20

Transformers

More than meets the eye

20.1
Pre-Transformer Era.

Background

- The transformer architecture, was first introduced in the groundbreaking paper "Attention Is All You Need"
- It revolutionized the field of natural language processing (NLP) and has had a significant impact on other areas of machine learning as well

Previously Dominant Architectures

- Before transformers, the dominant architectures in NLP were based on Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs)
- RNNs, especially their advanced variants like LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit), were particularly popular for sequence-to-sequence tasks (like translation)

Limitations

- **Sequential Processing**: RNNs process data sequentially, which prevents parallelization within training examples. This becomes a bottleneck in terms of computational efficiency and training time
- **Long-Range Dependencies**: While LSTMs were designed to handle long-range dependencies better than basic RNNs, they still struggled with very long sequences, making it hard to capture context effectively in large documents

20.2
Introduction to Transformers.

Motivation

- The primary motivations behind the development of the transformer architecture were:
  - To increase the training speed by enabling parallel processing
  - To improve the ability to capture long-range dependencies
  - Simplify various RNN-specific techniques

Main Features of Transformer Architecture

- The transformer model addresses the limitations faced by the prior models
- It had a few key features such as:
    - Self-Attention Mechanism
    - Parallelization
    - Scalability

Self-Attention Mechanism

- The core idea is the attention mechanism, which allows the model to weigh the importance of different parts of the input data
- It's particularly effective in understanding the context and relationships between words in a sentence, regardless of their positional distance

Parallelization

- Unlike RNNs, transformers process entire sequences at once, not sequentially
- This allows for much more efficient training as operations can be parallelized

Scalability

- Transformers are highly scalable, which means they can be trained with a large amount of data, and their capacity can be increased to improve performance

Impact

- Transformers quickly set new records in a wide range of NLP tasks. It directly helped to create various state-of-the-art models
- They led to the development of models like BERT (Bidirectional Encoder Representations from Transformers) and the GPT (Generative Pre-trained Transformer) series
- The architecture has been adapted for use in other fields like computer vision and audio processing

20.3
Transformer **Architecture**.

Encoder-Decoder Structure

- A transformer consists of two main parts:
  - **Encoder**: The encoder processes the input data and transforms it into a rich, abstract representation
  - **Decoder**: The decoder takes the output of the encoder and generates the final output sequence
- Both are made up of multiple layers, and each contain the key component of self-attention

Self-Attention Mechanism

- Self-attention, sometimes called intra-attention, is a mechanism that allows each position in the input sequence to attend to all positions in the same sequence
- This is particularly powerful for understanding the context and relationships within the input data

Self-Attention Mechanism

- **Attention Score**: For each word in a sentence, the model computes a score that signifies how much focus to put on other parts of the sentence as the model processes that word
- **Scaled Dot-Product Attention**: The transformer computes the dot products of the query with all keys, divides each by the square root of the dimensionality (for scaling), and applies a softmax function to get the weights on the values

Multi-Head Attention

- Multi-head attention is an extension of the self-attention mechanism
- Rather than performing a single attention mechanism, the model does it multiple times in parallel - these are the "heads"
- Each head focuses on different parts of the input sequence, allowing the model to simultaneously attend to information from different representation subspaces

Multi-Head Attention

- Each head can potentially focus on different aspects of the input sequence, leading to a more comprehensive understanding
- Multiple heads can process the data simultaneously, making the model more efficient
- With multiple attention perspectives, the model can potentially learn more complex patterns

Overview of how a transformer works

- In the encoder, self-attention layers help the model to look at other words in the input sequence for better understanding context
- In the decoder, self-attention layers also look at the words in the output sequence to better predict the next word
- The multi-head attention attends to the encoder's output to understand how each word in the output sequence relates to each word in the input sequence

20.4
Transformer-based models.

Bidirectional Encoder Representations from Transformers

- Short for BERT, it is developed by Google
- BERT is adept at understanding the context of a word in a sentence, which improves its performance on tasks like sentiment analysis, named entity recognition, and question answering
- Use Case: Google uses a version of BERT to improve search query understanding

Generative Pre-trained Transformer

- Short for GPT, developed by OpenAI
- GPT models are known for generating coherent and contextually relevant text, making them suitable for tasks like content creation, story generation, and even code writing
- They are used in chatbots and virtual assistants for generating human-like responses

Vision Transformer (ViT)

- Developed by Google Research
- ViT applies the Transformer model, originally designed for text, directly to images
- Images are split into fixed-size patches, which are then linearly embedded. The sequence of these embedded patches (along with positional encodings) is fed into a standard Transformer encoder
- It has shown great success in image classification

Wave2Vec 2.0

- Developed by Facebook AI
- This model focuses on self-supervised learning from raw audio data. It captures the contextual representations of audio data, which are crucial for tasks like speech recognition
- It's primarily used for automatic speech recognition, enabling models to learn from unlabeled audio data effectively

Temporal Fusion Transformers (TFT)

- Developed by Google Cloud AI and Imperial College London
- TFT is designed specifically for interpreting and predicting time-series data
- The model can learn complex temporal relationships and handle missing data, variable input space, and correlation between time-series
- Useful in various time-series forecasting tasks

END.