Hide menu

Question Answering

Video: Week Introduction
41 sec

Video: Week 3 Overview
6 min

Reading: Week 3 Overview
10 min

Video: Transfer Learning in NLP
6 min

Reading: Transfer Learning in NLP
10 min

Video: ELMo, GPT, BERT, T5
7 min

Reading: ELMo, GPT, BERT, T5
10 min

Video: Bidirectional Encoder Representations from Transformers (BERT)
4 min

Reading: Bidirectional Encoder Representations from Transformers (BERT)
10 min

Video: BERT Objective
2 min

Reading: BERT Objective
10 min

Video: Fine tuning BERT
2 min

Reading: Fine tuning BERT
10 min

Video: Transformer: T5
3 min

Reading: Transformer T5
10 min

Video: Multi-Task Training Strategy
5 min

Reading: Multi-Task Training Strategy
10 min

Video: GLUE Benchmark
2 min

Reading: GLUE Benchmark
10 min

Lab: SentencePiece and BPE
2h

Hugging Face

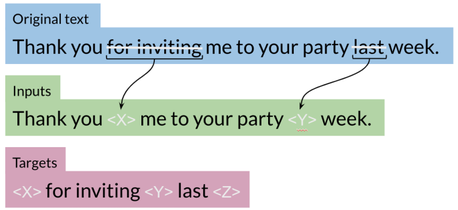
Lecture Notes (Optional)

Practice Quiz

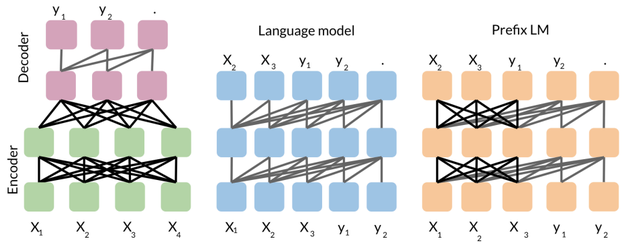
Assignment

Transformer T5

One of the major techniques that allowed the T5 model to reach state of the art is the concept of masking:



For example, you represent the “for inviting” with <X> and last with <Y> then the model predicts what the X should be and what the Y should be. This is exactly what we saw in the BERT loss. You can also mask out a few positions, not just one. The loss is only on the mask for BERT, for T5 it is on the target.



So we start with the basic encoder-decoder representation. There you have a fully visible attention in the encoder and then causal attention in the decoder. So light gray lines correspond to causal masking. And dark gray lines correspond to the fully visible masking.

In the middle we have the language model which consists of a single transformer layer stack. And it's being fed the concatenation of the inputs and the target. So it uses causal masking throughout as you can see because they're all gray lines. And you have X1 going inside, you get X2, X2 goes into the model and you get X3 and so forth.

To the right, we have prefix language model which corresponds to allowing fully visible masking over the inputs as you can see with the dark arrows. And then causal masking in the rest.