# Attention Is All You Need

A groundbreaking paper introducing the Transformer model.

# The Transformer: A Novel Architecture

We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely.

### Attention-Based

Relies entirely on attention mechanisms.

### No Recurrence/Convolutions

Dispenses with traditional RNNs and CNNs.

### Faster Training

More parallelizable and significantly less training time.

# Superior Performance in Machine Translation

The Transformer achieves state-of-the-art results on major machine translation tasks.

## 28.4

### BLEU Score (EN-DE)

WMT 2014 English-to-German, outperforming existing best results by over 2 BLEU.

## 41.8

### BLEU Score (EN-FR)

WMT 2014 English-to-French, setting a new single-model state-of-the-art.

# Model Architecture Overview

The Transformer follows an encoder-decoder structure, utilizing stacked self-attention and fully connected layers.

### Encoder Stack

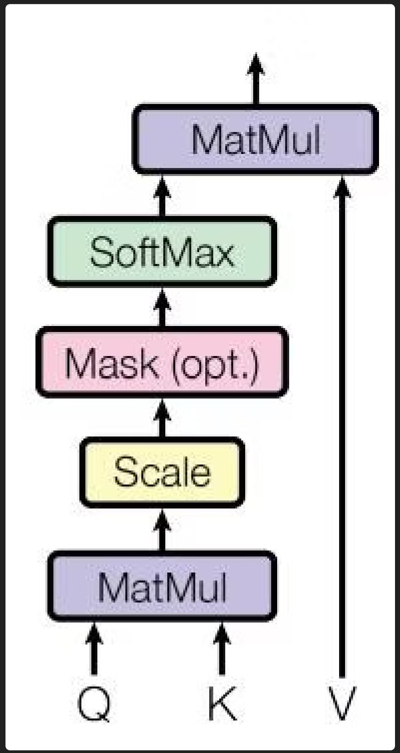
Composed of 6 identical layers with multi-head self-attention and feed-forward networks.

### Decoder Stack

Also 6 identical layers, with an additional multi-head attention over encoder output and masked self-attention.

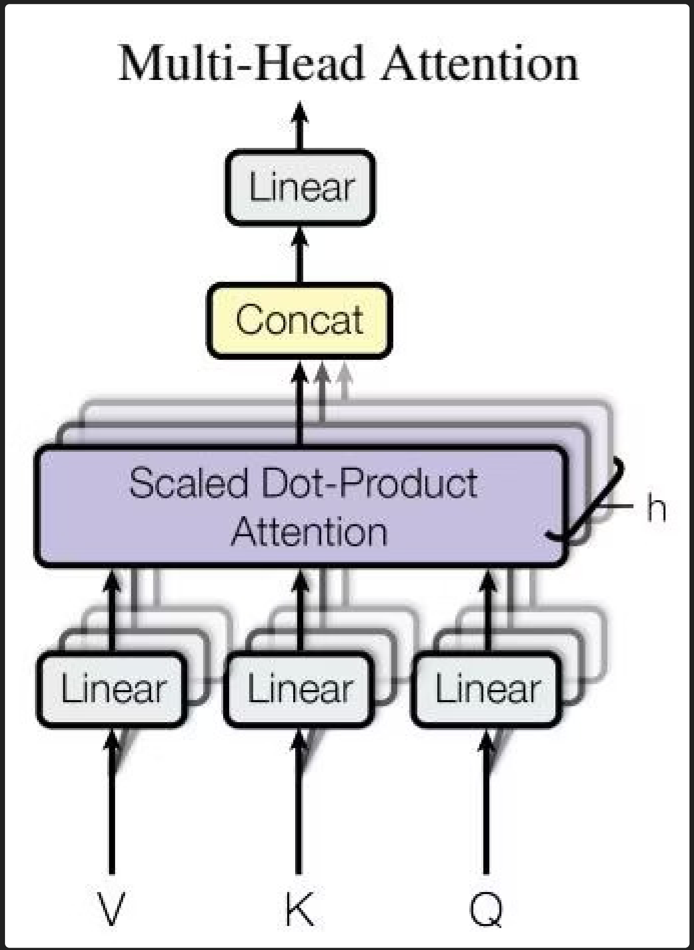
# Understanding Attention Mechanisms

Attention functions map a query and key-value pairs to an output, computed as a weighted sum of values.




### Scaled Dot-Product Attention

Efficiently computes attention weights by scaling dot products of queries and keys before softmax.




### Multi-Head Attention

Allows the model to jointly attend to information from different representation subspaces in parallel.

# Why Self-Attention?

Self-attention offers significant advantages over recurrent and convolutional layers, particularly in computational efficiency and handling long-range dependencies.

O(1)

### Sequential Operations

Self-attention connects all positions with a constant number of operations.

O(1)

### Maximum Path Length

Shorter paths between input/output positions facilitate learning long-range dependencies.

O(n^2 * d)

### Complexity per Layer

Faster than recurrent layers when sequence length (n) is smaller than representation dimension (d).

# Training and Regularization

Our models were trained on extensive datasets with specific optimization and regularization techniques.

1

### Training Data

WMT 2014 English-German (4.5M sentence pairs) and English-French (36M sentences).

2

### Hardware & Schedule

Trained on 8 NVIDIA P100 GPUs for 12 hours (base) to 3.5 days (big models).

3

### Optimizer

Adam optimizer with a varied learning rate schedule, including a warmup phase.

4

### Regularization

Residual Dropout and Label Smoothing to prevent overfitting and improve accuracy.

# Generalization and Future Directions

The Transformer generalizes well to other tasks, demonstrating its versatility and potential for future advancements.

### Constituency Parsing

Successfully applied to English constituency parsing, outperforming many previous models.

### Future Applications

Plans to extend to other input/output modalities like images, audio, and video.

### Research Goals

Investigating local attention mechanisms and less sequential generation.

# A Quote From One Of The Greatest Writers And Thinkers Of Our Time:

"You just want Attention"

- Charlie Puth .c 2017

# THANK YOU :)