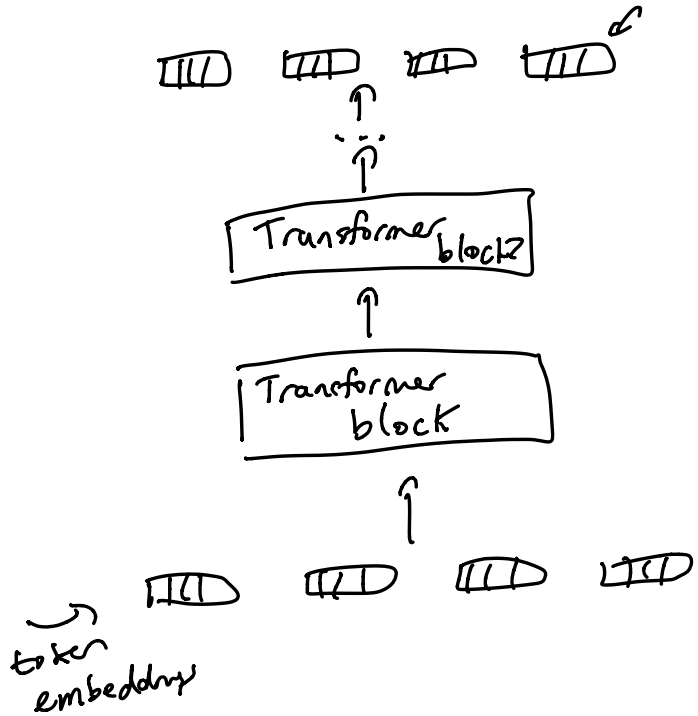
Today:

- "trained from scratch" vs. "fine-tuned"
- text-to-text transfer learning
- decoding algorithms to generate text

BERT for downstream tasks:

model: deep Transformer



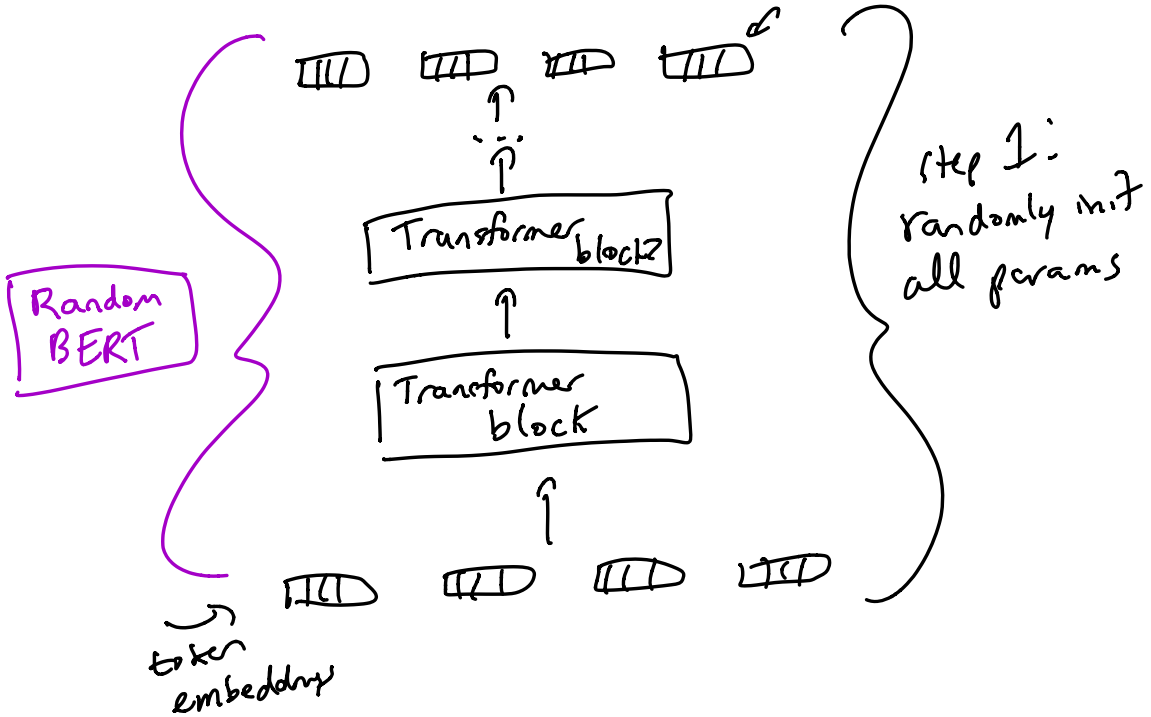
What are the model params?

W_q, W_k, W_v } for each head, each block

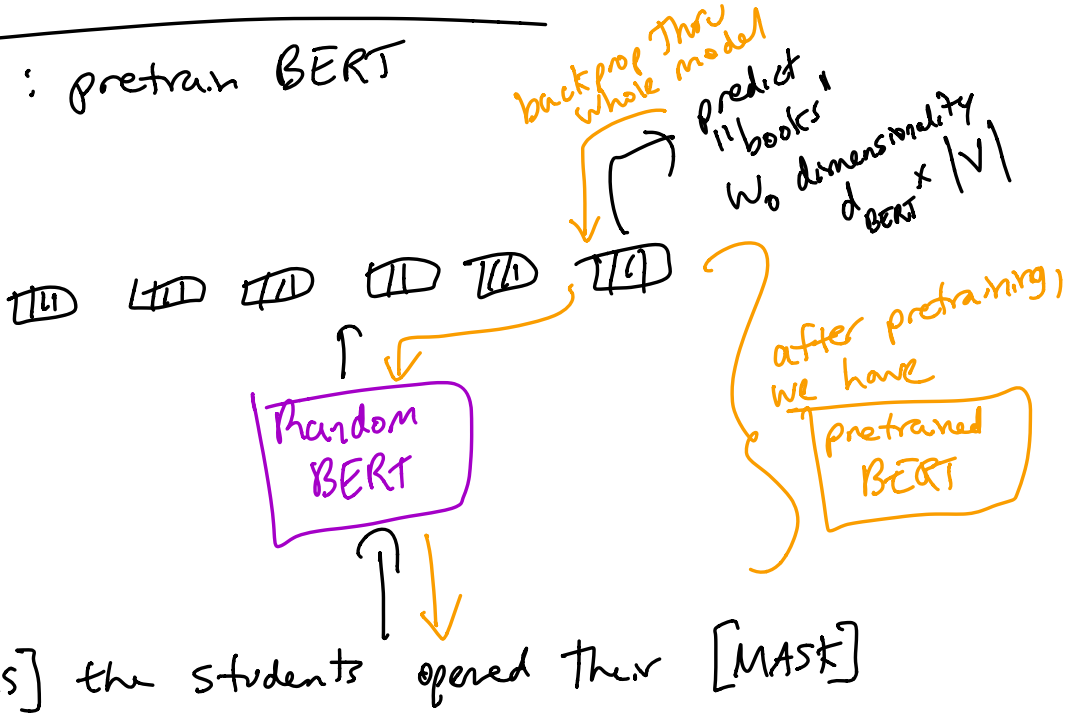
W_{ff} } feedforward matrix for each block

word embeddings (maybe pos embeddings)

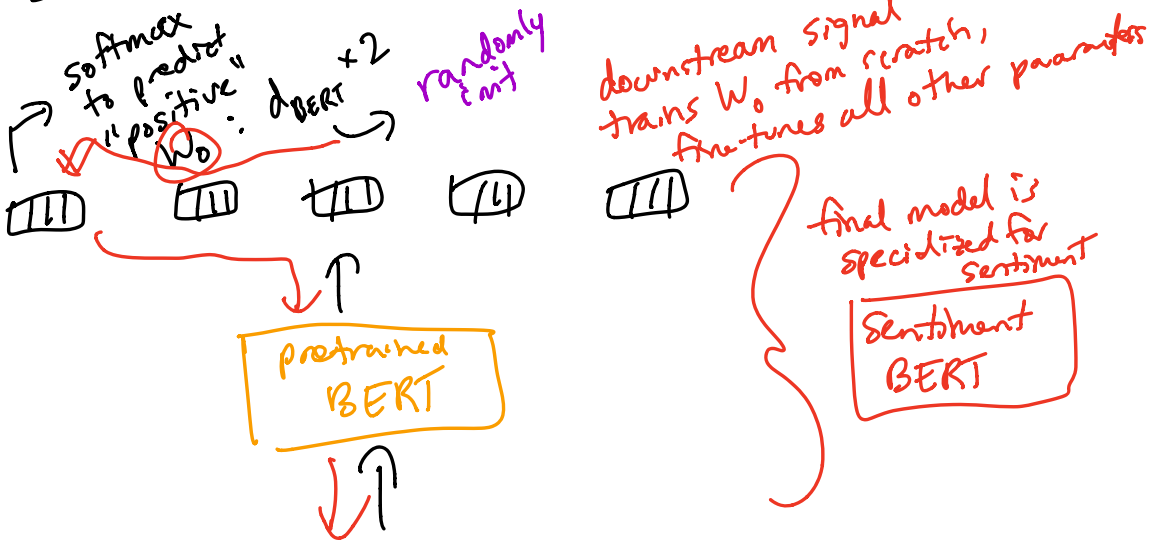
softmax matrix W_o



Step 2: pretrain BERT



Step 3: fine-tune BERT



Limitations of BERT:

What tasks can BERT not solve?

↳ text generation

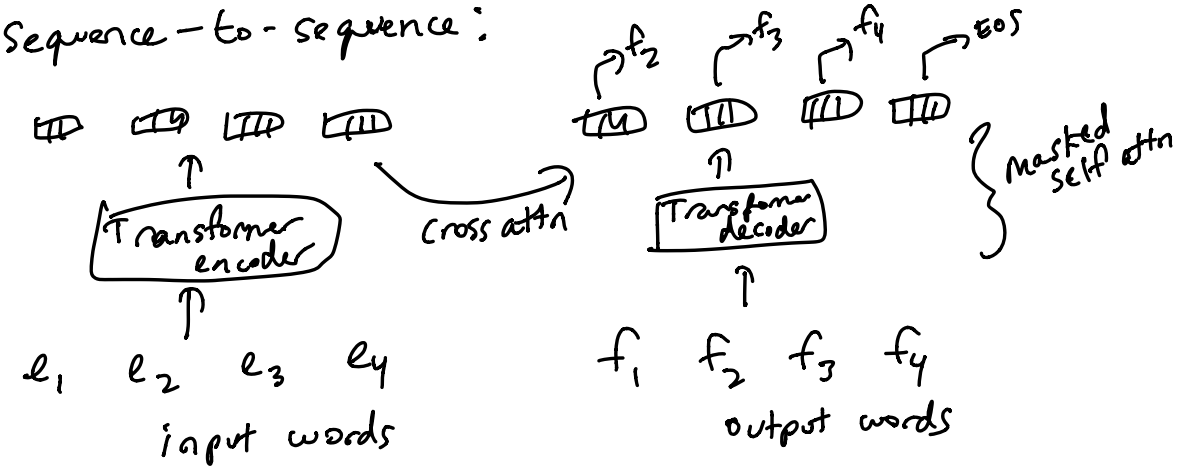
↳ translation, summarization

Can we develop self-supervised pretraining obj that covers all types of NLP tasks

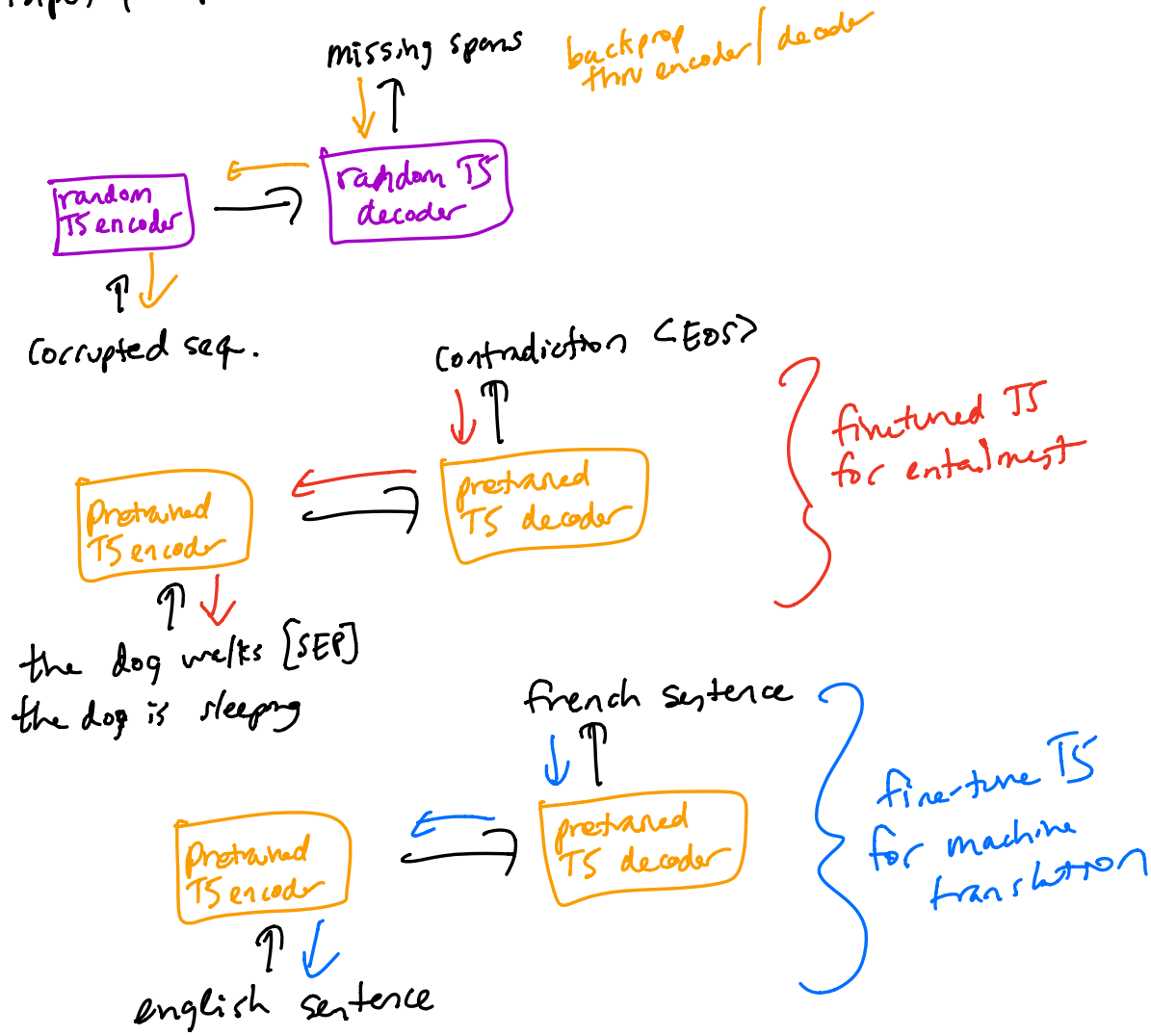
↳ classification, seq. labeling, generation

T5 paper: reformulate every NLP task as a generation problem, "text-to-text"

Sequence-to-sequence:



input/output for T5:



decoder-only version of T5:

<SEP>

thanks for <X> me to your party <X> for inviting <EOS>

masked self-attention, no cross attention