Attention is all you need 2017

Jan 6 25

Core:

the paper introduces Transformers, a neural network for sequence modeling (translation) that removes recurrence and Conv entirely and relies only on attention mechanism

motivation:

RNN based models:

- slow training
- vanishing gradient
- hard to parellelize

Transformer:

- process sequence at once
- uses attention to connect tokens to each other
- trains faster and performs better

Architecture:

Transformer is still a Seq2Seq encoder-decoder model

Encoder:

Stack of N identical layers, each:

- multi-head self attention
- pos wise feed forward network
- residual conn + layer norm around each

Decoder:

Stack of N identical layers each:

- masked multi head self attention (prevent looking ahead)
- encoder decoder attention (cross attention)
- feed forward
- res conn + layer norm