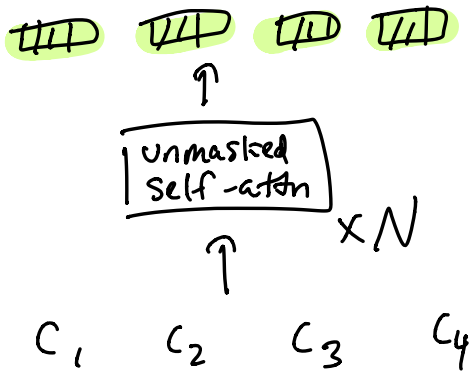
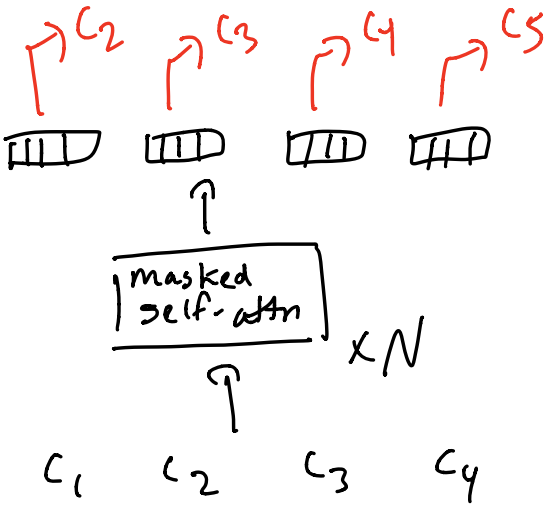
Common types of Transformers:

Transformer encoders (e.g., BERT)



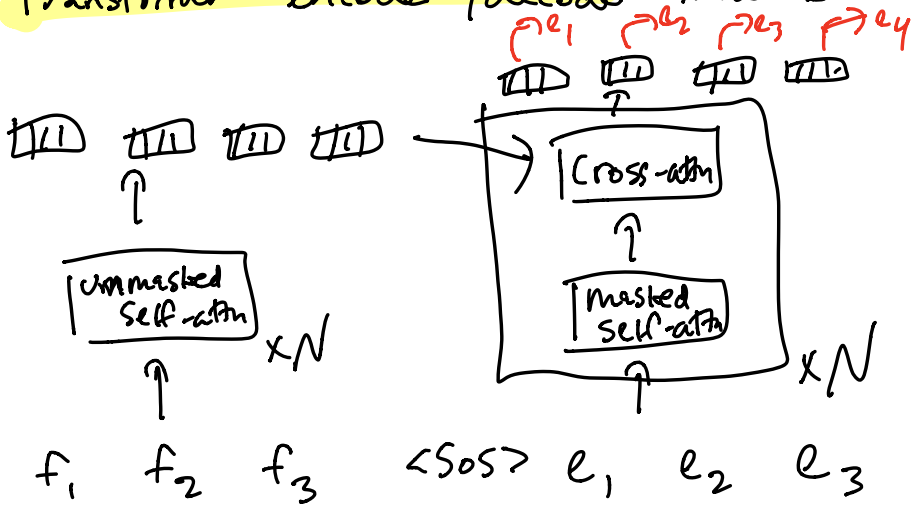
- useful if we're interested in using the final-layer token representations to solve downstream NLP tasks
- not useful for generation

Transformer decoder ("Transformer LM", GPT-2)



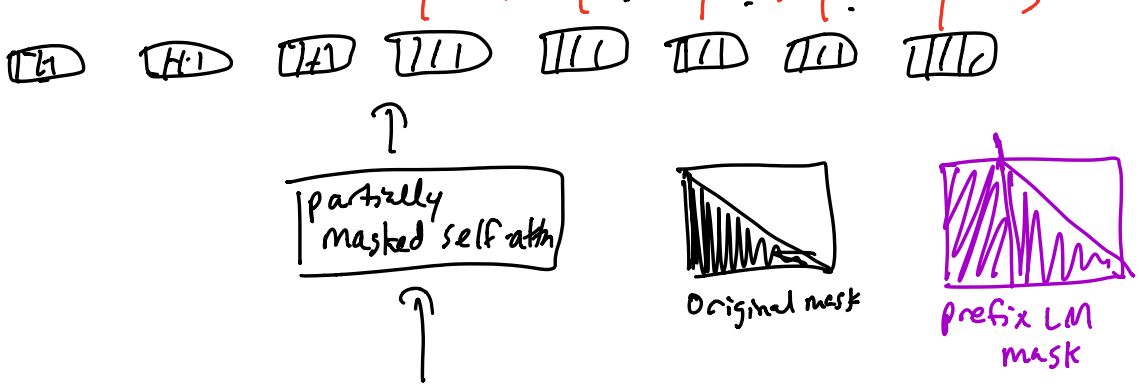
- useful to generate text, trained via LM objective

Transformer encoder/decoder model (e.g. T5):



- useful for conditional text generation

Prefix LM:

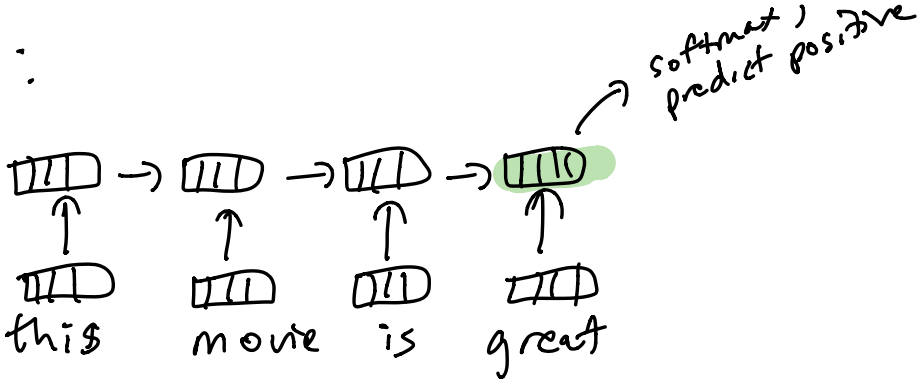


- useful for conditional text generation
  alternative to encoder/decoder approach w/ just 1 model instead of 2

Using neural LMs for transfer learning!

let's consider sentiment analysis

2013:



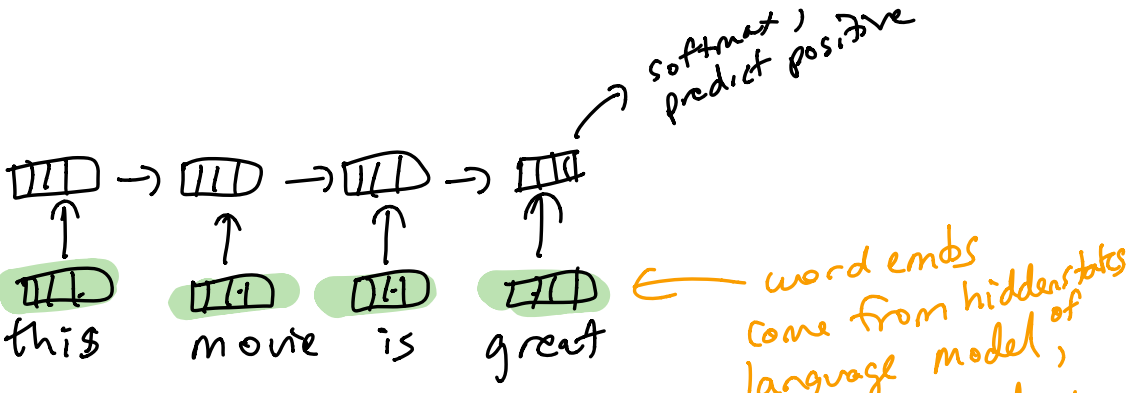
all params of RNN are trained from scratch on labeled sentiment data

issues:

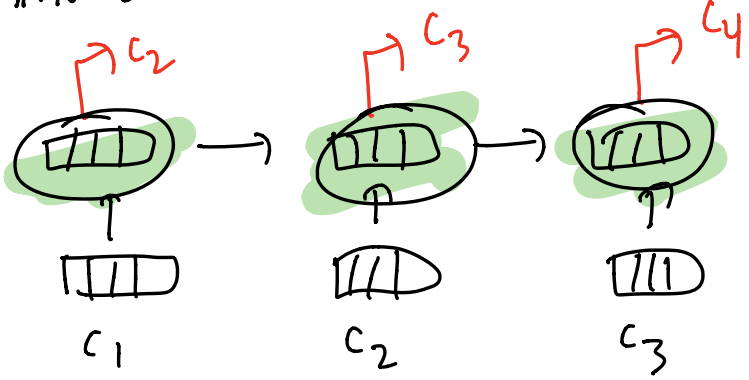
- we are forcing the RNN to learn composition and world knowledge just from a tiny labeled dataset

- what if we repurpose a large-scale neural LM to solve this task?

ELMo: embeddings from LM



RNN LM:



- Language models can be trained on huge datasets b/c labels come for free