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:

- Next self-attention
- Predict "open"
- Unmasked MHA
- "masked language modeling"
- \( p(\text{words}|\text{mask}, \text{input}) \)

- [CLS] Students [MASK] their books

Occurs at the beginning of every input seq.
Fine-tuning: Sentiment analysis, input $\text{positive} \rightarrow \text{predict "positive"}$

$[CLS]$ the movie was good

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

$G$ uses far less data than pretraining
Fine-tuning a decoder-only LM:

The movie was good

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

No new params

The movie was good 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 685 528

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

Instruction/Input

pretend LM

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