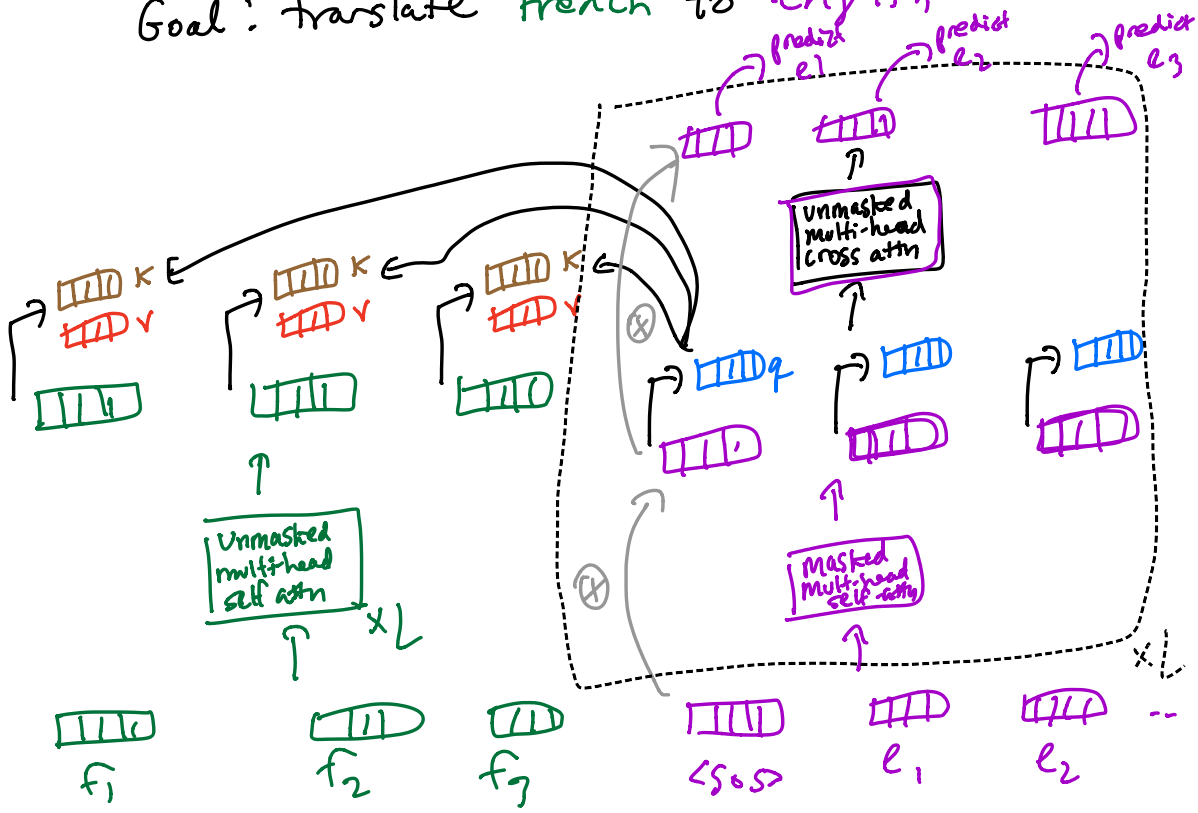
Different ways to use Transformer LMs:

- ↳ sequence-to-sequence models
  - ↳ encoder/decoder models ↳ T5
  - ↳ cross-attention
- ↳ encoder-only models
  - ↳ BERT

---

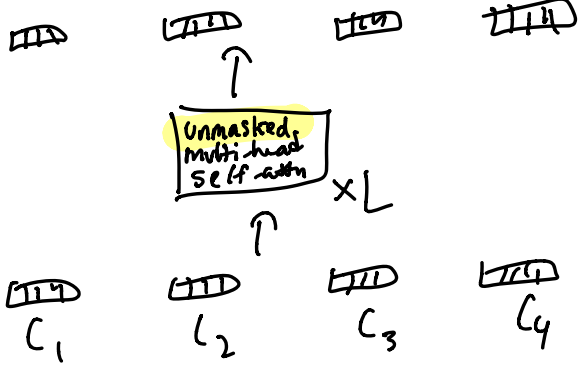
seq2seq

Goal: translate French to English



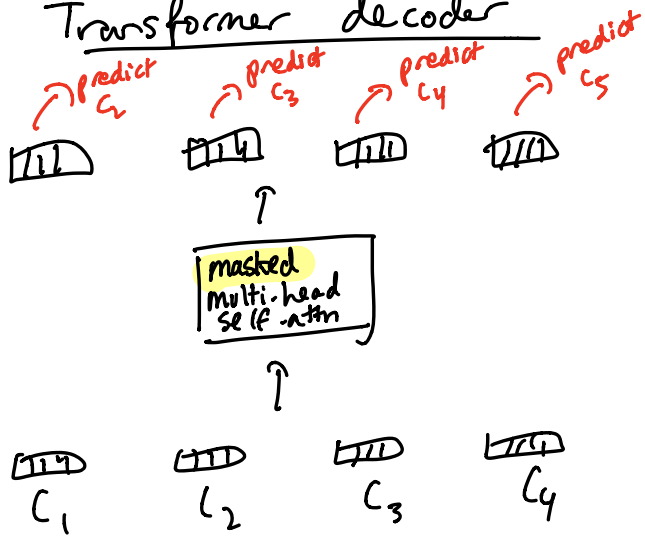
Common Transformer Configurations:

1. Transformer encoder:



} final layer token-level representations

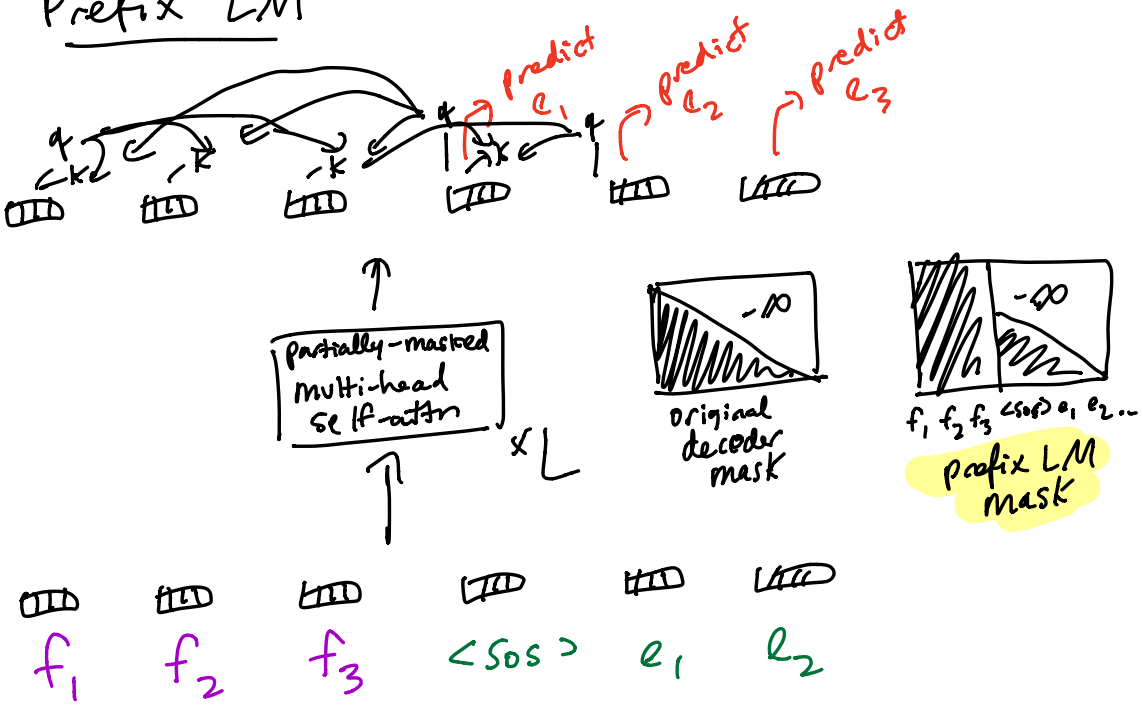
2. Transformer decoder



main differences:
- decoders can generate new text, encoders can't
- encoders observe a complete input, decoders only observe a prefix
- encoder's job is to produce powerful embeddings of the input, decoder's is to generate

3. Transformer encoder/decoder

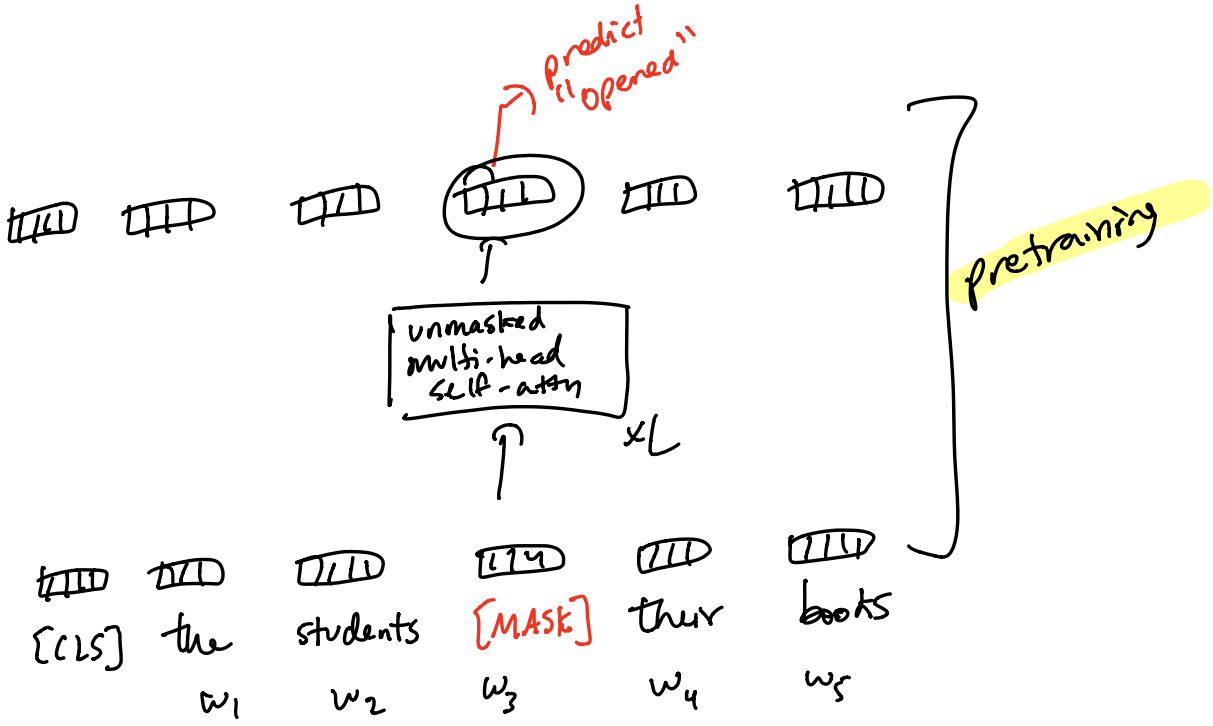
4. Prefix LM



---

pretraining an encoder-only Transformer:

↳ BERT: masked language modeling



instead of $p(w_3 = \text{opened} \mid w_1, w_2)$

in MLM we have $p(w_3 = \text{opened} \mid w_1, w_2, \{\text{MASK}\}, w_4, w_5)$

---

how do we USE a masked LM?

↳ **fine-tuning**: adjusting the params of a pretrained LM to adapt it to a single **downstream task**

ex: sentiment analysis:

- need a labeled dataset
- new softmax layer: $p = \text{softmax}(W_0 h_{[CLS]})$