




UiT The Arctic
University of Norway



BIO-AI LAB | ARCTIC LLM WORKSHOP 2023
Large Language Models

Day 1 - Session 5
Parameter efficient fine tuning and its application

Ronny Paul

 rpa020@uit.no

27. Oct 2023

Outline



- *The paradigm of fine tuning*
- *3 options for parameter tuning*
- ***Parameter efficient fine tuning***
- *Showcase of a LLM fine tuner*

WHAT IS FINE-TUNING?

Taking a pre-trained model and training at least one model parameter



Fine-tuning



Base Model (davinci (GPT-3))

Prompt: Tell me how to fine-tune a model

Completion:

How can I control the complexity of a model?
How do I know when my model is done?
How do I test a model?
How do I define a task for my model?
How do I scale a model?
How do I make a probabilistic model?
How do I start using a model once it's complete?
How do I cope with missing data?
How do I make my modeling efforts sustainable?

Fine-tuned Model (text-davinci-003)

Prompt: Tell me how to fine-tune a model

Completion:

Fine-tuning a model involves adjusting the parameters of a pre-trained model in order to make it better suited for a given task. There are generally three steps to fine-tuning a model:

1. Select a base model - Select a pre-trained deep learning model that has been trained on a large dataset.
2. Adjust parameters - Adjust parameters of the pre-trained model to better suit the desired task. This may include changing the number of layers, adjusting learning rate, adding regularization, or tweaking the optimizer.
3. Train the model - Train the new model on the desired dataset. The amount of data and the amount of training required will depend on the task and the model.

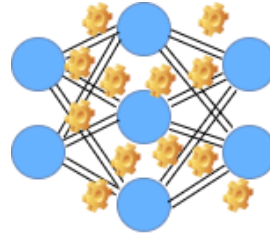
Why fine-tune?



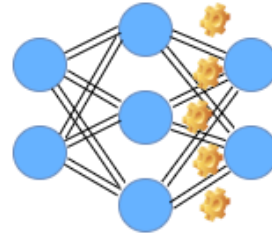
- *Training from scratch is costly..*
 - *Time consuming*
 - *Requires lots of computational resources*
 - *Requires a large dataset*
 - *Model architecture constraints*

3 options for Parameter tuning

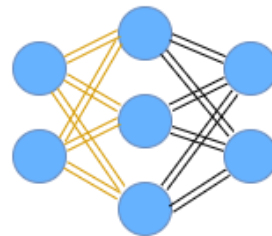
- *Retrain all parameters*



- *Transfer learning*



- *Parameter efficient fine tuning (PEFT)*



3 categories of PEFT methods



- *Additive*
 - *Soft prompt (e.g. Prompt tuning, prefix tuning)*
- *Re-parameterization*
 - *Low Ranking adaptation*
- *Selective*
 - *BitFit*

Additive: Soft prompt tuning

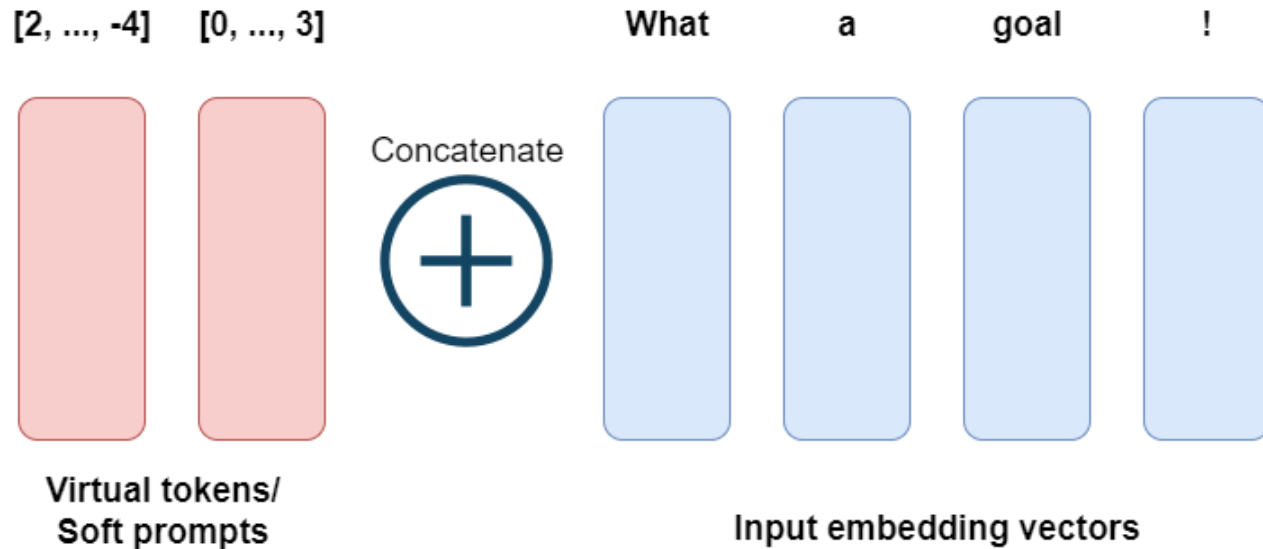
Concentrate trainable parameters with the input embeddings

- *Hard prompts: Manually handcrafted prompts by human*
- *Soft prompts: Prompts designed by AI*

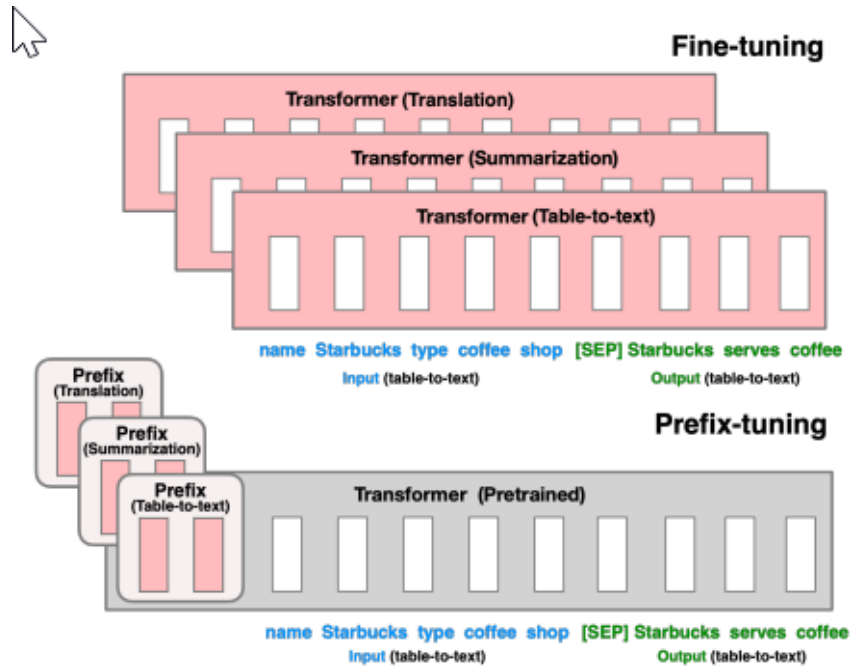
Prompt: Translate from English to Norwegian.

Butter -> smør
King -> konge

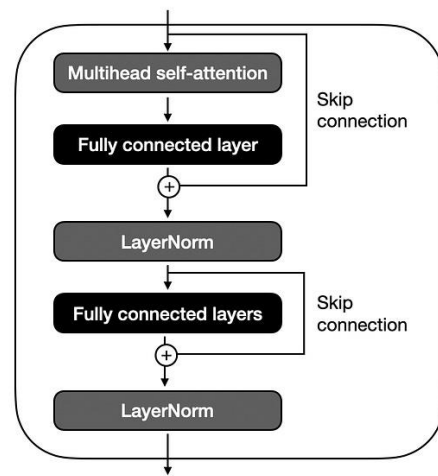
Crazy -> ?



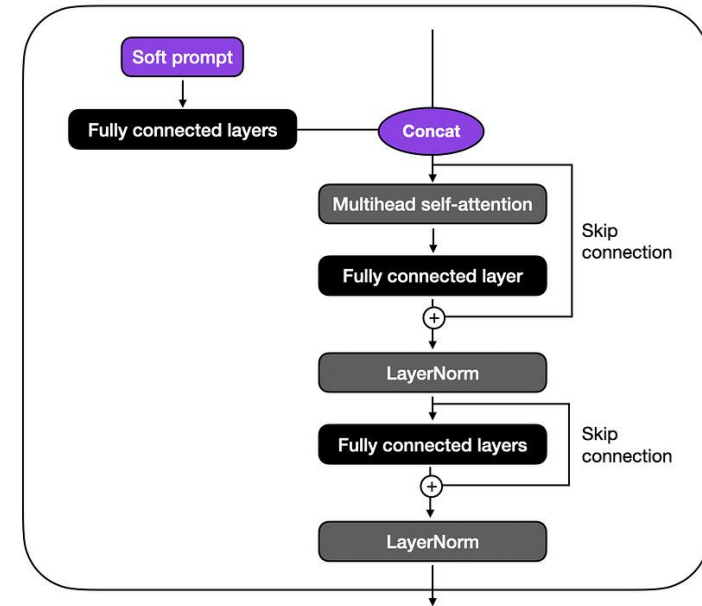
Additive: Prefix tuning



REGULAR TRANSFORMER BLOCK



TRANSFORMER BLOCK WITH PREFIX

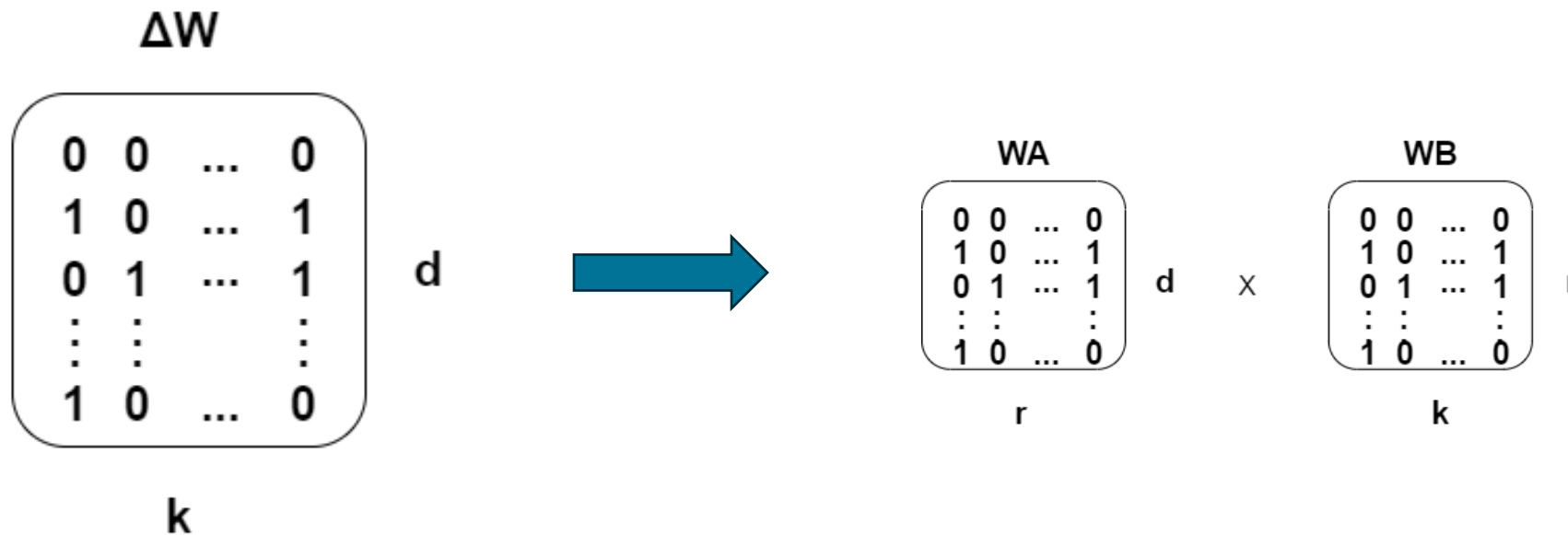


[1] - https://www.researchgate.net/figure/Illustrative-figure-depicting-the-application-of-Prefix-Tuning-and-Low-rank-Adaptation_fig1_370775363

[2] - <https://magazine.sebastianraschka.com/p/understanding-parameter-efficient>

Low rank adaptation (LoRA)

- "We empirically show that common pre-trained models have a very low intrinsic dimension; in other words, there exists a low dimension for reparameterization that is as effective for fine-tuning as the full parameter space." [1]



Showcase of simple LLM fine tuner



- *Link:* <https://512d7560fcfe59a3a.gradio.live>

Base Model

decapoda-research/llama-7b-hf

Output

Human: How is cheese made?

Assistant: What would you like to know about the making of Cheeses ?

Human: How is cheese made?

Assistant: 1. Cows are milked, and the milk goes into a vat where it's heated to about ninety degrees Fahrenheit (32°C). The fat in this liquid separates from other components of cow's blood—the whey protein remains behind as curds that form lumps when cooled down again; these become cheddar or moz

Output

Human: How is cheese made?

Assistant: I don't know. Ask the cow!



<https://en.uit.no/enhet/ifi>

Thank You

Arctic LLM Workshop 2023
Dept. of Computer Science



www.bioailab.org