Transformers

Motivation and Intuition

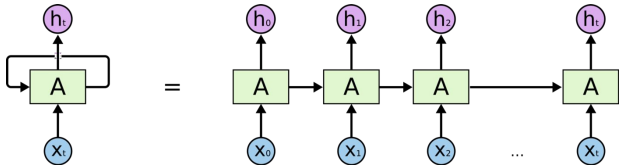
Overview

3 slide decks:

- **Motivation and Intuition**
- Architecture
- Popular Transformer-based models: BERT, GPT, RoBERTa, etc.

Motivations

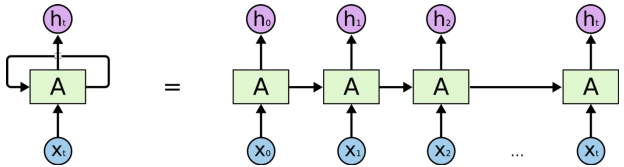
- Which disadvantages of recurrent models can you think of?



IMG: https://towardsdatascience.com/transformers-141e32e69591,

Motivations

- Which disadvantages of recurrent models can you think of?

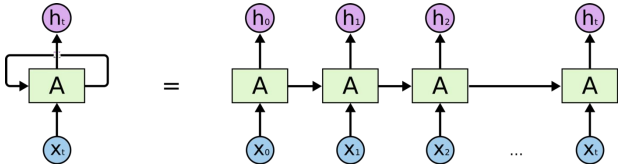


- Handling long distance dependencies: *Wer hat gestern Mittag kurz nach zwölf von meiner Palme die Kokosnuss geklaut?*
- Parallel processing: Faster implementations possible

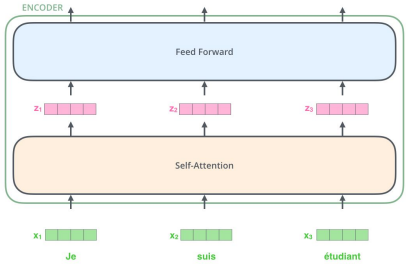
IMG: https://towardsdatascience.com/transformers-141e32e69591,

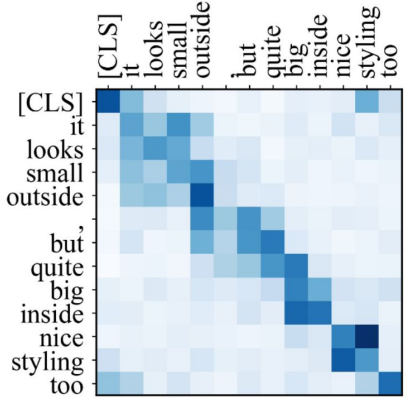
Motivation

- whole sentence is often available during processing
- So why go through the sentence word by word?
- Transfomer is more efficient and uses more information!



vs.

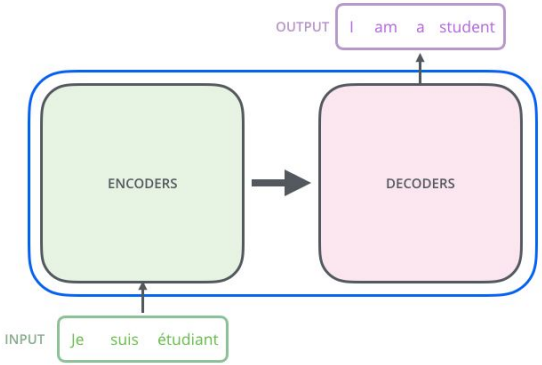




IMG 1: https://towardsdatascience.com/transformers-141e32e69591, IMG 2: http://jalammar.github.io/illustrated-transformer/
IMG 3: Cui et al. 2019 https://aclanthology.org/D19-1361.pdf

Transformer Architecture

- Designed for Translation: Attention is all you need, Vaswani et al. 2017
- Encoder-Decoder

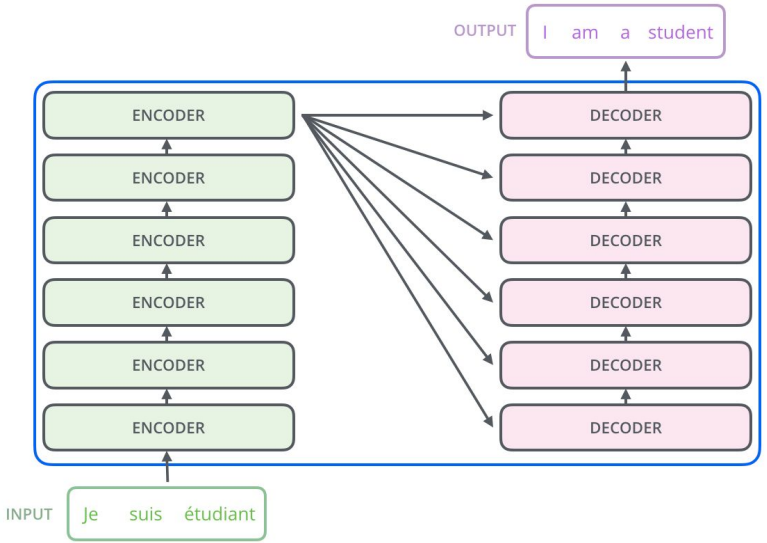


- Most LMs use only Encoder or Decoder

IMG: http://jalammar.github.io/illustrated-transformer/

Most important hyperparameters

- representation size *h* (768, 1024, 12888)
- number of layers *L* (12, 24, 96)
- number of heads *H* (12, 16, 96)
- number of parameters *P* (110M, 340M, 175B)



IMG: http://jalammar.github.io/illustrated-transformer/

Practical Session: Attention Heads

- Probability distribution over all words in a sentence
- Visualization: BertViz

Let’s try it out!