BERT:

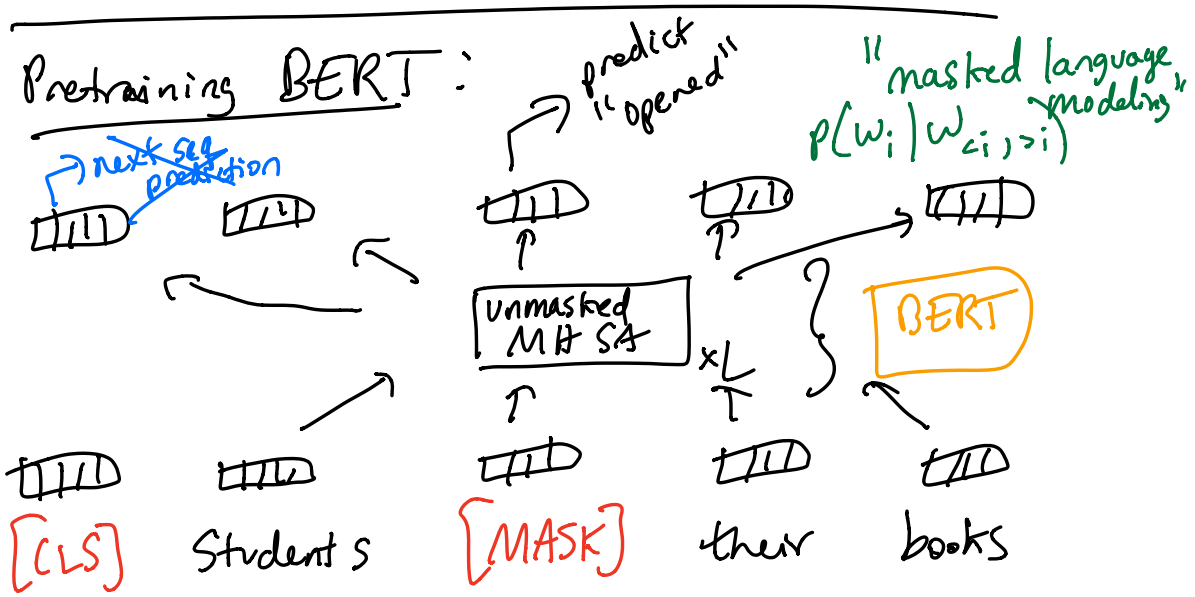
↳ example of an encoder-only Transformer

↳ Pretraining: training objective is self-supervised: "masked LM"

↳ fine-tuning: process of adapting a pretrained model to a particular downstream task

---

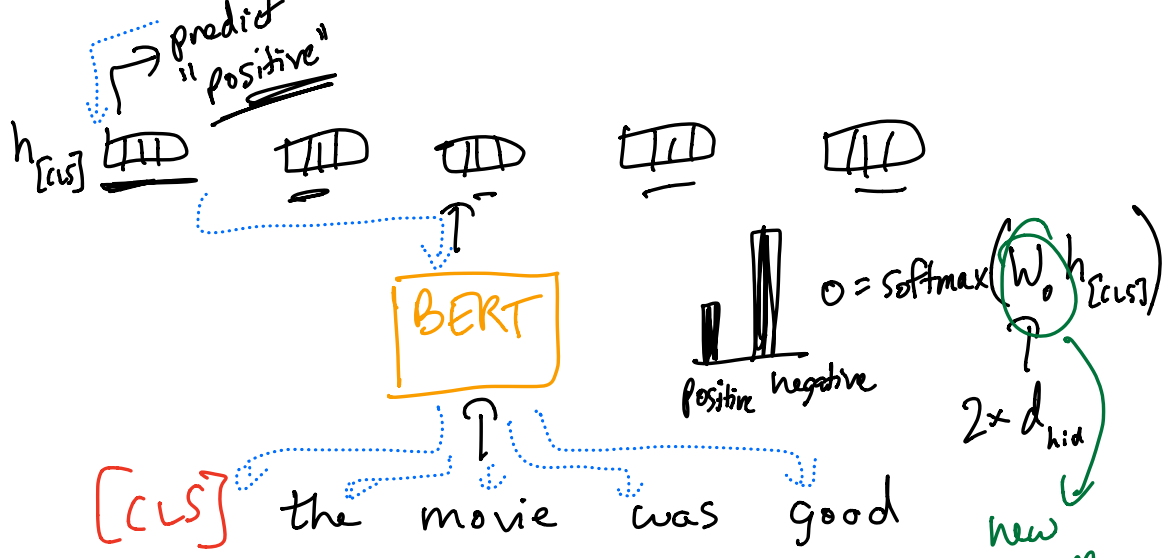
Pretraining BERT:



↳ occurs at the beginning of every input seq

---

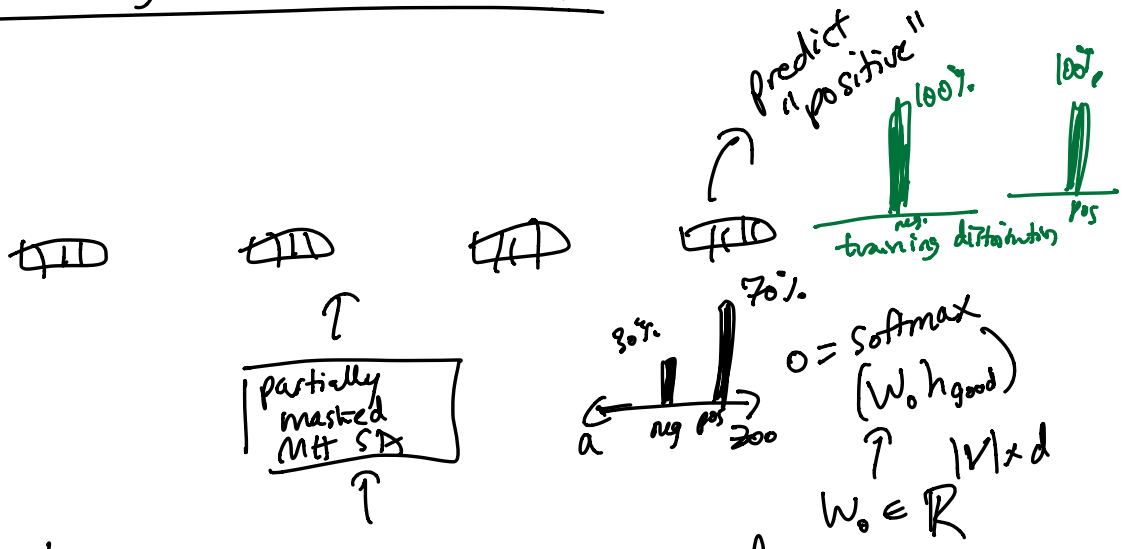
fine-tuning: Sentiment analysis, input → positive / negative



fine-tuning is NOT self-supervised! generally requires a labeled training dataset for the downstream task.

↳ uses far less data than pretraining

Fine-tuning a decoder-only LM:



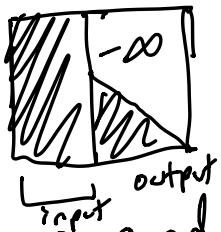
the movie was good

unmasked

LLaMA-2
GPT-2

fine-tuning a pretrained decoder model is useful for text generation tasks

no new params



the movie was good

input

positive because of "good"

Instruction tuning:

→ fine-tuning ("supervised fine-tuning"; SFT)

→ goal: make the pretrained model more capable of following instructions

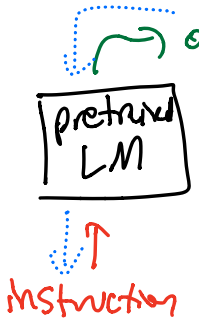
→ method: standard fine-tuning on a special dataset

1. collect a dataset of instructions on what tasks to solve, and outputs of that task for 1-2 examples

instruction: Please answer the following question and provide a detailed justification.

input: What was the avg of the 685 523 midterms?

output: I can't answer that because the Piazza is private.



- instruction tuning focuses on many diff. tasks at once, not just one
- instruction tuning improves generalization on tasks outside of the fine-tuning data