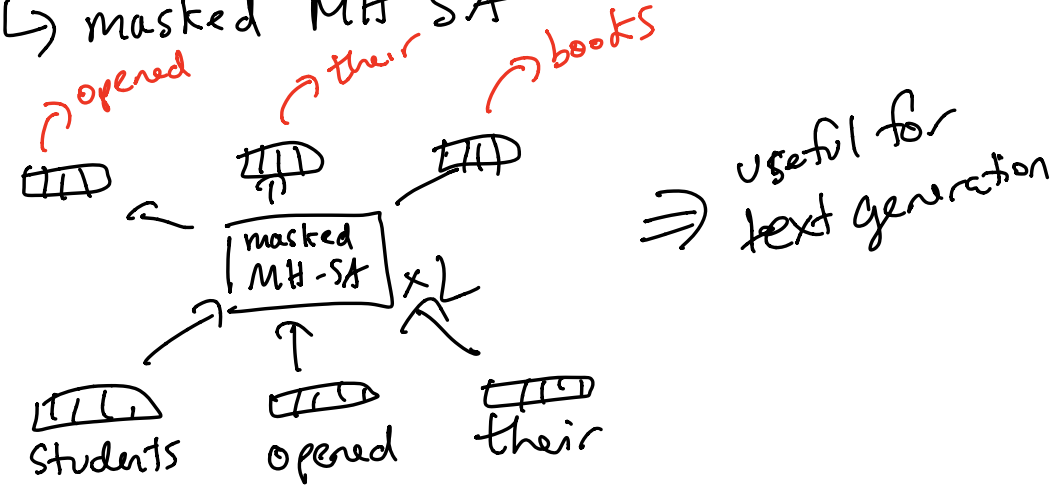
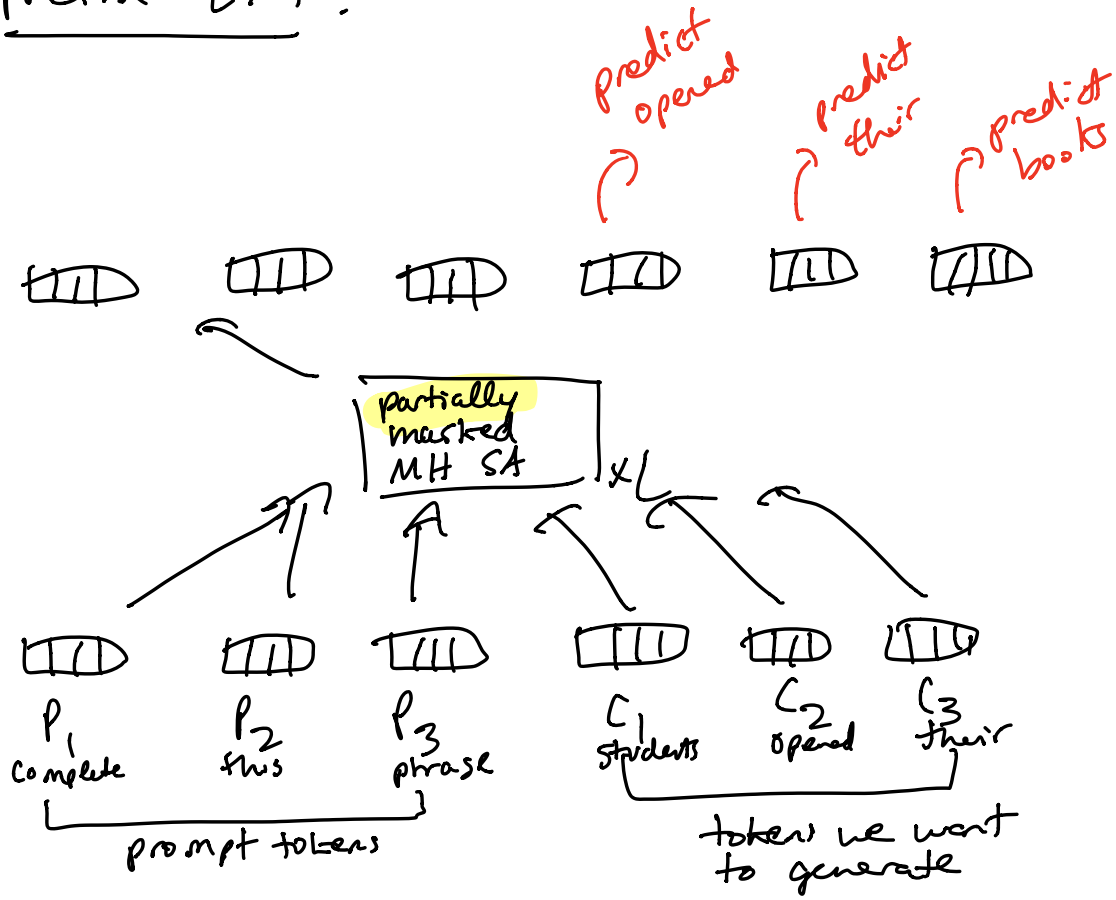
Transformer configuration:

Decoder:

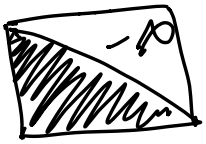
↳ masked MH SA



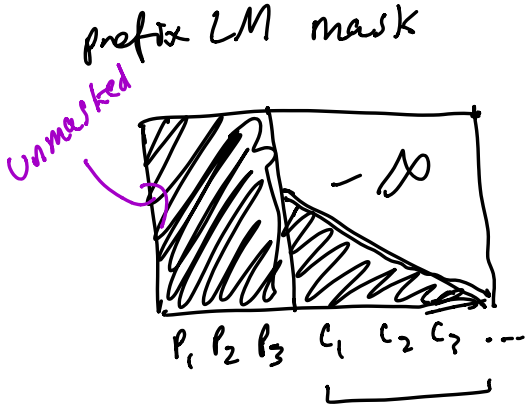
Prefix LM:



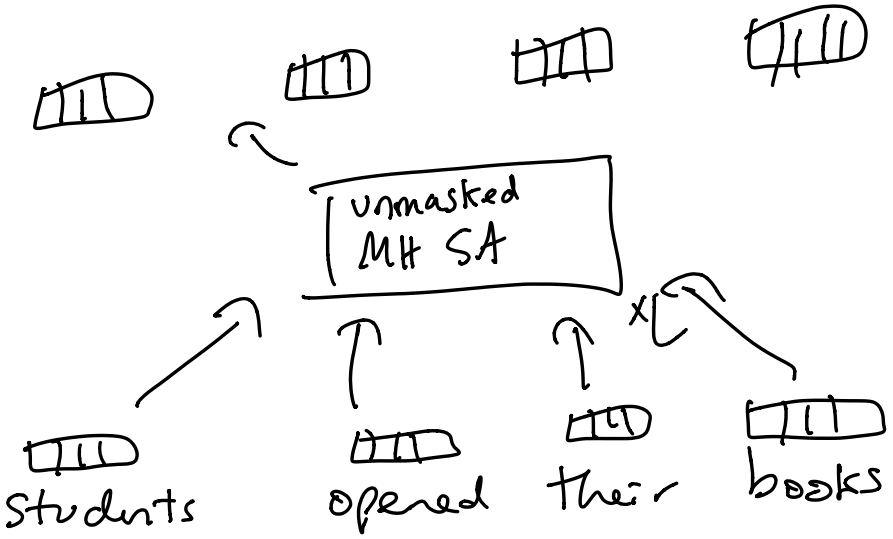
decoder mask



prefix LM mask



Encoder:



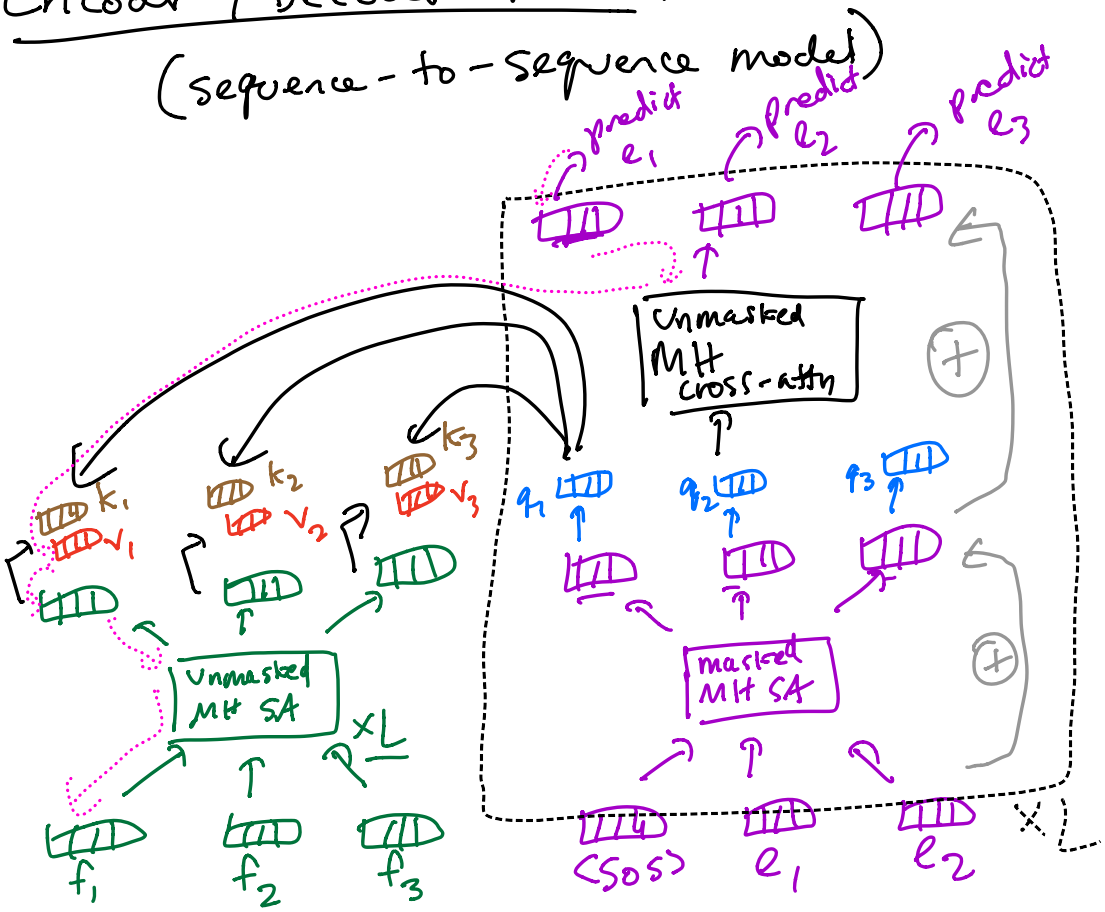
↳ useful for computing representations of a sequence of text that can then be used in other applications

↳ cannot generate text!

↳ ex: BERT, RoBERTa, ELECTRA

Encoder / Decoder model:

(sequence-to-sequence model)

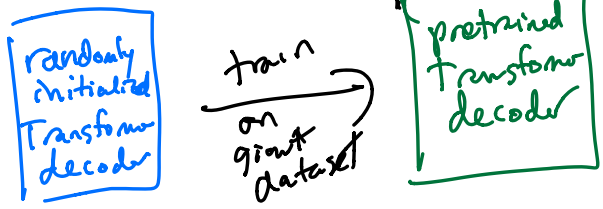


↳ Cross-attn always uses the representations from the final layer of the encoder

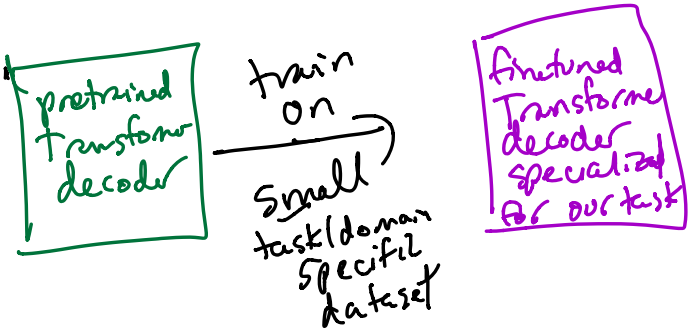
1. Pretraining
   - ↳ self-supervised objective
     - ↳ language modeling
   - ↳ use as much data as you can find
   - ↳ biggest model you can afford
   - ↳ goal: a model that understands many linguistic properties
     - ↳ grammar
     - ↳ world knowledge
       - "The President of the USA is ___"
     - ↳ "emergent properties"
   - ↳ we aren't focusing on a specific task or application

2. Fine-tuning
   - ↳ smaller labeled dataset corresponding to a single task/domain of interest
   - ↳ goal: maximize perf on this task/domain
   - ↳ parameter adaptation
     - ↳ parameter-efficient adaptation

Step 1 pretraining:



Step 2 finetuning.



Transfer learning

BERT:

↳ example of the encoder paradigm

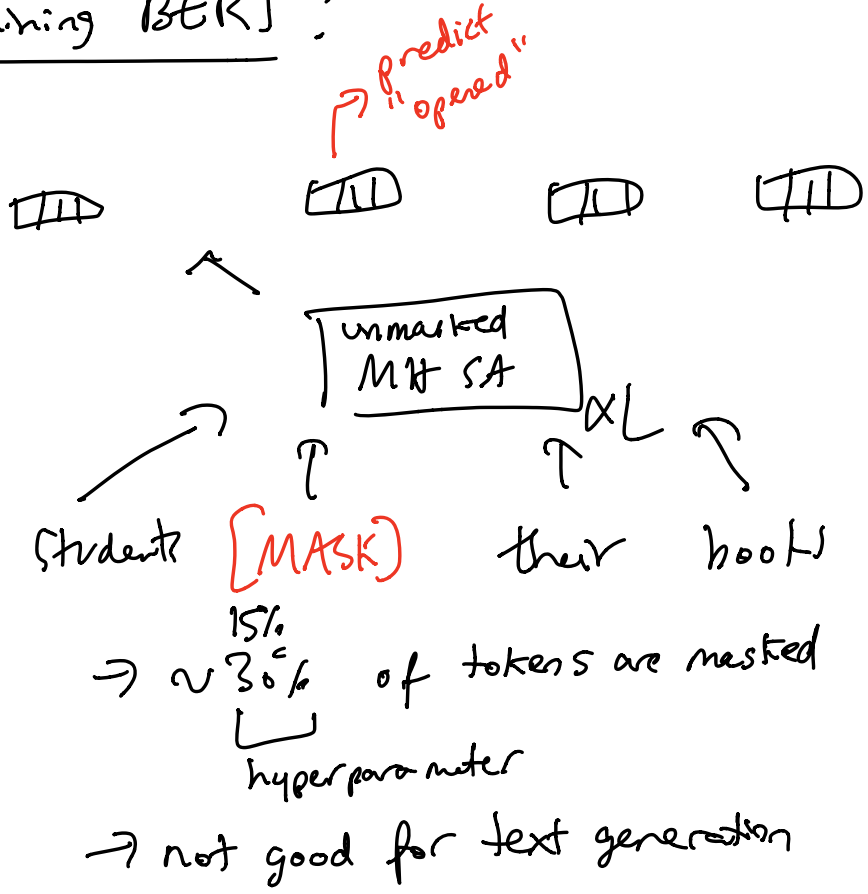
↳ pretraining

↳ masked LM

↳ finetuning

↳ adapt to "downstream" task

Pretraining BERT:



15%

→ ~30% of tokens are masked

hyperparameter

→ not good for text generation