downstream task: task that we care about solving (in this example, sentiment analysis)

pretraining:

[CLS] the students [MASK] the books

fine-tuning

predict positive

softmax trained from scratch

[CLS] this movie is good

data:
comes from huge volume of unlabeled text

data:
comes from labeled reviews (for sentiment)

BERT

predict "opened"
BERT for sentence pair classification:

NLI: natural language inference
  "textual entailment"

given two sentences \((s_1, s_2)\), a model must figure out if \(s_2\)
\(\not\)entails, contradicts, neutral\? to \(s_1\).

\(s_1 = "the dog walks"\)
\(s_2 = "the dog sits still"\)
\(\not\) contradiction

\(\Rightarrow\) SNLI, MNLI

\(\Rightarrow\) softmax, predict contradiction

\[\text{BERT} \]

\([\text{CLS}]\) the dog walks \(\text{[SEP]}\) the dog sits still
\(\Rightarrow\) special token in BERT vocab
BERT for extractive question answering:

- Input: question and a passage
- Goal: predict a contiguous span of text from the passage that answers the question

Examples: SQuAD, QuAC, CoQA, HotPotQA,

Q: Who starred in the Matrix as Neo?

A: (i, j)

Two binary classifiers:
- Predicts whether the token is the start of the answer
- Predicts whether it is the end of the answer

```
[CLS] Who starred in the Matrix [SEP] \text{\oe } \text{\oe } \text{\oe } Keanu Reeves
```
how do we select an answer span at test time?

→ find the span $P_{i...j}$ that maximizes

$P_{\text{start}(i)} \cdot P_{\text{end}(j)}$

→ exclude spans where $j < i$

→ exclude spans longer than some threshold

advanced variants of BERT:

→ pretraining improvements, ⇒ ROBERTA
  more data

→ longer sequences during pretraining
  - BERT = 512 tokens
  - XLNet = 900 tokens

→ more pretraining objectives
  ⇒ ELECTRA

→ smaller models
  ⇒ ALBERT, distilBERT, tiny BERT
**RoBERTa**: simple set of modifications

- train w/ bigger batches
  - smaller # of total batches
  - gradient accumulation
    - bypasses GPU mem. limitations
- has no pretraining task for [CLS]
- pretrain on more data
  - $16 \, \text{GB} \equiv 160 \, \text{GB}$
    - Common crawl
      - URLs from Reddit
- pretrained for longer
  - more total batches/epochs, 500k steps