Attenetion Is All You Need

Mohit Raj Aryal

Amnil Tech internship

Why was it needed?

- RNNs were slow.
- They struggled with long sequences.
- Parallel processing was difficult.

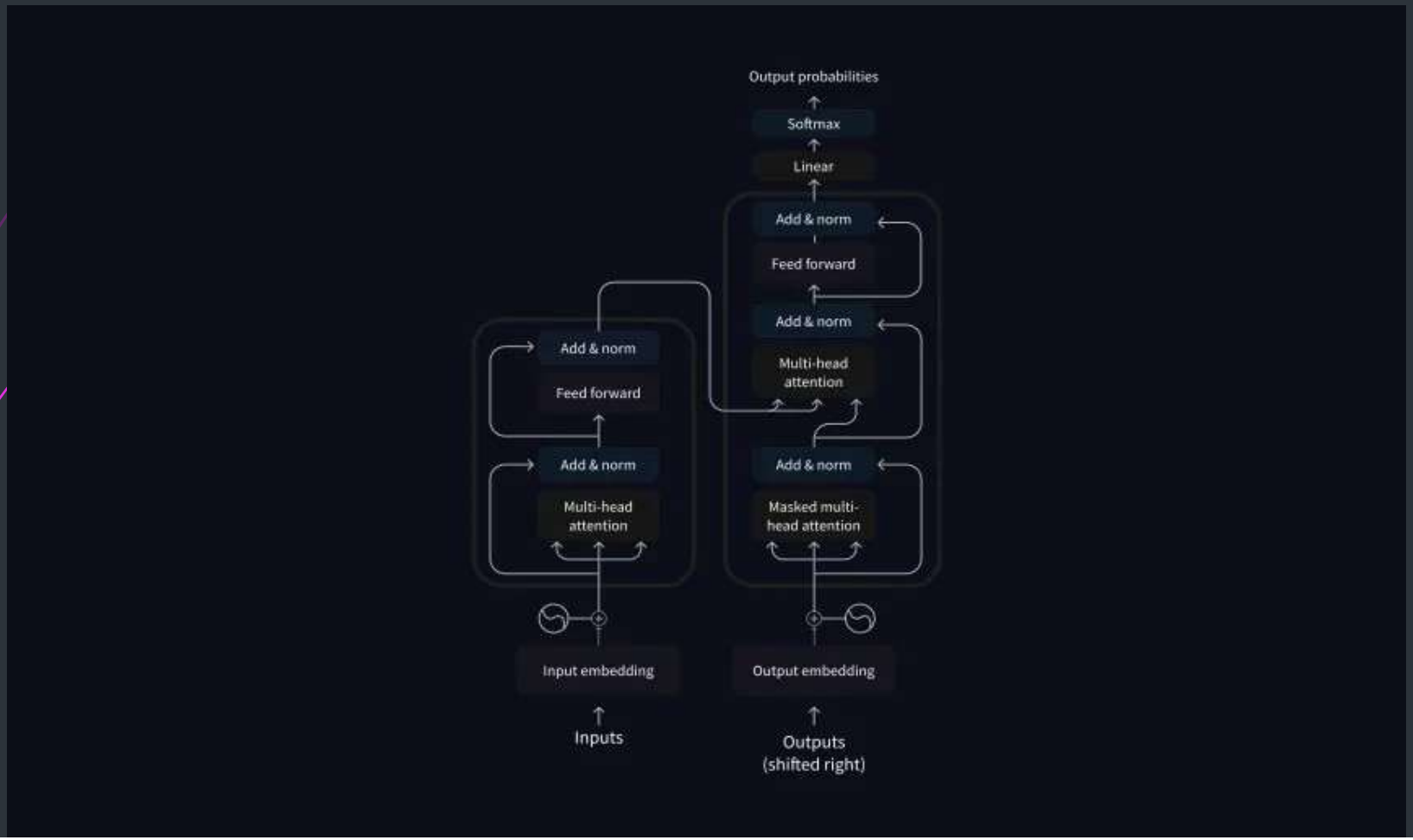
What was introduced?

- Remove recurrence.
- Use only attention.
- Enable full parallelism.

Transformer Architecture

- Encoder on the left.
- Decoder on the right.
- Both use attention blocks.
- Both use feed-forward layers.
- Both use residual + layer norm.

Proposed Architecture



Attention

- Core operation of the model.
- Measures how tokens relate.
- Helps focus on important words.

Scaled Dot-Product Attention

- Query × Key gives scores.
- Scores are scaled.
- Softmax gives weights.
- Weights × Value gives output.

Multi-Head Attention

- Many heads work in parallel.
- Each head learns a pattern.
- Captures richer relationships.
- Outputs are concatenated.

Positional Encoding

- No recurrence means no order.
- Positional encoding adds order.
- Uses sine and cosine waves.
- Helps model understand sequence.

Encoder

- Takes input tokens.
- Applies self-attention.
- Applies feed-forward layers.
- Produces context representations.

Decoder

- Takes target tokens.
- Uses masked self-attention.
- Looks at encoder output.
- Generates next tokens.

Why Masking?

- Prevents cheating.
- Model cannot see future words.
- Supports autoregressive generation.

Results in the Paper

- Outperformed RNNs.
- Faster training.
- Better translation quality.

Applications

- Translation.
- Summarization.
- Question answering.
- Chatbots.
- Many more.