Understanding the Transformer Architecture

IITG Winter School on Deep Learning, January 2025

Dr. Amit Awekar

Primary job of ANNs: Generate representation

Feed Forward ANNs are rigid

- Fixed number of inputs
- Fixed number of outputs

RNNs can accept input of any size

- One token at time processing
- Output at each step or final step

RNNs can be configured in multiple ways

- One to one: FF ANN
- One to many: Image captioning
- Many to one: Sentiment analysis
- Many to many without delay: Entity detection
- Many to many with delay: Translation

Encoder Decoder architecture can handle any data modality

- Input is a sequence of tokens
- Output is a sequence of tokens

RNNs have multiple limitations

- Sequential input processing
- Vanishing gradient

Attention mechanisms: Focus on important part of input

- Global attention: Consider all input
- Local attention: Select a window of input

Transformer = RNN - Input recurrence

- Encoder Decoder architecture
- Self attention
- Masked attention
- Encoder Decoder attention
- Position encoding
- Residual connections

Each encoder has four components

- Self attention
- Residual connection and normalisation
- Feed Forward NN
- Residual connection and normalisation
-

Self attention block generates context sensitive representation

- Query
- Key
- Value
- Attention weights

Self attention generalises the key value search in databases

- Select value from table where key = query
- Select weighted value from table where key is more similar to query

Each encoder has multiple attention heads

- Intuitively each attention head focuses on different aspects of input

Encoder has residual connections to blend old representation with new

- Old is sometimes gold!

Normalisation keeps the values from getting large

- We want to prevent overflow
- Large values can arbitrarily change output

Sequence of encoders generate final representation of input

- More encoders, more parameters, more complex function of the input
-

Decoder generates one output token at a time

- Input sequence and partially generated output is the input for the decoder

Each decoder has six components

- Masked multi head attention
- Residual connections and normalisation
- Encoder Decoder attention
- Residual connections and normalisation
- Feed Forward NN
- Residual connections and normalisation

Position embedding allow the model to know the order of input

- Otherwise input simply becomes a bag of words
- Input representation = Original representation + Position specific representation
- Usually done only at the first encoder and decoder

BERT is an encoder only transformer

LLMs are decoder only transformers

Summary

- Transformers get rid of input side recurrence
- They still have output side recurrence
- They have more refined attention mechanism
- Next token prediction has turned out to be a far more versatile tool than anyone could have expected before