AdaKron

an Adapter-based Parameter Efficient Model Tuning with Kronecker Product

m.braga@campus.unimib.it
marco.braga@polito.it

Marco Braga*°, Alessandro Raganato*, Gabriella Pasi*

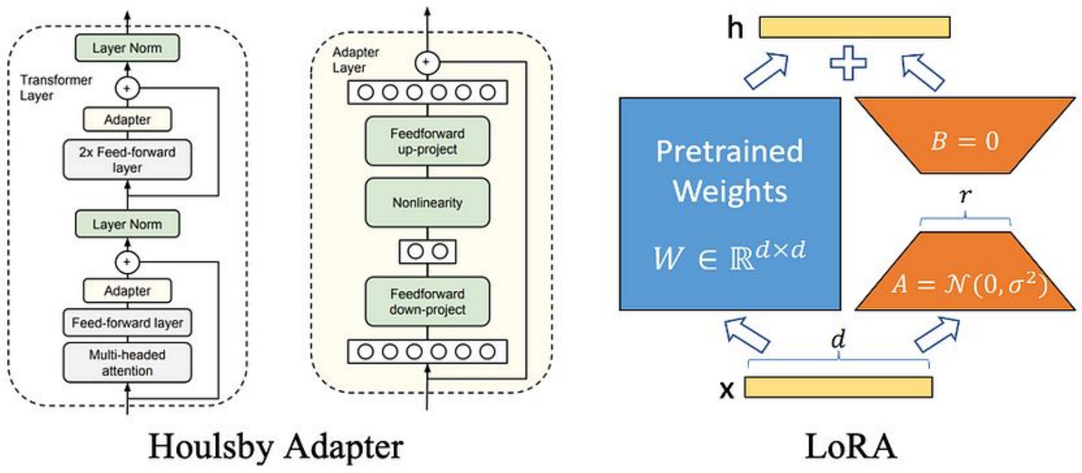
*University of Milano-Bicocca
°Politecnico di Torino

LREC-COLING 2024

Problem

- The conventional approach to **fine-tuning** for downstream tasks requires the training of all parameters of a neural model. However, with the recent increase of Large Pretrained Language Models reaching **billions** of parameters, the traditional fine-tuning process has become **challenging** due to large memory requirements.

- **Parameter-efficient** Fine-Tuning Techniques have emerged as a new paradigm. These methods allow us to train only a **fraction** of the original model parameters, while fine-tuning the model to a specific task and **keeping performance** levels **comparable** to traditional fine-tuning.

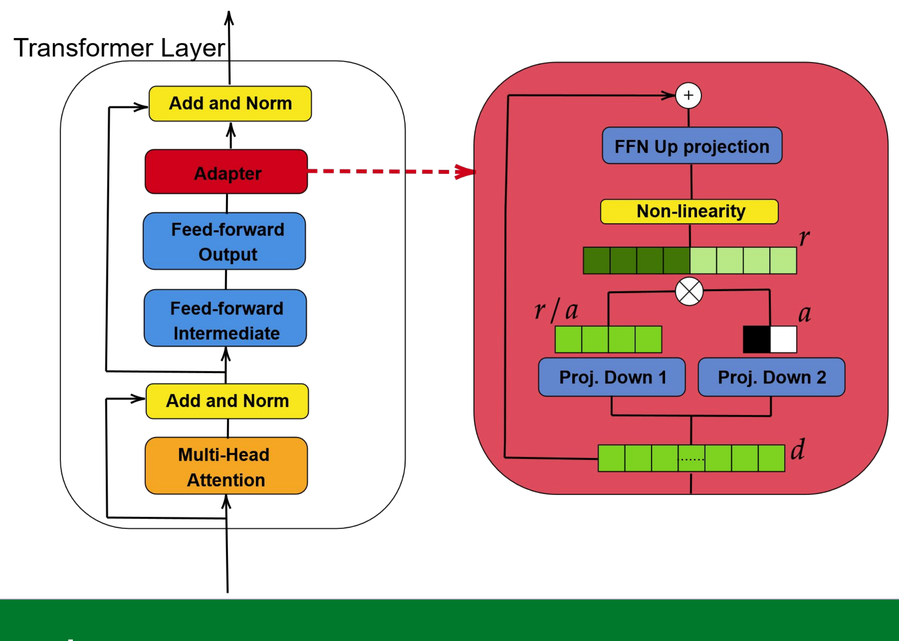


Proposed Approach: **AdaKron**

- The capability of an Adapter depends on its intermediate dimension, and recent empirical studies [1] suggest that **low-dimensional** Adapter modules can give better performances than high ones.

- **AdaKron** combines Adapters [2] with **Kronecker** Product [3] in a new and efficient way. AdaKron is composed of two Down Projection layers, whose outputs are then multiplied through Kronecker product.

- By training only **0.55%** of parameters, we reach performance on par with recent state-of-the-art PEFT methods that require more parameters to train.



Evaluation and Discussion

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
Prefix-tuning†	0.2	81.2	90.4	90.9	83.3	91.3	55.4	76.9	87.2	82.1
LoRA†	0.3	82.5	89.9	91.5	86.0	90.0	60.5	71.5	85.7	82.2
UNIPELT (AP)†	1.1	83.4	90.8	91.9	86.7	90.3	61.2	71.8	88.9	83.1
UNIPELT (APL)†	1.4	83.9	90.5	91.5	85.5	90.2	58.6	73.7	88.9	83.5
AdaMix Adapter△	0.9*	84.7	91.5	92.4	87.6	92.4	62.9	74.7	89.9	84.5
Pfeiffer Adapter$_{48}$	0.9	83.7	91.0	92.0	87.4	90.9	60.9	68.2	89.5	82.9
Pfeiffer Adapter$_{32}$	0.6	84.0	91.2	92.3	87.3	90.0	58.6	71.8	89.3	83.1
AdaKron$_{48}$	0.6	83.7	91.8	92.1	87.5	91.1	61.1	73.3	89.7	83.8
AdaKron$_{32}$	0.4	83.6	90.9	92.3	87.2	89.5	61.2	75.6	89.6	83.7

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

- We compare our method with different PEFT techinques:
  - **Houlsby and Pfeiffer Adapter**: requires defining and training new modules, each of one composed by two linear layers;
  - **BitFit**: requires training only the bias parameters of the model;
  - **Prefix-Tuning**: prepends a sequence of continuous task-specific vectors to the input, which are the only trained parameters;
  - **LoRa**: requires training two low-rank matrices to upgrade attention weights;
  - **UNIPELT**: is a combination of the previous methods;
  - **AdaMix**: combines Mixture of Experts (with random routing) and Adapters.

- AdaKron shows on average **better performance** compared to the full Fine-Tuning, and Houlsby Adapter. Moreover, AdaKron achieves an average one-point **improvement** over smaller PEFT methods like BitFit, Prefix-tuning, and LoRA. Interestingly, our approach also achieves **better performance** than UNIPELT, which uses twice the amount of parameters compared to AdaKron.

Future Improvement

- Improve our approach by incorporating it within a **Mixture of Experts** framework [3] with a linguistic/task-based gating function.

- Inject **user-related information** into the Adapter to define user-specific experts.

- Extending evaluation to Language **Generation** and Multilingual tasks.

References

[1] Chen, Guanzheng, et al. "Revisiting parameter-efficient tuning: Are we really there yet?." *arXiv preprint arXiv:2202.07962* (2022).
[2] Houlsby, Neil, et al. "Parameter-efficient transfer learning for NLP." *International Conference on Machine Learning*. PMLR, 2019.
[3] Henderson, Harold V., Friedrich Pukelsheim, and Shayle R. Searle. "On the history of the Kronecker product." *Linear and Multilinear Algebra* 14.2 (1983): 113-120.
[4] Chung, Hyung Won, et al. "Scaling instruction-finetuned language models." *arXiv preprint arXiv:2210.11416* (2022).
[5] Asghar, Nabiha. "Yelp dataset challenge: Review rating prediction." *arXiv preprint arXiv:1605.05362* (2016).
[6] Sclar, Melanie, et al. "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting." *arXiv preprint arXiv:2310.11324* (2023).

For More Information

Marco Braga