

Topic 4: Large Model Adaptation

1. Low-Rank Adaptation paper: <https://arxiv.org/abs/2106.09685>
2. Explains LoRA for large language models:
https://huggingface.co/docs/peft/main/en/conceptual_guides/lora
3. Blog post survey of PEFT methods, including adapters, soft-prompts, prefix tuning, LoRA; good mid-level overview:
<https://huggingface.co/blog/samuellimabraz/peft-methods>
4. How LoRA reduces compute/memory cost:
<https://heidloff.net/article/efficient-fine-tuning-lora/>
5. Describing adapter modules and showing how they enable multi-task learning without catastrophic forgetting:
<https://research.google/pubs/parameter-efficient-transfer-learning-for-nlp/>
6. Community article explaining how to use the PEFT library, walking through setting up a base model and then adding LoRA adapters:
<https://huggingface.co/blog/ariG23498/workflow-peft>
7. The code implementation of parameter-efficient fine-tuning methods:
<https://github.com/huggingface/peft>
8. Introduces adapter modules:
<https://medium.com/dair-ai/adapters-a-compact-and-extensible-transfer-learning-method-for-nlp-6d18c2399f62>