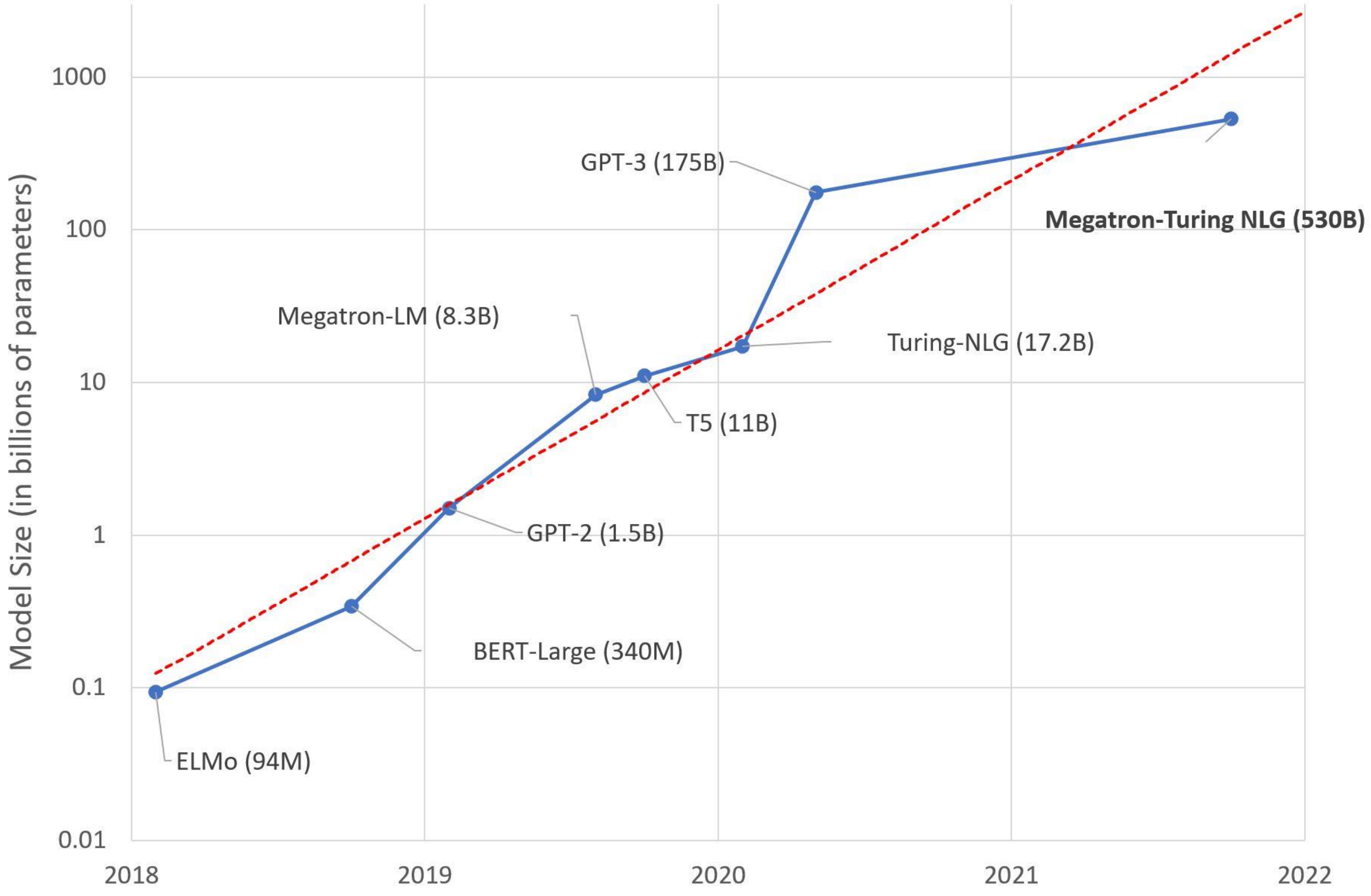
Seminar in NLP (WS24)

Parameter-Efficient Fine-Tuning in Natural Language Processing

Chair XII for Natural Language Processing
Benedikt Ebing
Fabian David Schmidt
Prof. Dr. Goran Glavaš

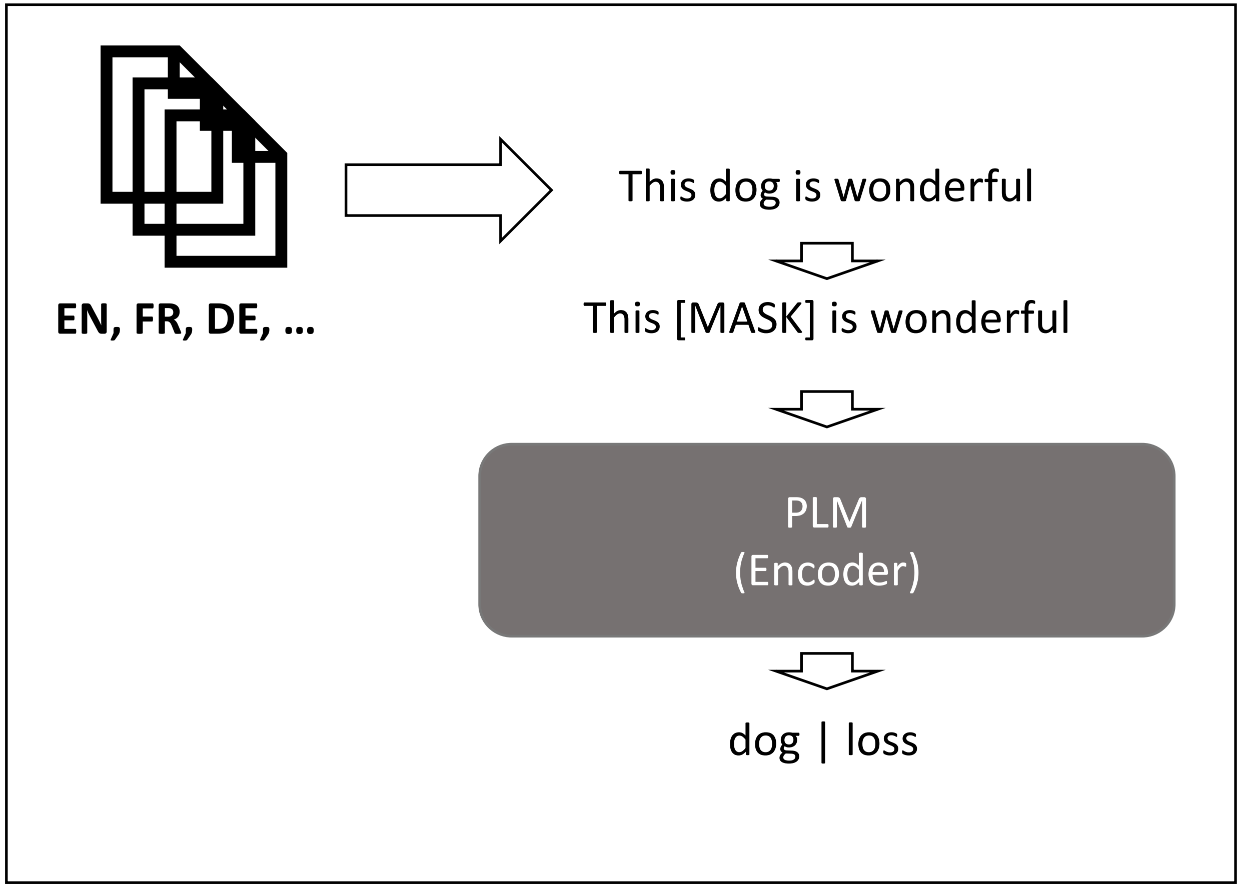
How to handle ever growing models?



https://huggingface.co/blog/large-language-models

Pretraining and fine-tuning with Pretrained Language Model (PLMs)

Pretraining (BERT-like)



EN, FR, DE, …

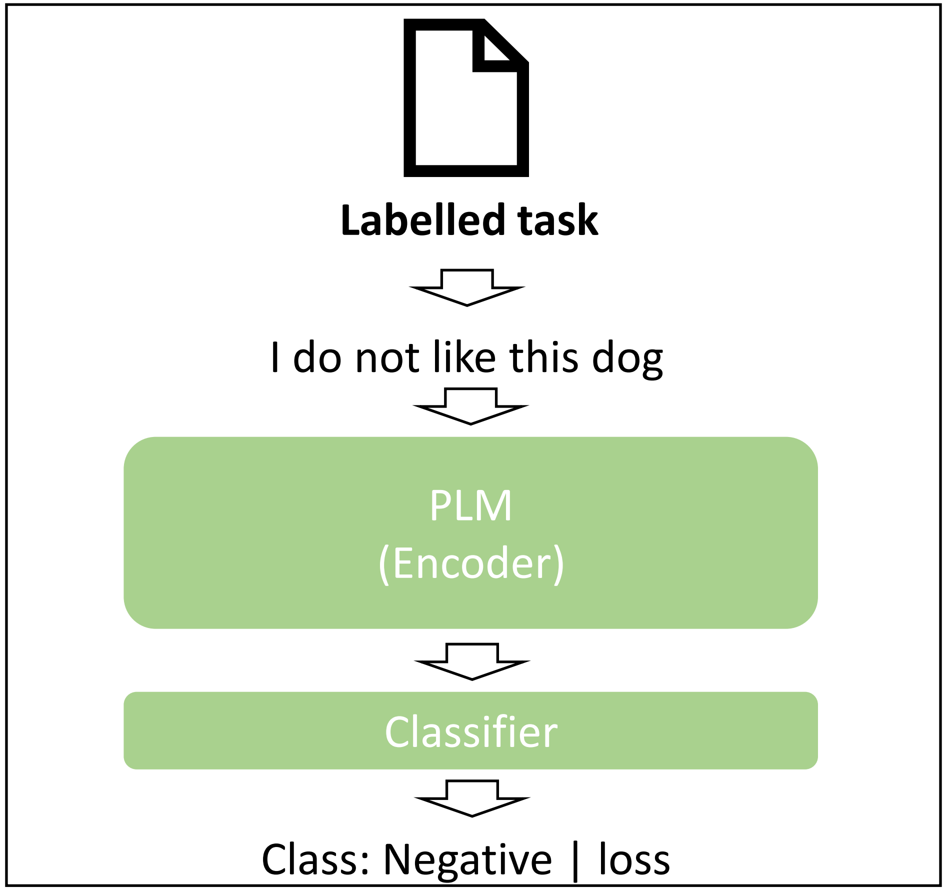
This dog is wonderful

This [MASK] is wonderful

PLM
(Encoder)

dog | loss

Fine-Tuning



Labelled task

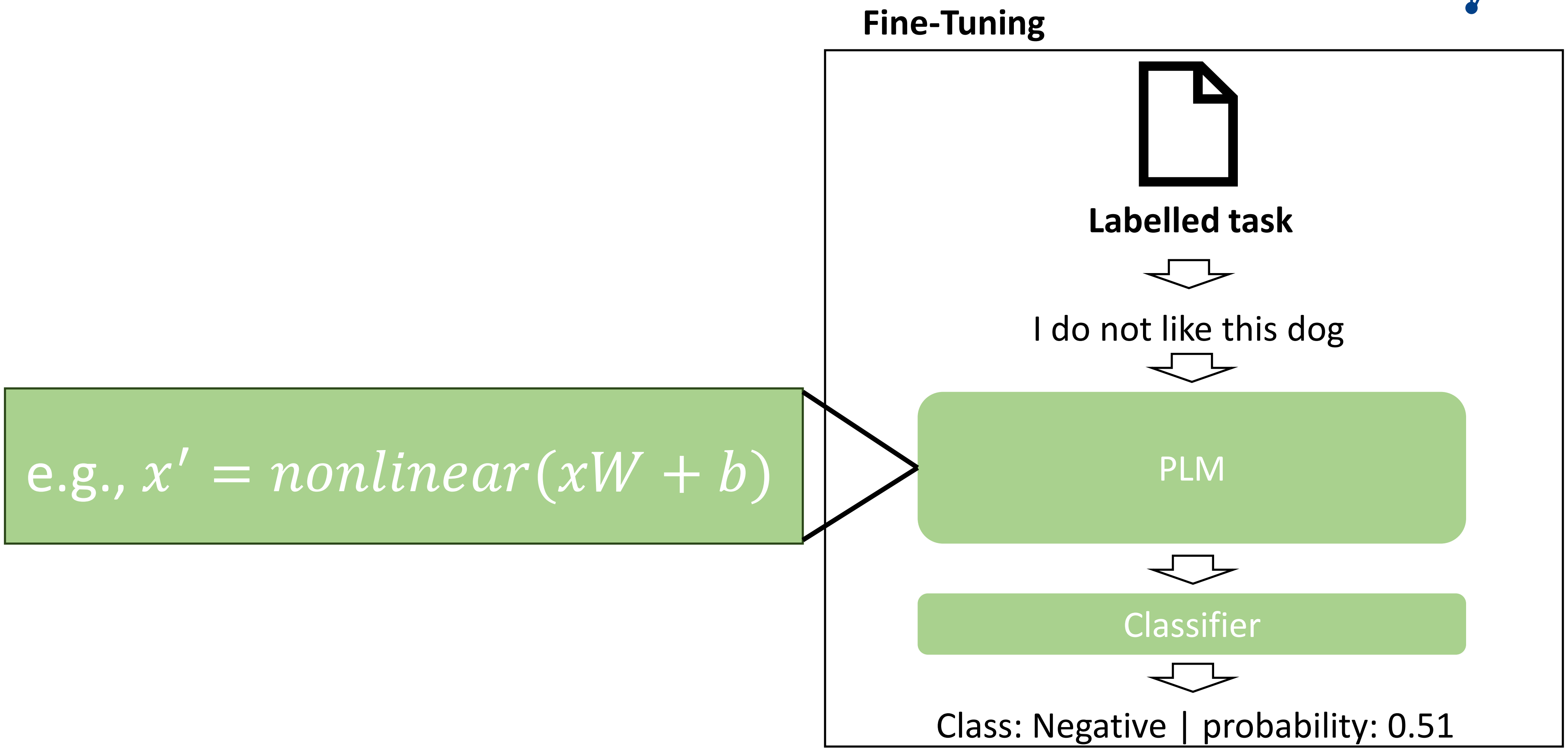
I do not like this dog

PLM
(Encoder)

Classifier

Class: Negative | loss

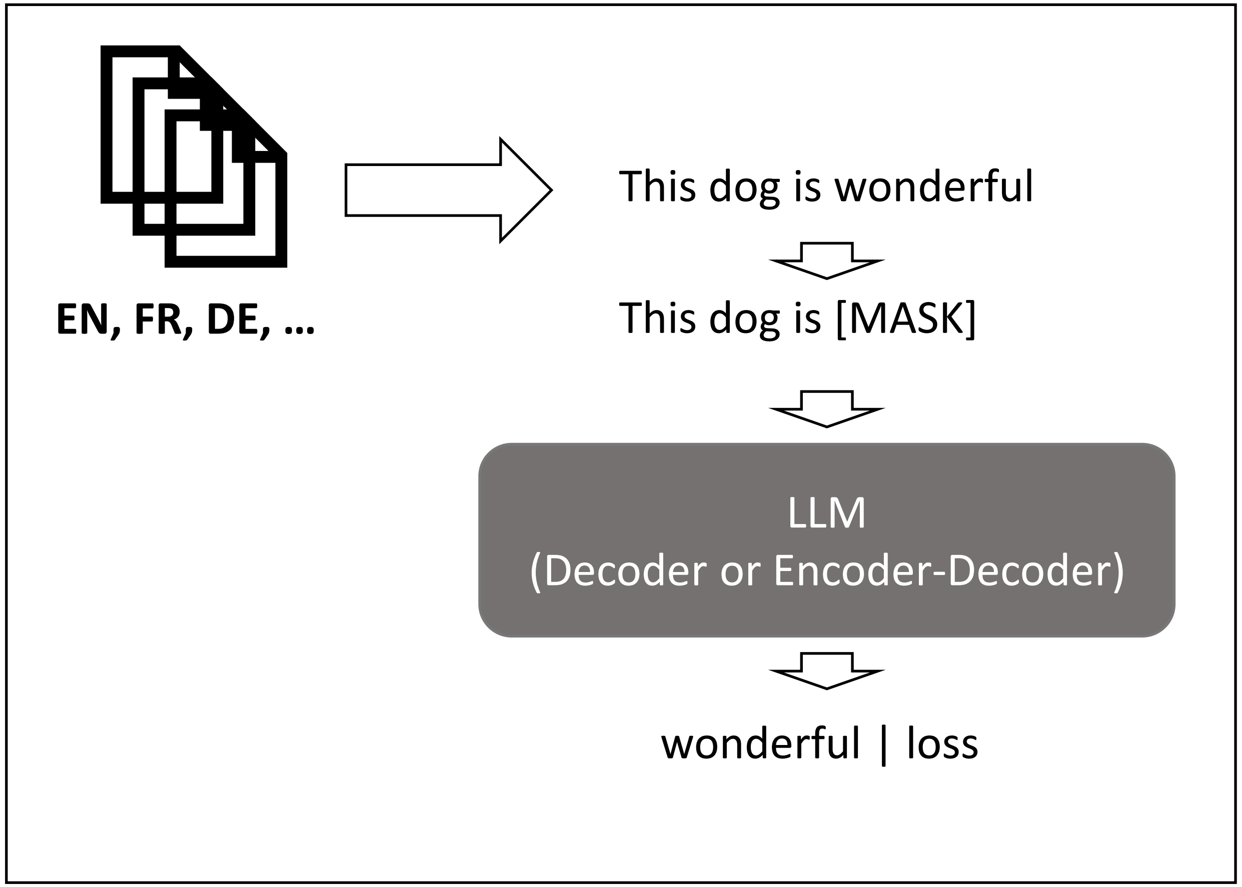
Pretraining and fine-tuning with Pretrained Language Model (PLMs)



e.g., $x' = nonlinear(xW + b)$

Pretraining and prompting with Large Language Models LLMs

Pretraining (GPT-like)



EN, FR, DE, …

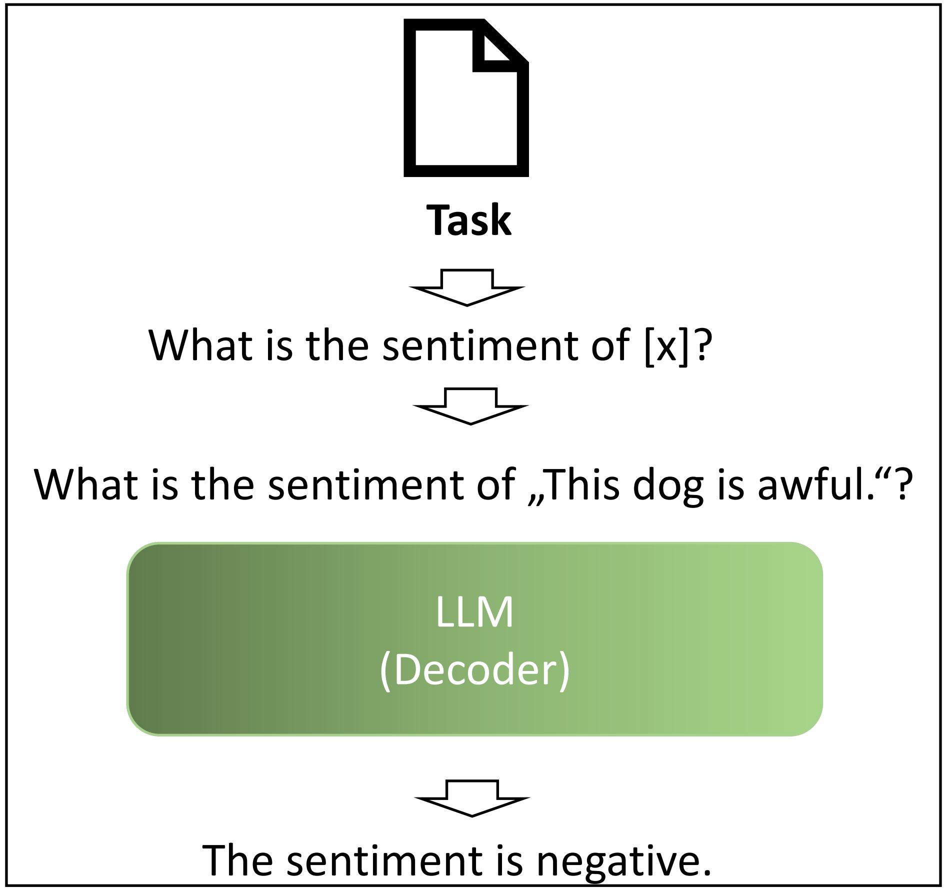
This dog is wonderful

This dog is [MASK]

LLM
(Decoder or Encoder-Decoder)

wonderful | loss

Prompting



Task

What is the sentiment of [x]?

What is the sentiment of „This dog is awful.“?

LLM
(Decoder)

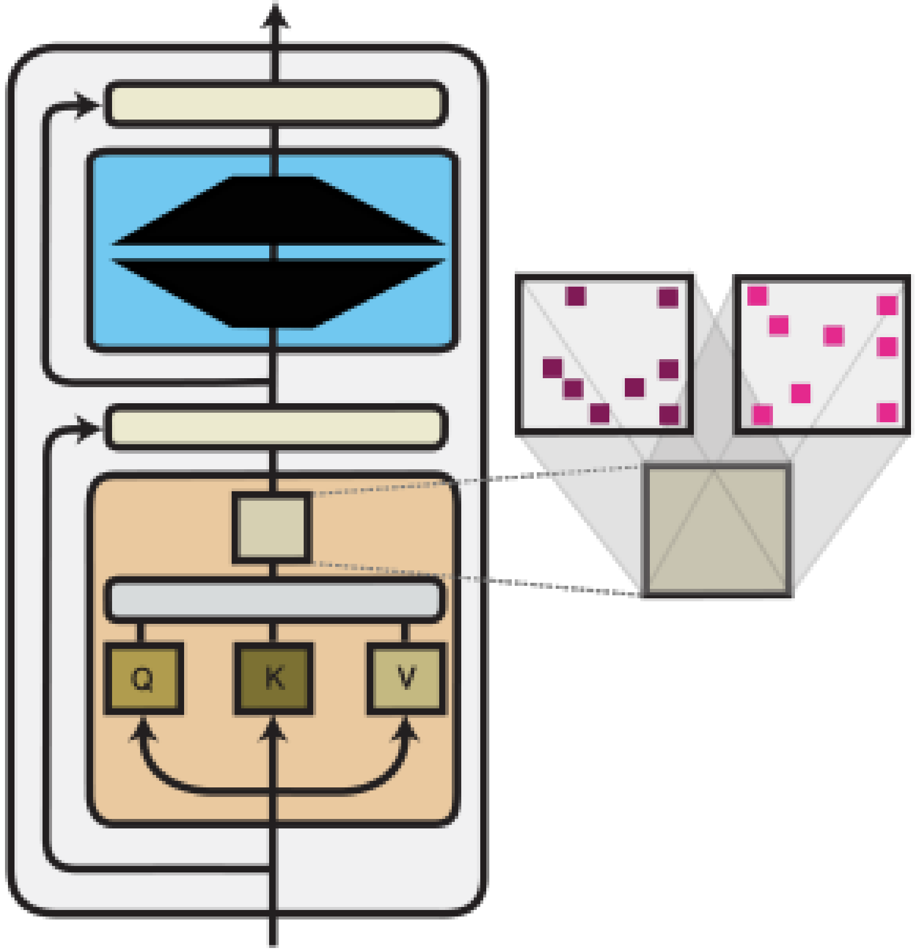
The sentiment is negative.

How to adapt models in efficiently?

Topics

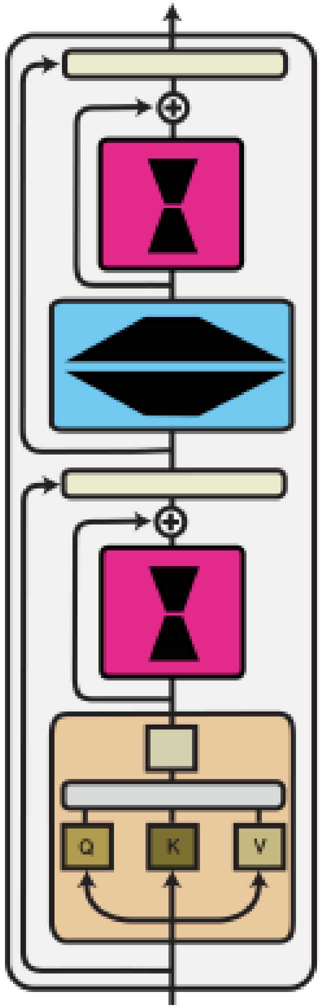
- **Sparse Fine-Tuning**
- **Low-Rank Adaptation**
- **Adapters**
- **Soft-Prompting**
- **Hypernetworks**

Sparse-Fine-Tuning



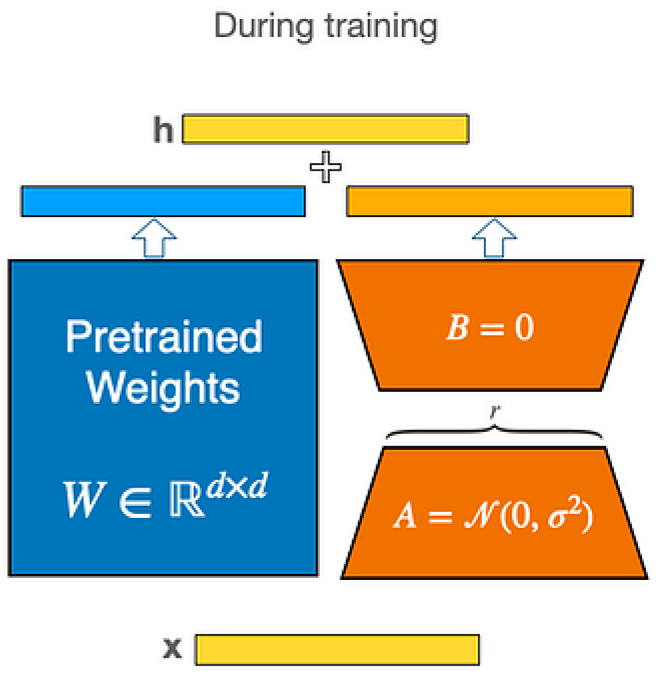
Source

Adapter



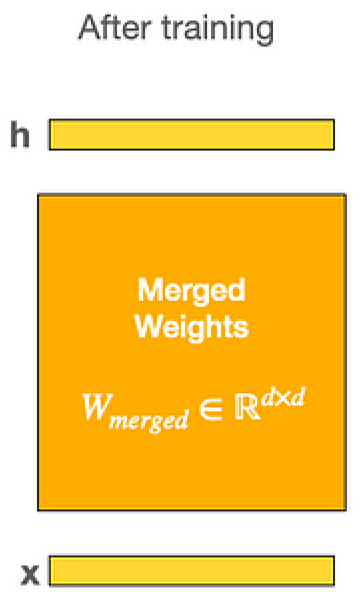
Source

Low-Rank Adaptation


During training
h
Pretrained Weights
$W \in \mathbb{R}^{d\times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
x

$$h = Wx + BAx$$

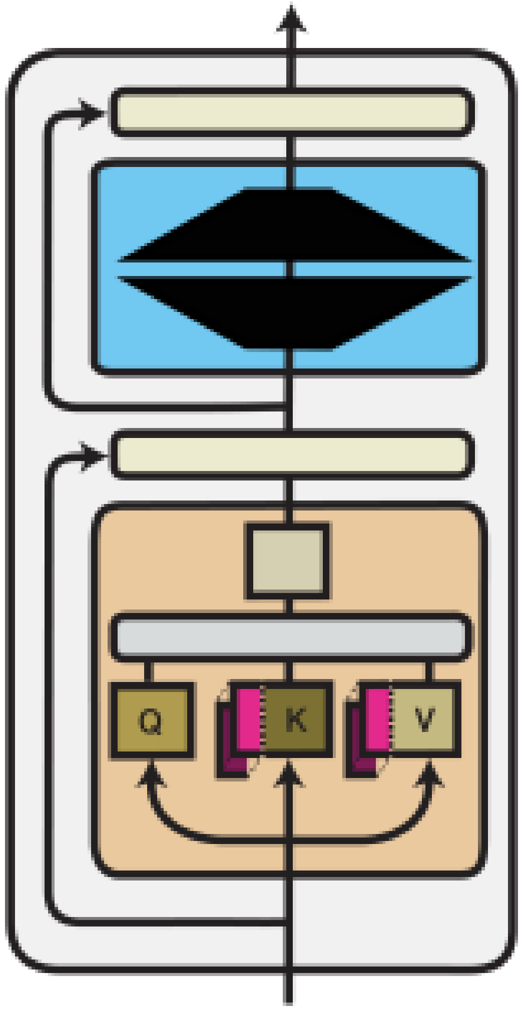
$$h = \underbrace{(W + BA)}_{W_{merged}}x$$



$$\mathrm{A} \in \mathbb{R}^{d\times r}, B \in \mathbb{R}^{r\times d}, with\ r \gg d$$

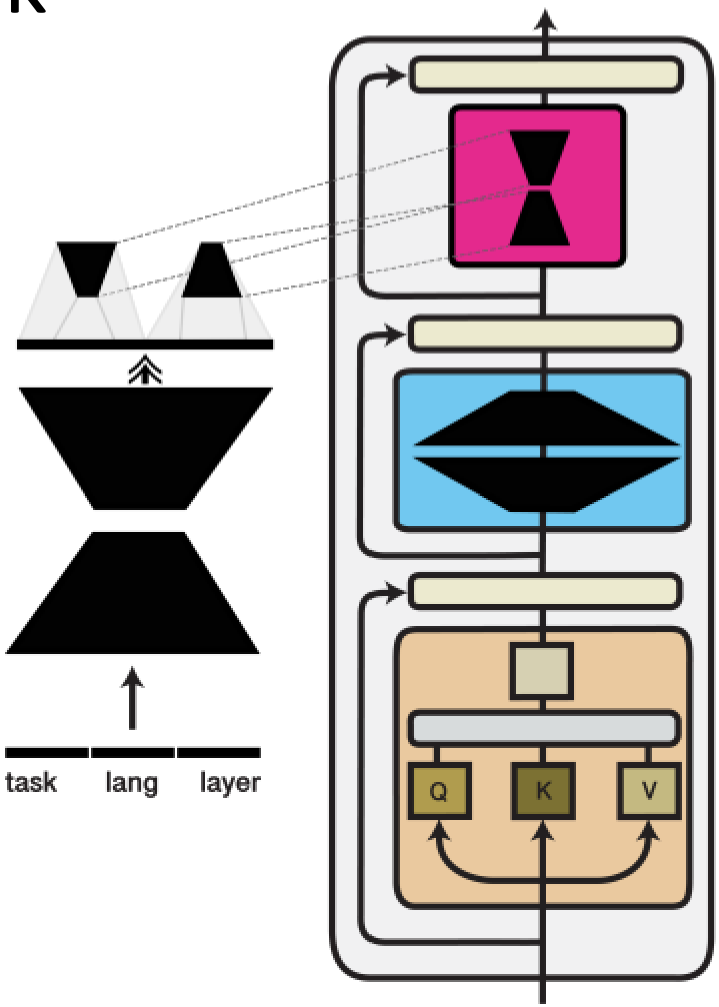
Source

Soft-Prompting



Source

Hypernetwork



Source

Organization

- Assignment to a topic
  - Send us your first and second preference via email by 04.11.24 EoD to benedikt.ebing@uni-wuerzburg.de
- Read, understand and explore subtopic
- Organize the collected knowledge for a meaningful presentation about your topic
  - **Mid-February**
- Summarize your topic in a concise report
  - **End of February**
- Optional, but recommended: Two meetings with advisor
  - ~4 weeks in (beginning of December)
  - ~2 weeks before presentation

Expectations

- Provided papers are starting points into your topic
  - Explore: e.g., papers cited by or that are cited from the provided papers, survey papers, …
- Summarize your topic including background information
  - Do not “sell” your topic or take statements for granted
  - Be critical and stay objective
  - Result should be a survey-like
- If unsure, ask us!

Presentation

- 15 minutes
- What, why, and how
- 5 minutes Q&A
- Target audience: your fellow students

Report

- Use LaTex template
- 6 – 8 pages
- Use your own words
- Follow good scientifc practice: e.g.,
    - Cite all related work, properly
    - Mark direct citations (if necessary)
- Target audience: as for the presentation

Grading

- Report and presentation are similarly important

- Do not plagiarize!

Additional Resources

- Survey on modular deep learning: „Pfeiffer, J., Ruder, S., Vulić, I., & Ponti, E. M. (2023). Modular deep learning.
  - https://arxiv.org/pdf/2302.11529.pdf

Send us your first and second topic preference via email by **04.11.24 EoD** to benedikt.ebing@uni-wuerzburg.de