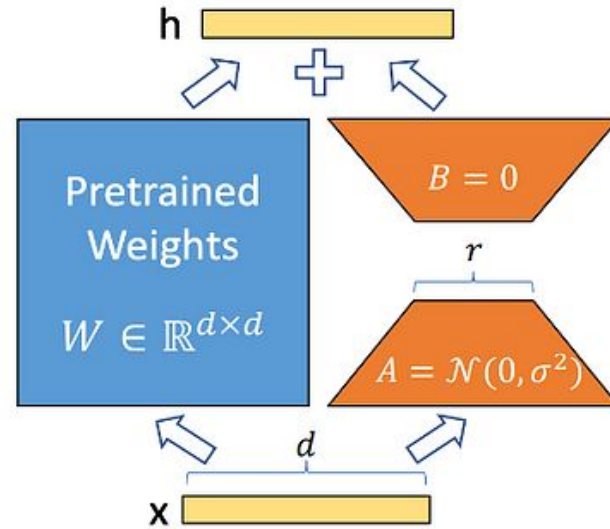
- Recent upsurge of LLMs, characterized by **billions** of parameters, has introduced profound computational **challenges** to the fine-tuning process.
- Intensive research on **Parameter-Efficient** Fine-Tuning (PEFT) techniques, usually involving the training of a selective **subset** of the original model parameters.



Parameter Efficient Fine-Tuning



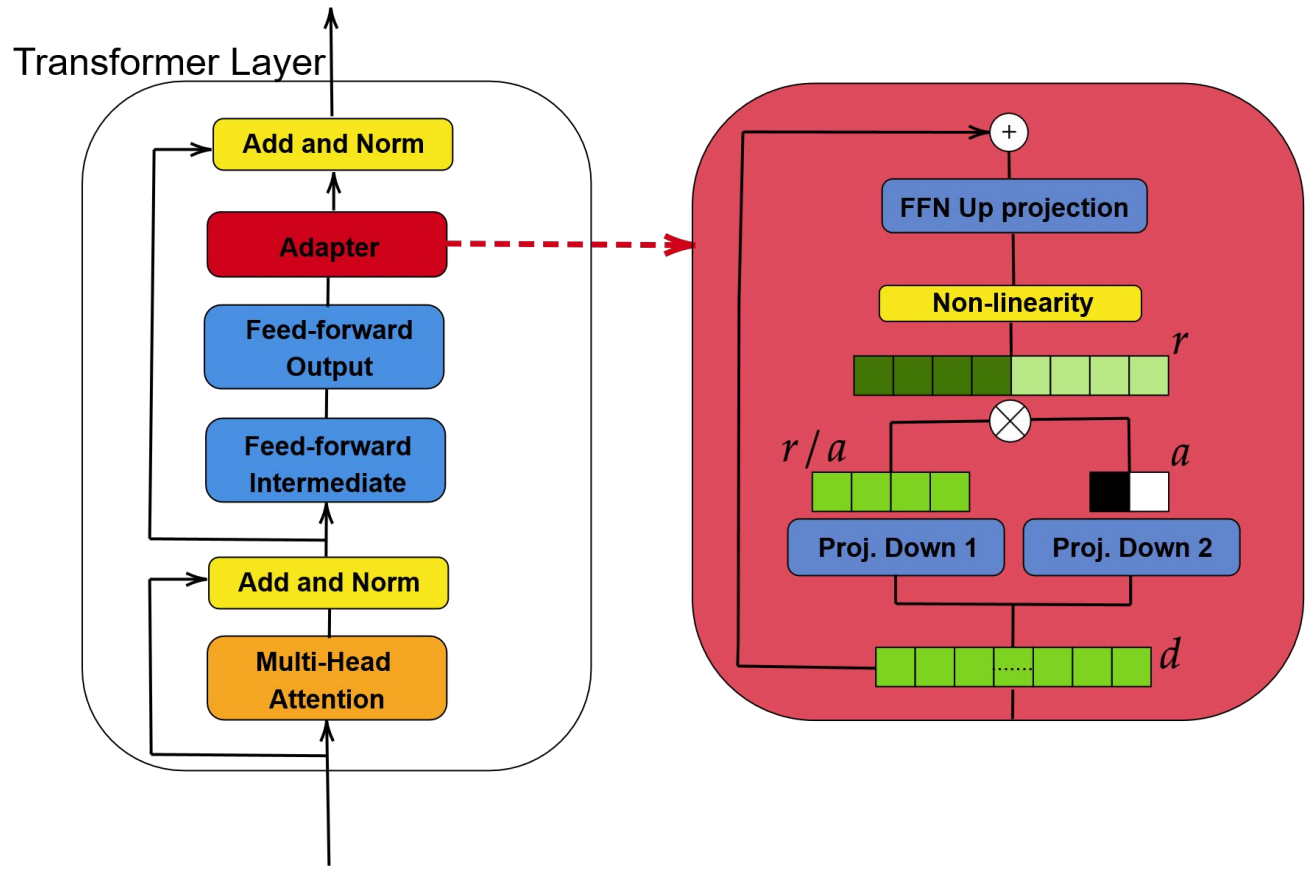
Houlsby Adapter



LoRA



Our Approach: AdaKron



AdaKron

- Employ the Kronecker product between output vectors of two feed-forward networks (FFNs), which compose the down projection of the Adapter.
- The output vector of the Kronecker product has a dimension equal to the product of the dimensions of the input vectors.
- Train **fewer parameters** in the down projection layer compared to a single layer.
- Example:
 - Let be d the dimension of the input vector, 48 be the intermediate dimension of the adapter and 4 the dimension of the second down projection. Therefore, AdaKron requires the training of

$$d*(48/4 + 4) = d*16$$



Experimental Setup

- We test against:
 - **Fine-Tuning**: it requires training all the parameters of a model
 - Houlsby and Pfeiffer **Adapters** [1,2]: they add few new parameters to a model and train only them
 - **Bit-Fit**, which trains only bias parameters of the model [3]
 - **LoRA** [4]: it requires training two low-rank matrices to update the attention parameters of a model
 - **AdaMix** [5]: combines Mixture of Experts (with random routing) and Adapters.
- **Datasets: GLUE** [6], composed of eight different Language Inference tasks



Results

Model	# Params (M)	MNLI Acc	QNLI Acc	SST2 Acc	QQP F1	MRPC F1	CoLa Mcc	RTE Acc	STS-B Pearson	Avg.
Fine-Tuning	110	83.2	90.0	91.6	87.4	90.9	62.1	66.4	89.8	82.7
Houlsby Adapter [†]	0.9	83.1	90.6	91.9	86.8	89.9	61.5	71.8	88.6	83.0
BitFit [◇]	0.1	81.4	90.2	92.1	84.0	90.4	58.8	72.3	89.2	82.3
LoRA [†]	0.3	82.5	89.9	91.5	86.0	90.0	60.5	71.5	85.7	82.2
AdaMix Adapter [△]	0.9*	84.7	91.5	92.4	87.6	92.4	62.9	74.7	89.9	84.5
AdaKron ₄₈	0.6	83.5	91.1	92.0	87.1	90.8	61.1	73.8	89.4	83.6
AdaKron ₃₂	0.4	83.7	90.9	92.2	87.1	89.5	60.7	74.1	89.5	83.5

Results on the **GLUE** development set with **BERT**-base

Params (M) refers to the number of update parameters (in Millions)



Results

Model	# Params (M)	MNLI Acc	QNLI Acc	SST2 Acc	QQP Acc	MRPC Acc	CoLa Mcc	RTE Acc	STS-B Pearson	Avg.
Fine-Tuning	355	90.2	94.7	96.4	92.2	90.9	68	86.6	92.4	88.9
Houlsby Adapter [†]	6	89.9	94.7	96.2	92.1	88.7	66.5	83.4	91	87.8
Houlsby Adapter [†]	0.8	90.3	94.7	96.3	91.5	87.7	66.3	72.9	91.5	86.4
Pfeiffer Adapter [†]	3	90.2	94.8	96.1	91.9	90.2	68.3	83.8	92.1	88.4
Pfeiffer Adapter [†]	0.8	90.5	94.8	96.6	91.7	89.7	67.8	80.1	91.9	87.9
LoRA [†]	0.8	90.6	94.8	96.2	91.6	90.2	68.2	85.2	92.3	88.6
AdaMix Adapter [△]	0.8*	90.9	95.4	97.1	89.8	94.1	70.2	89.2	92.4	89.9
AdaKron ₁₆	0.6	90.2	94.8	96.9	91	90.9	69.2	87.4	92.1	89.1
AdaKron ₃₂	1.0	90.2	94.4	96.1	87.8	93.1	69.9	86.1	92.2	88.7

Results on the **GLUE** development set with **RoBERTa**-Large

Params (M) refers to the number of update parameters (in Millions)



Conclusions and Future Work

- We manage to fine-tune **only 0.55%** of the original BERT parameters, while consistently achieving **competitive performance results** comparable to other state-of-the-art PEFT methods, even with larger parameter counts.
- As future work, we plan to improve our approach by incorporating it within a **Mixture of Experts** framework, extending our evaluation to different datasets and tasks.



Bibliography

1. N. Hounsby, A. Giurgiu, S. Jastrzebski, Bruna Morrone, A. Gesmundo, M. Attariyan, S. Gelly, "Parameter-efficient transfer learning for nlp.", International Conference on Machine Learning (2019)
2. Pfeiffer, Jonas, et al. "AdapterHub: A Framework for Adapting Transformers." EMNLP 20
3. Ben Zaken et al., "BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models" ACL 2022
4. Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." *ICLR 21*
5. Wang, Y., & Agarwal, S. (2022, January). AdaMix: Mixture-of-Adaptations for Parameter-efficient Model Tuning. In *EMNLP 22*
6. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018, September). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*.

