Limitations of BERT:

- What NLP task cannot be solved w/ BERT?
  - ↳ text generation
    - ↳ translation, summarization etc.
- Can we develop a self-supervised pretraining objective that covers all types of NLP tasks?
  - ↳ in addition to text gen, we'd like our model to handle classification, QA, sequence labeling, etc.
- T5 paper: reformulate every NLP task as a text generation problem, "text-to-text"

---

input/output for T5:

students opened their books

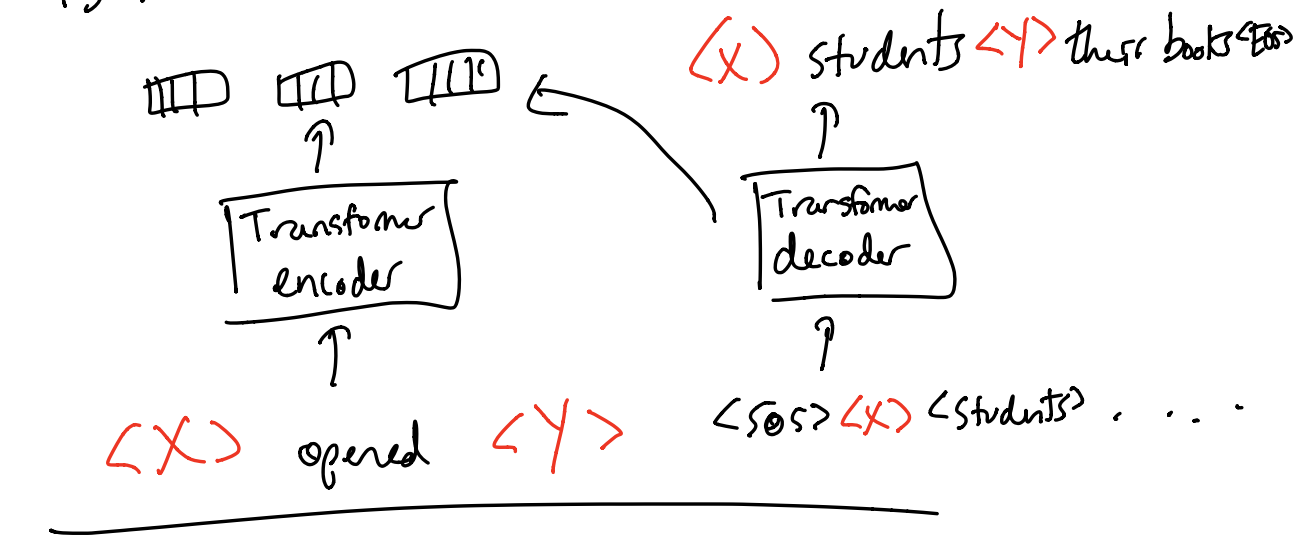
masked LM:

students [MASK] their books

predict "opened"

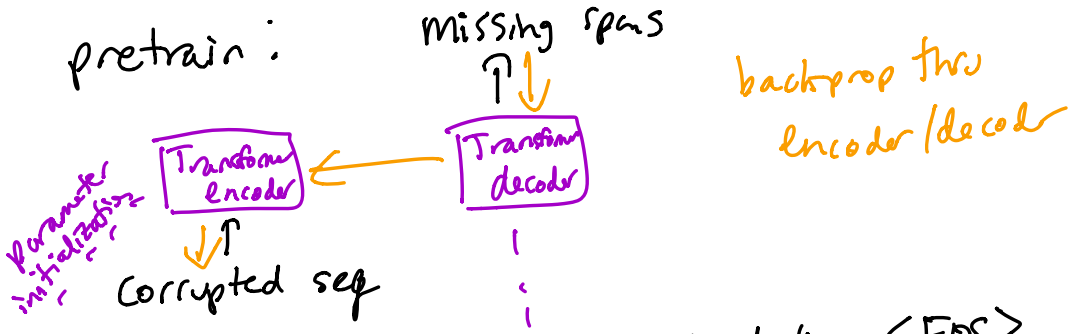
(Transformer encoder)

T5:

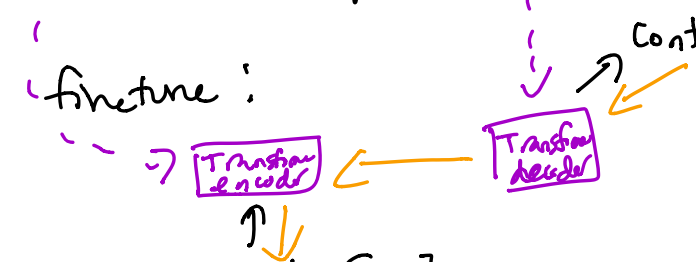


---

pretrain:



finetune:

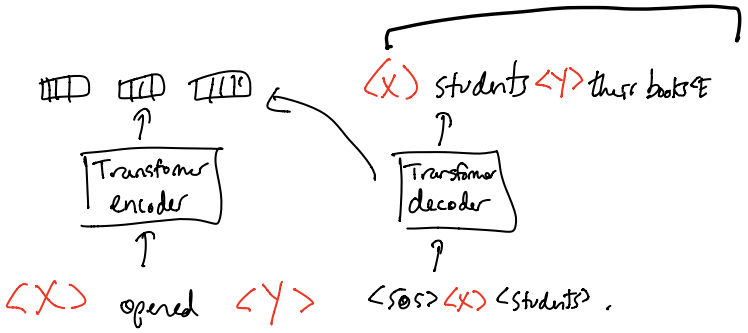


the dog walks [SEP]
the dog is sleeping

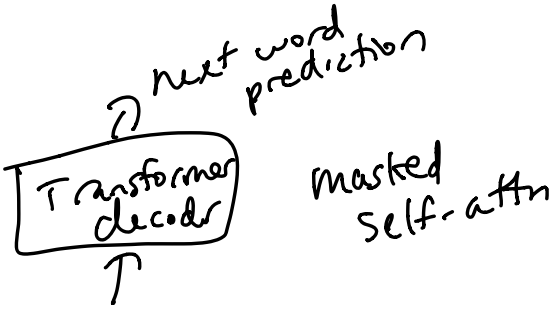
the decoder is actually generating the text of the label.
not a 3-way classification problem like in BERT.

T5 variants: (Table 2 of T5 paper)

encoder-decoder:



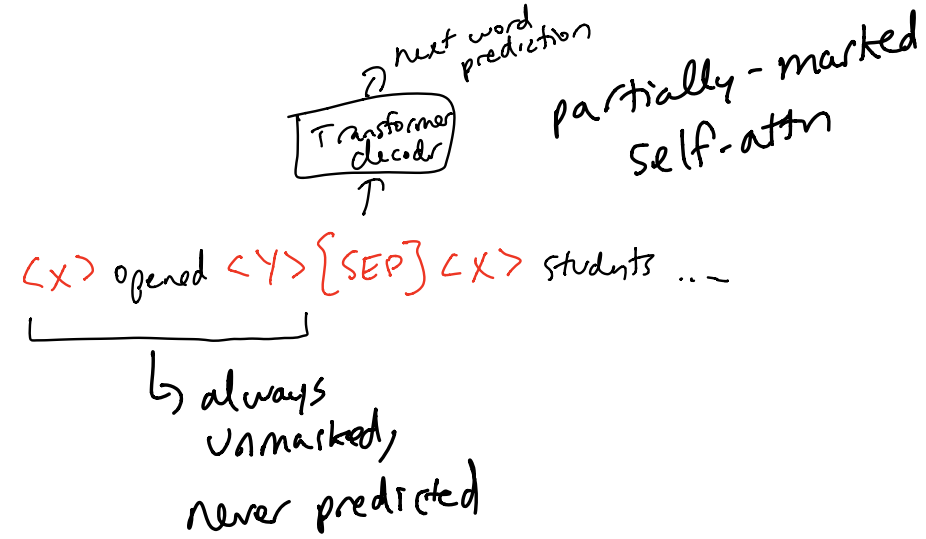
decoder-only:



<X> opened <Y> [SEP] <X> students ...

↳ decoder LM also predicts next word for the input sequence

prefix LM!


next word prediction
Transformer decoder
partially-marked self-attn
<X> opened <Y> [SEP] <X> students ...
always unmasked, never predicted