BERT review:

Pretraining task: Masked LM

Downstream task: Ex: Sentiment analysis

Applying BERT for text classification:

Parameters "fine-tuned" thru downstream task signal

predict whether $s_1$ precedes $s_2$

[CLS] the students opened ...

[SEP] then they started the exam ...

[CLS] this movie is good

Special token used for classification tasks
**terminology:**

**pretrain:** start w/ randomly init. model, train it using a self-supervised obj
- LM, masked LM
- data is free
  - train big models on big datasets

**freeze:** do not backprop into the params of the pretrained model using the downstream training obj.

**fine-tuning:** backprop into pretrained model using task-specific signal

---

**BERT for sentence pair classification:**

**NLI:** natural language inference
  - "textual entailment"
    - given two sentences $s_2, s_2$, does $s_2$ entail, contradict, neutral $s_2$

  - e.g. "The dog walks" \(\neq\) contradiction
    - "The dog sits still"

  - SNLI, MNLI
BERT for question answering (extractive)

- input: question and a passage
- goal: predict a contiguous span of text from passage that answers the question

ex: SQuAD, QuAC, CoQA, ...

Q: Who starred in The Matrix?
P: ...
A: [Keanu Reeves]
(t: Keanu Reeves (i, j))
Two binary classifiers: one predicts whether a token is the start of answer, other predicts -- end of answer.


How to select answer at test time?

→ Find span pᵢ...pⱼ that maximize

STARTᵢ · ENDⱼ

Advanced variants of BERT:

→ Pretraining improvements \( \Rightarrow \) RoBERTa
→ More data

→ Longer sequences during pretraining \( \Rightarrow \) TransformerXL
→ XLNet

→ More efficient pretraining obj
→ ELECTRA
→ Smaller models \( \Rightarrow \) ALBERT
**RoBERTa**: simple collection of modifications

- train with bigger batches
  - smaller total # of batches,
  - larger batch size
  - gradient accumulation
    - bypasses GPU memory limitations

- remove [CLS] pretraining task of next sentence prediction
  - [CLS] token gets no special treatment

- pretrain on more data
  - 16 GB \(\Rightarrow\) 60 GB
    - Common Crawl
    - URLs from Reddit

- pretrain for longer
  - more total batches/epochs; SQuAD

---

**TransformerXL**: BERT/RoBERTa have a fixed max seq limit ↓ 512 tokens

- what if we add a recurrent mechanism that connects adjacent segments
- no gradient flow to previous segment
- practical limit of context size for TransformerXL is 900 tokens
ELECTRA: Jane goes to baseball practice

binary classifier: is token real or fake

Transformer

Jane goes to tree practice

how do i decide which words to replace and with what

L) "generator" model which is essentially BERT