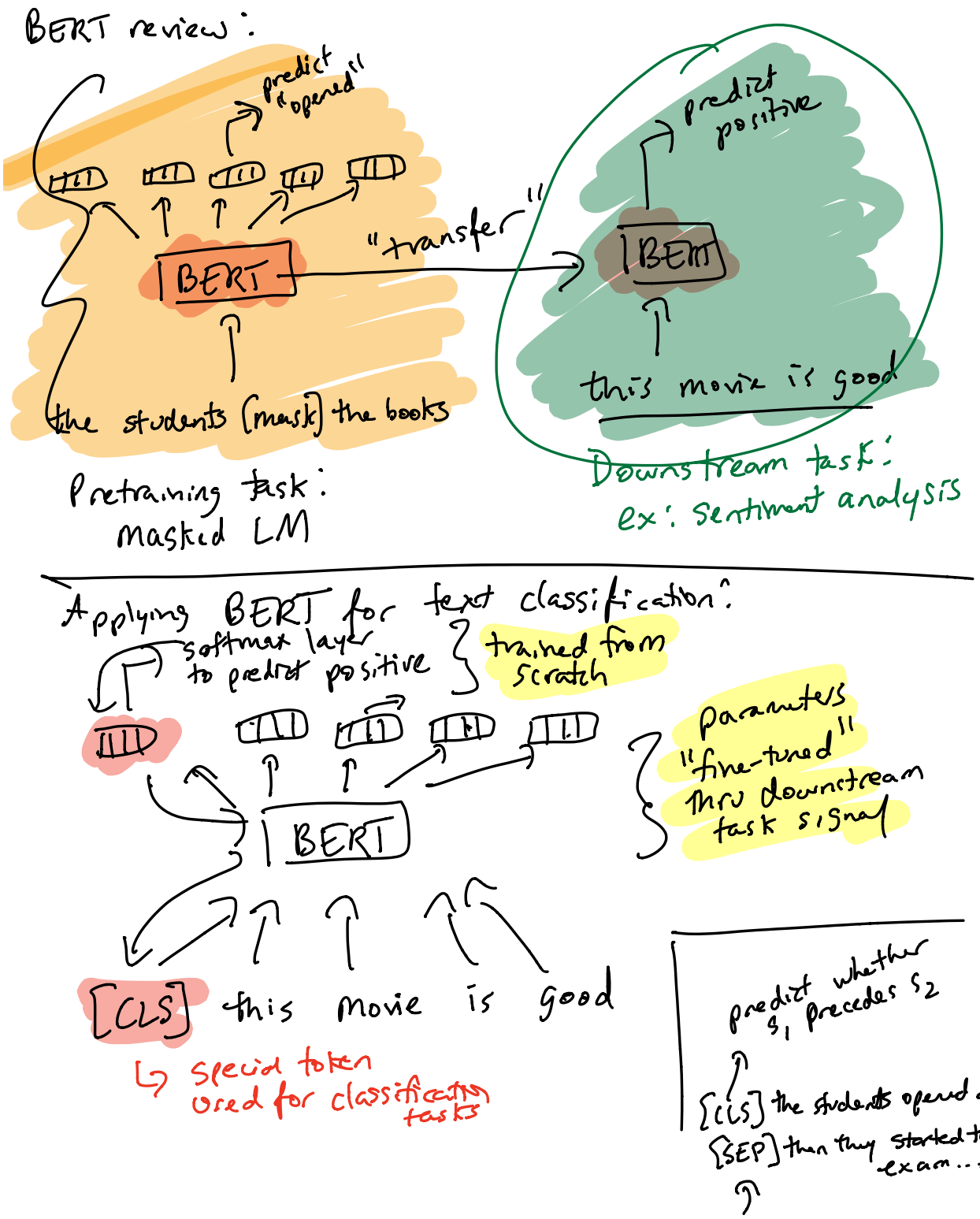

BERT review:
predict "opened"
BERT
"transfer"
the students [mask] the books
Pretraining task:
masked LM
predict positive
BERT
this movie is good
Downstream task:
ex: sentiment analysis
Applying BERT for text classification:
softmax layer to predict positive
trained from scratch
parameters "fine-tuned" thru downstream task signal
BERT
[CLS] this movie is good
↳ special token used for classification tasks
predict whether $s_1$ precedes $s_2$
[CLS] the students opened ...
[SEP] then they started the exam ...

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

---

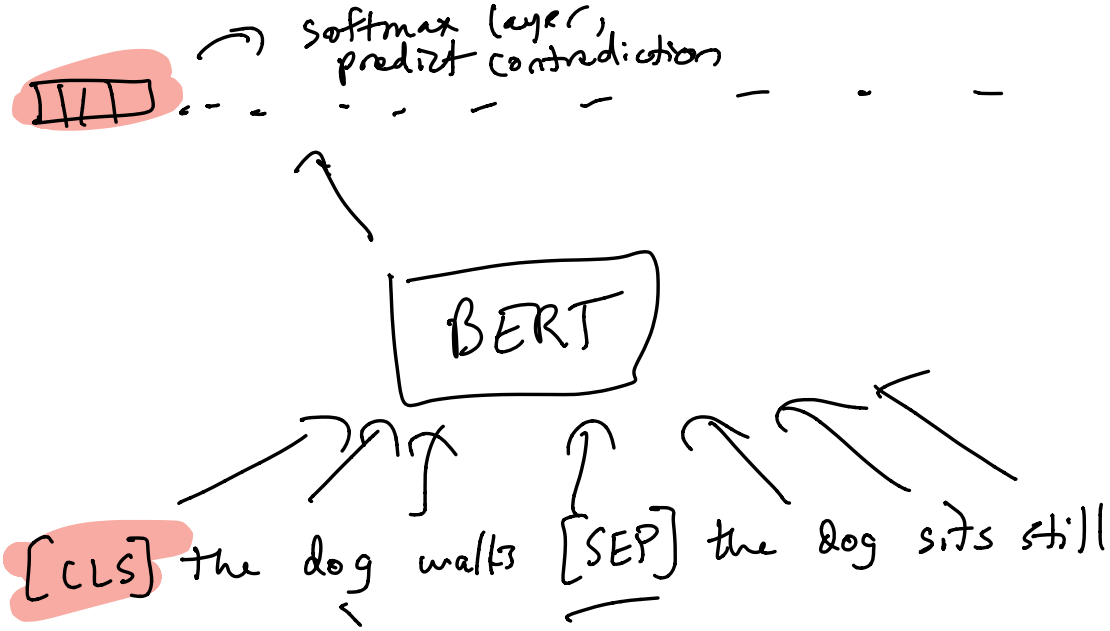
BERT for sentence pair classification:

NLI: natural language inference
"textual entailment"

given two sentences $s_2, s_2$, does
$s_2$ {entail, contradict, neutral} $s_1$

e.g. "The dog walks" } contradiction
"The dog sits still" }

↳ SNLI, MNLI




---

BERT for question answering (extractive)

↳ input: question and a passage

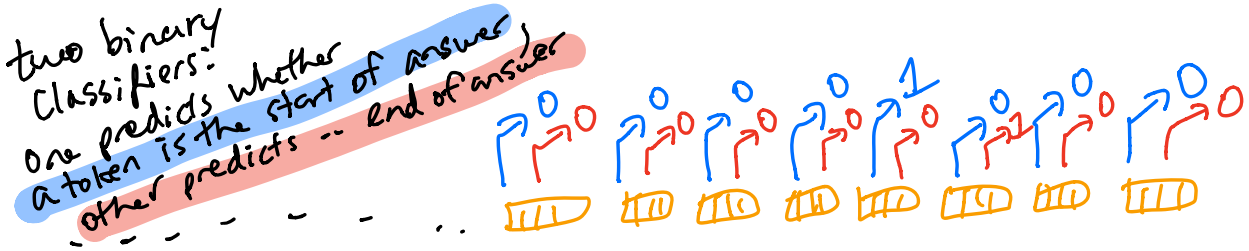
↳ goal: predict a contiguous span of text from passage that answers the question

↳ ex: SQuAD, QuAC, CoQA, ...

Q: Who starred in The Matrix?

P: ... — — — [Keanu Reeves] - -
- — — — — ? — . - ? - .
i j

A: Keanu Reeves
(i, j)




BERT

[CLS] Who starred in Matrix [SEP] $p_1$ $p_2$ $p_3$ $p_4$ Keanu Reeves $p_6$ $p_7$

how to select answer at test time?

→ find span $p_{i \ldots j}$ that maximize

$\text{START}_i \cdot \text{END}_j$

---

advanced variants of BERT:

- ↳ pretraining improvements ⇒ RoBERTa
  more data
- ↳ longer sequences during pretraining ⇒ TransformerXL
  XLNet
- ↳ more efficient pretraining obj;
  - ↳ ELECTRA
- ↳ smaller models ⇒ ALBERT

RoBERTa: simple collection of modifications

- train w/ 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 ⇒ 160 GB
    - ↳ common crawl URLs from reddit
- pretrain for longer
  - ↳ more total batches / epochs, 500K steps

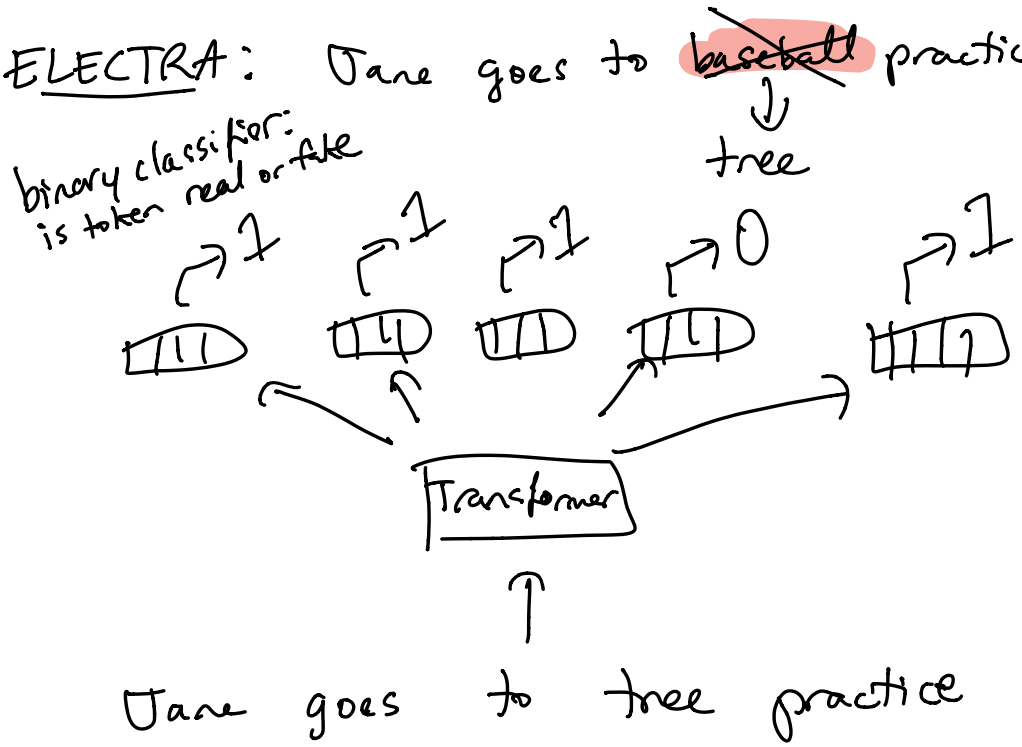
---

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~~ tree practice




how do i decide which words to replace and with what

↳ "generator" model which is essentially BERT