Budapest Semesters in Mathematics
Directed Reading Program (Autumn 2025)

A Mathematical Description of Transformers via Kronecker Products

Vasudev Menon
Teaching Assistant: Attila Jung

1 Introduction

Transformer [5] models are a popular neural network architecture suited for a variety of tasks, most famously in OpenAI's GPT large language model (LLM) [2]. As these models scale, they become increasingly difficult to understand and interpret, especially in regards to harmful, unexpected, and often confusing behavior that they may exhibit. To remediate these issues, the field of mechanistic interpretability aims to reveal how individual components of a model contribute to its overall behavior, which is critical for ensuring AI alignment and safety [1].

In this project, we will explain the framework introduced by Elhage et al. [3] of Anthropic, which provides an introductory mathematical formulation of transformer circuits and decomposes their operations into interpretable linear and nonlinear components. By analyzing transformers in this way, we can develop a more precise mathematical understanding of their inner workings.

Objective 1. Our objective is to express the computations performed by a transformer layer using standard linear algebraic operations, then reformulate them in terms of the Kronecker product for compactness and structural clarity.

2 Preliminaries

Transformers are neural network architectures designed to process sequences of token embeddings by combining *self-attention* and *feedforward networks (MLPs, i.e., multilayer perceptrons)*. Let

$$X = (x_1, x_2, \ldots, x_n) \in (\mathbb{R}^{d_{\text{model}}})^n$$

denote a sequence of n token embeddings, where each $x_i \in \mathbb{R}^{d_{\text{model}}}$ represents the embedding of the i-th token.

A standard transformer layer consists of:

(i) **Self-attention:** allows each token to attend to all other tokens in the sequence, producing a context-aware representation.

(ii) **MLP:** a positionwise nonlinear transformation applied independently to each token.

Formally, a single layer can be expressed as

$$T(X) = \text{MLP}\left(\text{Attn}(X)\right),$$

where

$$\text{Attn}(X) = X + W_O\Big[\,\text{softmax}\left((W_Q X)(W_K X)^\top/\sqrt{d_k}\right)(W_V X)\Big].$$

Here, W_Q, W_K, W_V, W_O are learned weight matrices, and the addition of X represents a residual (or "skip") connection[1]. The full transformer architecture also includes multiple attention heads, layer normalization, and an unembedding layer for producing token logits.

[1] where the original input x is added elementwise back to the output $(F(x) + x)$

Each attention head is an independent self-attention mechanism parameterized by its own projection matrices (W_Q, W_K, W_V, W_O). Conceptually, each head attends to a different subspace of the embedding, and their outputs are concatenated and linearly combined to produce the layer output.

For simplicity, our derivations focus on the mathematics of a single head, which captures the essential structure of self-attention. For a complete description of the standard transformer, including these omitted components, we refer the reader to Vaswani et al. [5] and Brown et al. [2].

For the purposes of this paper, we focus on a *simplified, attention-only transformer* that preserves the essential structure of self-attention while remaining mathematically tractable. In particular, we make the following assumptions:

1. **No MLPs:** We omit MLPs to isolate attention dynamics.
2. **Linear maps only:** All affine transformations are linear; biases can be absorbed by augmenting input vectors with a constant component[2].
3. **No layer normalization:** Normalization can be approximately absorbed into adjacent linear maps.
4. **Single attention head and no scaling:** We drop the $\sqrt{d_k}$ factor and restrict to a single head.

Definition 1 (Simplified Transformer)**.** Under these assumptions, a single simplified transformer layer is

$$\mathrm{Attn}(X) = X + W_O A(X), \quad A(X) = \mathrm{softmax}\big((W_Q X)(W_K X)^\top\big) W_V X.$$

An L-layer simplified transformer is the composition

$$T^{(L)}(X) = T_L \circ T_{L-1} \circ \cdots \circ T_1(X).$$

This minimal model captures the core structure of self-attention and is sufficient for the Kronecker product formulation discussed in this paper.

3 Kronecker Product

We leverage the Kronecker product [4] operation to express attention matrix computations in a concise and elegant manner.

Definition 2 (Kronecker Product)**.** Let $A = [a_{ij}] \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$. The *Kronecker product* of A and B, denoted $A \otimes B$, is defined as

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix} \in \mathbb{R}^{mp\times nq}.$$

In words, $A \otimes B$ is the block matrix obtained by replacing each entry a_{ij} of A with the matrix $a_{ij}B$.

Remark 3. When applying Kronecker products to matrices (as opposed to vectors), we interpret them as column-wise operators acting jointly on both dimensions. Concretely, for matrices P and Q of compatible sizes, $(P \otimes Q)$ acts by applying Q along the feature dimension (rows) and P along the sequence dimension (columns). This convention aligns with the identity proved later in Property 5, and will be used consistently throughout.

[2]Consider an affine transformation $x \mapsto Ax + b$ with weight A and bias b. We "fold" the bias b into A by defining $A' = \left(\begin{array}{c|c} A & b \\ \hline 0\ldots0 & 1 \end{array}\right)$ and letting $x' = \begin{bmatrix} x \\ 1 \end{bmatrix}$. We then have $A'x' = \begin{bmatrix} Ax+b \\ 1 \end{bmatrix}$, which holds in general. Thus, we can recover the same affine transformation without explicitly needing the vector b.

Example 4. The following is a brief visual example of the Kronecker Product. Let

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 5 \\ 6 & 7 \end{bmatrix}.$$

Then

$$A \otimes B = \begin{bmatrix} 1B & 2B \\ 3B & 4B \end{bmatrix} = \begin{bmatrix} 0 & 5 & 0 & 10 \\ 6 & 7 & 12 & 14 \\ 0 & 15 & 0 & 20 \\ 18 & 21 & 24 & 28 \end{bmatrix}.$$

In fact, there are some elegant properties we can use to further simplify our mathematical formulation of attention.

Property 5 (Mixed-Product Property)**.** *Let $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p\times q}$, $C \in \mathbb{R}^{n\times r}$, and $D \in \mathbb{R}^{q\times s}$. Then*

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \in \mathbb{R}^{mp\times rs}.$$

Proof. We interpret the Kronecker product as a block matrix and compute using block multiplication rules. By definition,

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix}, \qquad C \otimes D = \begin{bmatrix} c_{11}D & c_{12}D & \cdots & c_{1r}D \\ c_{21}D & c_{22}D & \cdots & c_{2r}D \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1}D & c_{n2}D & \cdots & c_{nr}D \end{bmatrix}.$$

The (i,k)-th block of $(A \otimes B)(C \otimes D)$ is

$$\begin{aligned} \big[(A \otimes B)(C \otimes D)\big]_{ik} &= \sum_{j=1}^{n} (a_{ij}B)(c_{jk}D) \\ &= \sum_{j=1}^{n} a_{ij}c_{jk}(BD) \\ &= (AC)_{ik}(BD). \end{aligned}$$

Hence the result is

$$(A \otimes B)(C \otimes D) = \begin{bmatrix} (AC)_{11}(BD) & (AC)_{12}(BD) & \cdots & (AC)_{1r}(BD) \\ (AC)_{21}(BD) & (AC)_{22}(BD) & \cdots & (AC)_{2r}(BD) \\ \vdots & \vdots & \ddots & \vdots \\ (AC)_{m1}(BD) & (AC)_{m2}(BD) & \cdots & (AC)_{mr}(BD) \end{bmatrix},$$

which, by definition, equals $(AC) \otimes (BD)$. □

Property 6 (Standard Matrix Identity)**.** *For conformable matrices M, X, N,*

$$MXN = (I \otimes M)\,(N^{\top} \otimes I)\,X,$$

where each Kronecker product acts naturally on the row and column spaces of X as indicated. Equivalently, the combined action of left-multiplying by M and right-multiplying by N is represented by the Kronecker product $N^{\top} \otimes M$.

Proof. We verify that both sides act identically on the standard basis.

Let E_{ij} denote the elementary matrix with a 1 in position (i, j) and zeros elsewhere. Then any X can be written as $X = \sum_{i,j} X_{ij} E_{ij}$. Hence it suffices to check the identity for E_{ij}.

Compute the left-hand side:

$$ME_{ij}N = M(e_i e_j^\top)N = (Me_i)(N^\top e_j)^\top.$$

This produces a rank-one matrix whose columns are scalar multiples of Me_i, and whose column scalars are given by entries of $N^\top e_j$.

Now examine the right-hand side. By the defining property of the Kronecker product,

$$(N^\top \otimes M)\, E_{ij}$$

acts by the same rule: it multiplies each column of E_{ij} by the corresponding scalar from $N^\top e_j$, and replaces the standard basis vector e_i with its image Me_i. Thus the two expressions coincide entrywise for every i, j.

Since the equality holds on all basis elements E_{ij} and both sides are linear in X, it holds for all X. □

Claim 7. *Let $X \in \mathbb{R}^{d\times n}$, and let*

$$W_V \in \mathbb{R}^{d_k\times d}, \qquad W_O \in \mathbb{R}^{d\times d_k}, \qquad A \in \mathbb{R}^{n\times n}$$

be the usual value, output and attention matrices for a single head. Define

$$V := W_V X \in \mathbb{R}^{d_k\times n}, \qquad R := VA^\top \in \mathbb{R}^{d_k\times n}, \qquad H := W_O R \in \mathbb{R}^{d\times n}.$$

Then the head output H can be written as a single Kronecker action on X:

$$H \;=\; (A \otimes (W_O W_V))\, X$$

where the operator $A \otimes (W_O W_V)$ acts on matrices by the rule $(P \otimes Q)\, X = Q\, X\, P^\top$. Equivalently (reading the ordinary matrix product),

$$H \;=\; W_O W_V X A^\top.$$

Proof. First note the three elementary operator identities (all verified columnwise):

- $(I_n \otimes W_V)X = W_V X = V$ (apply W_V to each column of X),
- $(A \otimes I_{d_k})V = VA^\top = R$ (mix columns of V according to A),
- $(I_n \otimes W_O)R = W_O R = H$ (apply W_O to each column of R).

Chaining these three operators gives the representation

$$H = (I_n \otimes W_O)\,(A \otimes I_{d_k})\,(I_n \otimes W_V)\,X.$$

Now apply Property 5:

$$(P_1 \otimes Q_1)(P_2 \otimes Q_2) = (P_1 P_2) \otimes (Q_1 Q_2).$$

Using this twice,

$$(I_n \otimes W_O)(A \otimes I_{d_k})(I_n \otimes W_V) = (I_n A I_n) \otimes (W_O I_{d_k} W_V) = A \otimes (W_O W_V).$$

Therefore

$$H = (A \otimes (W_O W_V))\, X,$$

which is exactly the identity. Writing this back in ordinary matrix multiplication yields the equivalent form $H = W_O W_V\, X\, A^\top$, as claimed. □

Finally, recall from Definition 1 that

$$\mathrm{Attn}(X) = X + W_O \,\mathrm{softmax}\big((W_Q X)(W_K X)^\top\big)\, W_V X.$$

Denoting

$$A := \mathrm{softmax}\big((W_Q X)(W_K X)^\top\big)\,,$$

our previous derivation shows that this transformation can be expressed as a single Kronecker operator acting on X:

$$H = (A \otimes (W_O W_V))\, X.$$

Thus, the complete attention formula can be equivalently written as

$$\mathrm{Attn}(X) = X + (A \otimes W_O W_V)X.$$

where A governs interactions across the sequence dimension and $W_O W_V$ acts across the feature dimension. The Kronecker product $(A \otimes W_O W_V)$ therefore captures the separable yet coupled nature of attention, a key mathematical insight for interpretability.

References

[1] L. Bereska and E. Gavves. Mechanistic interpretability for ai safety – a review, 2024.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

[3] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

[4] A. Graham. *Kronecker products and matrix calculus*. Mathematics and its Applications. Ellis Horwood Ltd, Publisher, Harlow, England, Nov. 1981.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.