

# Understanding and Training Language Models: **Task-Specific Models**

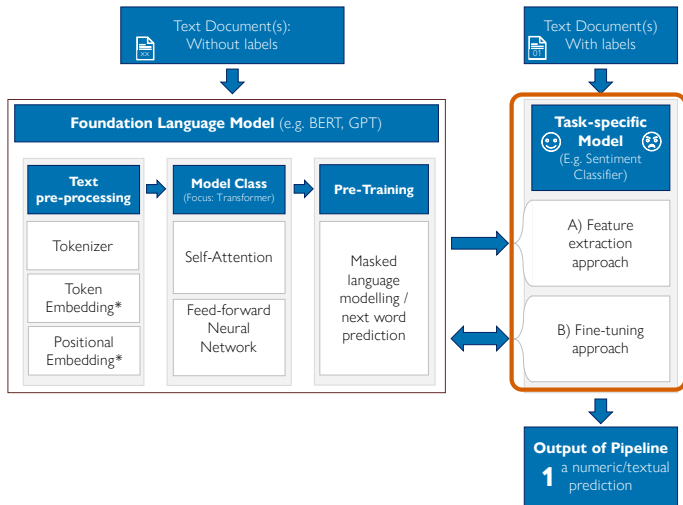
Erik-Jan Senn

Faculty of Mathematics and Statistics, University of St. Gallen

---

CSH Autumn School at University of Hohenheim  
September/October 2024

# Where are we?



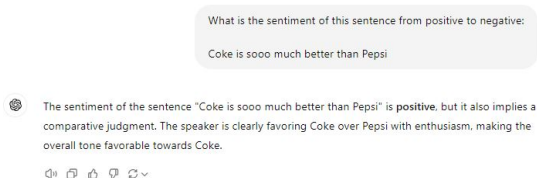
Language Modelling Pipeline

# Sentiment Analysis as Example Application

- ▶ Typically a **sequence classification** task  $f : X \mapsto Y$  with  $X \in \mathbb{R}^{l \times v}$  and  $Y \in \{1, 2, \dots, k\}$  with  $k$  classes.
- ▶ Training a *good* classifier from scratch is difficult:
  - ▶ Too **few noisy labels** are available.
  - ▶ Sentiment analysis is a **complex task** that requires contextual understanding.

# Prompting for Sentiment Analysis

- ▶ Provide an instruction and the text as prompt (context) to a generative model.
- ▶ The LM does iterative next word prediction to create a response.
- ▶ Evaluate the generated text.



Prompt to *Chat-GPT-4* <https://chatgpt.com/>.

# Transfer learning for Sentiment Analysis

**Transfer learning** is the idea of improving on something by using knowledge from a different domain.

# Transfer learning for Sentiment Analysis

**Transfer learning** is the idea of improving on something by using knowledge from a different domain.

For LMs, we can **transfer language understanding** of **pretrained LMs** to help for **specific tasks** (even if these are not masked word prediction).

# Transfer learning for Sentiment Analysis

**Transfer learning** is the idea of improving on something by using knowledge from a different domain.

For LMs, we can **transfer language understanding** of **pretrained LMs** to help for **specific tasks** (even if these are not masked word prediction).

**Simplification** of the task-specific model:

- ▶ **Dimensionality reduction**: Replace the one-hot-encoded sequence  $X \in \mathbb{R}^{l \times v}$  by the trained **hidden representation**  $Z \in \mathbb{R}^{l \times d}$  of the LM.
- ▶ **Language understanding**:  $Z$  should reflect the **general semantic and contextual understanding** of texts, *hopefully* making it a better predictor.

# Transfer learning for Sentiment Analysis

**Transfer learning** is the idea of improving on something by using knowledge from a different domain.

For LMs, we can **transfer language understanding** of **pretrained LMs** to help for **specific tasks** (even if these are not masked word prediction).

**Simplification** of the task-specific model:

- ▶ **Dimensionality reduction**: Replace the one-hot-encoded sequence  $X \in \mathbb{R}^{l \times v}$  by the trained **hidden representation**  $Z \in \mathbb{R}^{l \times d}$  of the LM.
- ▶ **Language understanding**:  $Z$  should reflect the **general semantic and contextual understanding** of texts, *hopefully* making it a better predictor.

**Approach**: We will use (**extract**) or slightly adapt (**fine-tune**) the **hidden representation**  $Z$  and estimate a classification model using these (better) **inputs**.



# Feature Extraction

Standard supervised learning setting as in Lecture 2.

- ▶ Find a function  $f : Z \rightarrow Y$  from a class  $f \in \mathcal{H}$  that minimizes a loss  $\mathcal{L}$ .
- ▶ We use any learning algorithm, e.g. OLS, gradient descent
- ▶ Everything learned there can be applied.

## Fine-tuning the LM

**Goal:** Improve performance on the downstream task compared to simple feature extraction.

**Fine-tuning** is the process of **specializing the LM hidden states  $\mathbf{Z}$**  to a specific downstream task by continuing to train on task-specific labelled data.

## Fine-tuning the LM

**Goal:** Improve performance on the downstream task compared to simple feature extraction.

**Fine-tuning** is the process of **specializing the LM hidden states  $\mathbf{Z}$**  to a specific downstream task by continuing to train on task-specific labelled data.

Approaches:

- ▶ Train all model parameters of classifier and LM jointly.
- ▶ Train the classifier and only few parameters for the LM.
  - ▶ Only train a subset of parameters of the LM, keeping the others **frozen** (fixed). E.g. only train the classifier and the last two transformer blocks (potentially with a lower learning rate).
  - ▶ Low-Rank Adaptation introduces a new low-dimensional parameter set to train instead of optimizing all LM parameters (Hu et al., 2021).

k

## Fine-tuning the LM

**Goal:** Improve performance on the downstream task compared to simple feature extraction.

**Fine-tuning** is the process of **specializing the LM hidden states  $\mathbf{Z}$**  to a specific downstream task by continuing to train on task-specific labelled data.

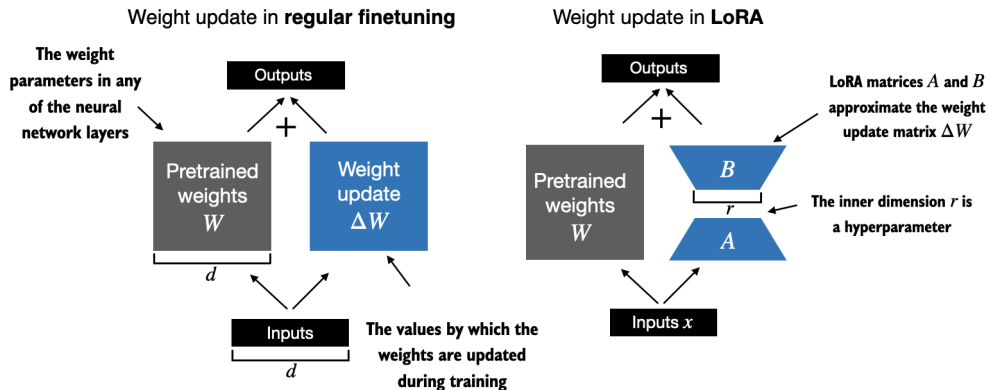
Approaches:

- ▶ Train all model parameters of classifier and LM jointly.
- ▶ Train the classifier and only few parameters for the LM.
  - ▶ Only train a subset of parameters of the LM, keeping the others **frozen** (fixed). E.g. only train the classifier and the last two transformer blocks (potentially with a lower learning rate).
  - ▶ Low-Rank Adaptation introduces a new low-dimensional parameter set to train instead of optimizing all LM parameters (Hu et al., 2021).

k

*Note:* There are also domain specific LMs trained on masked language modelling, e.g. *FinBERT* (Huang et al., 2023) for financial texts.

# Fine-tuning the LM - LoRA



Low Rank Adaptation from [Hu et al. \(2021\)](#), [Source](#)

## An important detail: What is $\mathbf{Z}$ exactly?

$\mathbf{Z}$  should be the **last low-dimensional hidden representation**  $\mathbf{Z} \in \mathcal{R}^{l \times d}$  of the LM, which should contain the most contextualized information.

- ▶ If the LM comes with the token classifier which maps to  $l \times v$  (e.g. in GPT), make sure to extract the hidden representation of the previous layer
- ▶ For sequence classification, a causal mask is not needed.
- ▶ Often practice, the dimensionality  $l$  of  $\mathbf{Z}$  is collapsed to 1, e.g. by taking a mean over  $l$ , or by taking the first (BERT [CLS] token) or the last (GPT-2 <endoftext> token).

# Questions ?

## References

- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL <https://arxiv.org/abs/2106.09685>.
- Allen H. Huang, Hui Wang, and Yi Yang. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, 2023. doi: <https://doi.org/10.1111/1911-3846.12832>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/1911-3846.12832>.