Transformer Architecture: Attention, Normalization, Positional Encoding, and Structural Variants

October 23, 2025

1 Attention Mechanisms

Self-attention enables transformers to capture pairwise interactions across a sequence without recurrence. Given input matrix $\mathbf{X} \in \mathbb{R}^{T \times d_{\text{model}}}$, linear projections yield queries $\mathbf{Q} = \mathbf{X}\mathbf{W}^{Q}$, keys $\mathbf{K} = \mathbf{X}\mathbf{W}^{K}$, and values $\mathbf{V} = \mathbf{X}\mathbf{W}^{V}$. Scaled dot-product attention computes

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}, \tag{1}$$

where d_k denotes key dimension. The scale $\sqrt{d_k}$ stabilizes gradients. Causal masking applies an additive $-\infty$ mask to enforce autoregressive structure.

1.1 Multi-Head Attention

Multi-head attention partitions embeddings into H subspaces, applying attention independently and concatenating:

$$\text{MHA}(\mathbf{X}) = \text{Concat}(\mathbf{O}_1, \dots, \mathbf{O}_H)\mathbf{W}^{O}, \tag{2}$$

$$\mathbf{O}_h = \text{Attention}\big(\mathbf{X}\mathbf{W}_h^{Q}, \mathbf{X}\mathbf{W}_h^{K}, \mathbf{X}\mathbf{W}_h^{V}\big). \tag{3}$$

Benefits include richer representation capacity, directional specialization, and the ability to model heterogeneous relations. Key variants include:

- **Relative attention** (Transformer-XL, T5) injecting distance-dependent bias.
- **Sparse attention** (Longformer, BigBird) restricting receptive fields for efficiency.
- **FlashAttention** computing attention in tiles to reduce memory bandwidth.

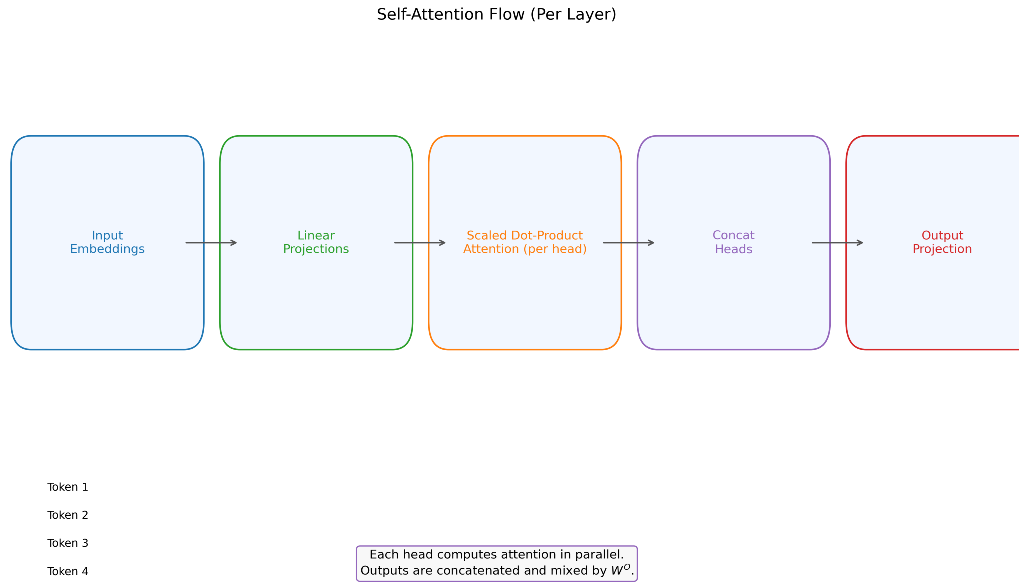
Figure **??** visualizes the multi-head pipeline, highlighting projection, attention, concatenation, and output mixing.

1.2 Cross-Attention

Cross-attention generalizes self-attention by using external context $\mathbf{Y}$ to provide keys and values:

$$\text{CrossAttn}(\mathbf{X}, \mathbf{Y}) = \text{Attention}(\mathbf{X}\mathbf{W}^{Q}, \mathbf{Y}\mathbf{W}^{K}, \mathbf{Y}\mathbf{W}^{V}). \tag{4}$$

Encoder-decoder transformers leverage cross-attention to condition the decoder on encoder outputs, enabling sequence-to-sequence modeling.



Each head computes attention in parallel. Outputs are concatenated and mixed by W^O.

Figure 1: Multi-head self-attention pipeline: projection into query/key/value spaces, attention computation per head, concatenation, and output mixing.

2 Residual Connections and Layer Normalization

Deep transformers rely on residual pathways and normalization to maintain gradient flow and stable activations.

2.1 Residual Paths

Each sublayer (attention or feed-forward network) adds its output to the input:

$$\mathbf{y} = \mathbf{x} + \mathrm{Sublayer}(\mathbf{x}). \tag{5}$$

Residuals mitigate vanishing gradients and allow the network to learn perturbations around identity mappings. Pre-activation placement (Pre-LN) applies normalization before the sublayer, offering training stability for deep stacks.

2.2 Layer Normalization

Layer normalization (LayerNorm) standardizes activations per token:

$$\hat{\mathbf{h}} = \frac{\mathbf{h} - \mu}{\sqrt{\sigma^2 + \epsilon}}, \tag{6}$$

$$\mathrm{LayerNorm}(\mathbf{h}) = \gamma \odot \hat{\mathbf{h}} + \beta, \tag{7}$$

where μ and σ^2 are the mean and variance across features, and γ, β are learnable parameters. Unlike batch normalization, LayerNorm is insensitive to batch size and suits autoregressive decoding. Modern variants (RMSNorm, ScaleNorm) adjust normalization formulae to reduce parameter count or improve stability.

2.3 Feed-Forward Networks

Position-wise feed-forward networks (FFN) complement attention:

$$\mathrm{FFN}(\mathbf{x}) = \sigma(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2. \tag{8}$$

Gated linear units (GLU), SwiGLU, and GEGLU introduce multiplicative gating, yielding better throughput-accuracy trade-offs in large models.

3 Positional Encoding Strategies

Since attention is permutation-invariant, positional signals inject order information.

3.1 Sinusoidal Encoding

Original transformers employed deterministic sinusoids:

$$\mathrm{PE}(t, 2i) = \sin\left(\frac{t}{10000^{2i/d_{\text{model}}}}\right), \tag{9}$$

$$\mathrm{PE}(t, 2i+1) = \cos\left(\frac{t}{10000^{2i/d_{\text{model}}}}\right). \tag{10}$$

These encodings generalize to unseen lengths and allow relative distance computation via phase differences.

3.2 Rotary Position Embeddings (RoPE)

RoPE rotates query/key vectors by position-dependent angles. For complex embedding $\mathbf{z}_t$, multiplication by a rotation matrix yields

$$\mathrm{RoPE}(\mathbf{z}_t) = \mathbf{z}_t \cdot e^{i\theta_t}. \tag{11}$$

RoPE preserves relative offsets inside inner products, enabling linear extrapolation and efficient extrapolation to longer contexts. Implementations operate on real-valued pairs using rotation matrices.

3.3 ALiBi Bias

Attention with Linear Biases (ALiBi) introduces a slope-based bias to the attention logits:

$$\mathrm{Score}(i, j) = \frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d_k}} - m_h(i - j), \tag{12}$$

where m_h is a head-specific slope. ALiBi extends context length without additional memory and improves extrapolation by penalizing distant positions in a head-dependent manner.

4 Encoder and Decoder Structures

4.1 Encoder-Decoder (Seq2Seq)

Encoder-decoder transformers consist of an encoder stack producing latent representations $\mathbf{H}_{\text{enc}}$ and a decoder stack generating outputs conditionally. The decoder contains masked self-attention, cross-attention, and FFN layers. Such architecture excels in translation, summarization, and multi-modal fusion.

4.2 Decoder-Only Models

Decoder-only transformers (GPT-class models) drop cross-attention and rely solely on masked self-attention. Their simplicity suits large-scale autoregressive training, and the resulting models act as general-purpose sequence engines via prompting, in-context learning, and alignment fine-tuning.

4.3 Encoder-Only Models

Encoder-only architectures (BERT, RoBERTa, DeBERTa) retain bidirectional self-attention without decoding modules. They produce contextual embeddings for understanding tasks and serve as feature extractors or fine-tuned encoders.

4.4 Mixture and Hybrid Designs

Modern systems mix components: retrieval-augmented generators add encoder modules to decoder-only backbones; encoder-decoder models sometimes share weights or use prefix tuning for efficient adaptation. Figure ?? compares structural motifs across transformer families.

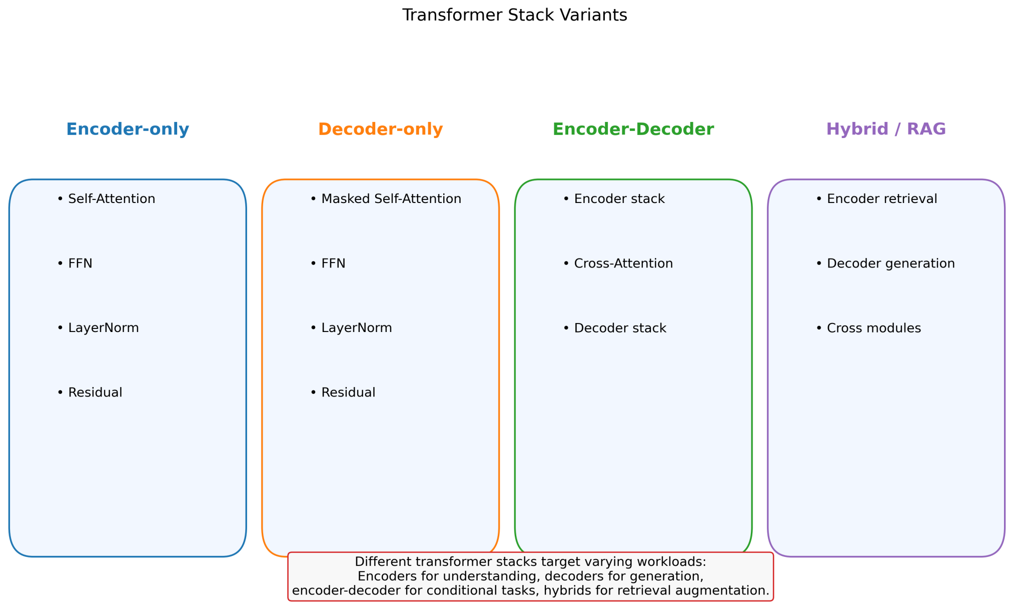


Figure 2: Structural variants of transformer stacks: encoder-only, decoder-only, encoder-decoder, and hybrid retrieval-augmented designs.

5 Implementation Notes

- **Scaling laws:** Depth, width, and head count scale with compute budget; post-norm vs. pre-norm choices impact convergence.
- **Regularization:** Dropout, stochastic depth, gradient noise, and weight decay stabilize large models.
- **Hardware:** FlashAttention, tensor parallelism, and quantization reduce memory and latency when deploying transformer blocks.

Further Reading

- Vaswani et al. “Attention is All You Need.” NeurIPS 2017.
- Shazeer. “Fast Transformer Decoding: One Write-Head is All You Need.” 2019.
- Xiong et al. “On Layer Normalization in the Transformer Architecture.” ICML 2020.
- Press et al. “Train Short, Test Long: Attention with Linear Biases.” ICLR 2022.
- Dao et al. “FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness.” NeurIPS 2022.