The Transformer Architecture: Attention Is All You Need

CS 482 — Advanced Machine Learning Dr. Elena Vasquez 9 July 2026 Session 6

Overview

The **Transformer** (Vaswani et al., 2017) revolutionised sequence modelling by replacing recurrence entirely with *attention mechanisms*, enabling massively parallel training and superior long-range dependency capture. Every major language model — BERT, GPT-4, T5, and beyond — is built on this backbone.

Key Concepts

1 Scaled Dot-Product Attention

Given queries Q, keys K, and values V (each a matrix of row-vectors), attention is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$$

The $\sqrt{d_k}$ scaling prevents vanishingly small softmax gradients when the embedding dimension is large.

2 Multi-Head Attention (MHA)

Rather than a single attention function, MHA runs h heads in parallel on learned linear projections, then concatenates and re-projects:

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\, W^O$$

Different heads capture complementary relational patterns: syntactic agreement, coreference, positional proximity, etc.

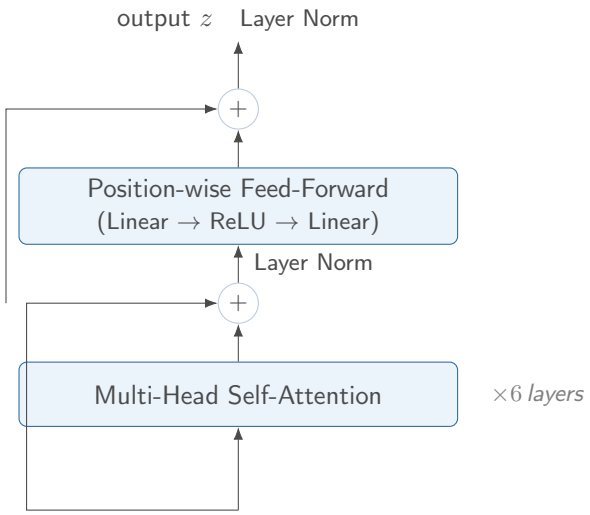
3 Positional Encoding

Attention is permutation-invariant, so position is injected via sinusoidal signals added to the input embeddings:

$$\text{PE}_{(p,2i)} = \sin\!\big(p/10000^{2i/d}\big), \qquad \text{PE}_{(p,2i+1)} = \cos\!\big(p/10000^{2i/d}\big)$$

Modern variants (RoPE, ALiBi) modify attention scores directly for better length generalisation.

Encoder Block — Bird's-Eye View



Trace: token embeddings enter the MHA sub-layer; a residual connection bypasses it; the sum is layer-normalised, fed into the feed-forward sub-layer, bypassed again, then layer-normalised to produce contextualised representation z.

Hyperparameter	Transformer-Base value
d_{model}	512
h **(heads)**	8
$d_k = d_v$	64
d_{ff}	2048
Encoder layers	6
Total parameters	65 M

Summary & Takeaways

- Transformers replace recurrence with **self-attention**, enabling full parallelisation across sequence positions.
- **Multi-head attention** lets the model jointly attend to information from different representation sub-spaces.
- **Positional encoding** is the sole source of sequential order in the architecture — removing it makes the model bag-of-words.
- Residual connections and **layer normalisation** stabilise deep stacks and mitigate vanishing gradients.
- The architecture scales predictably: more parameters + more data + more compute ⇒ log-linear performance gains (scaling laws).

Discussion Questions

Q1. Self-attention has $\mathcal{O}(n^2)$ complexity in sequence length. What architectural modifications have been proposed to reduce this cost, and what trade-offs do they introduce?

Q2. BERT uses bidirectional encoder-only Transformers; GPT uses decoder-only models with causal masking. How do these choices shape the tasks each family excels at?

Q3. Attention weights are often cited as evidence of what the model "focuses on." Critically evaluate this claim: when does the interpretation hold, and when does it mislead?

Q4. Positional encodings are needed because attention is permutation-invariant. Could permutation invariance itself be *exploited* as a feature in certain application domains?

Further Reading

1. Vaswani, A. et al. (2017). *Attention Is All You Need*. *NeurIPS 30*. arXiv:1706.03762
2. Devlin, J. et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. *NAACL-HLT*. arXiv:1810.04805
3. Kaplan, J. et al. (2020). *Scaling Laws for Neural Language Models*. arXiv:2001.08361
4. Tay, Y. et al. (2022). *Efficient Transformers: A Survey*. *ACM Computing Surveys* 55(6). arXiv:2009.06732
5. Elhage, N. et al. (2021). *A Mathematical Framework for Transformer Circuits*. Anthropic. transformer-circuits.pub

Prepared for CS 482 — Advanced Machine Learning, Westbrook University. Corrections: e.vasquez@westbrook.edu.