15-780: Graduate AI
Lecture 10: transformers continued

Aditi Raghunathan

Recap of self-attention

$$\underset{\sim}{X} = \begin{bmatrix} \leftarrow & x_1 & \rightarrow \\ \leftarrow & x_2 & \rightarrow \\ & \vdots & \\ \leftarrow & x_T & \rightarrow \end{bmatrix}$$

$\underset{\sim}{X} W$: treat rows independently

"batching"

$A \underset{\sim}{X}$

$A(\underset{\sim}{X}) \; \underset{\sim}{X}$

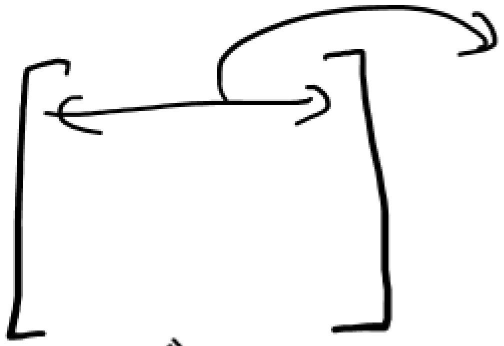
Recap of self-attention

$$A = \text{softmax}\left(\frac{\overbrace{X W_Q}^{Q}\,\overbrace{W_K^T X^T}^{K}}{\sqrt{d}}\right)$$

$Y \in \mathbb{R}^{T \times d}$ $\quad X \in \mathbb{R}^{T \times d}$

$$Y = (A)\, X W_V$$

colored quantities are "learnt" (by SGD)



A

one set of linear coefficients

$x \Leftarrow e_{the} W_Q W_K^T e_{quick}$

– the quick brown fox

– the fox brown quick

Properties of self-attention

- Order invariant: permutations don't change the mixing
- Multi-head attention
- Full mixing: every word can depend on everything else , masking

Position embeddings

$$T \times d \quad \begin{bmatrix} X \end{bmatrix} + \begin{bmatrix} \\ E \\ \\ \end{bmatrix}$$

E_{pos} depends on pos

$$A(X) \longrightarrow A(X+E) \neq A(X) + E$$

$$(X+E)\, W_Q\, W_K^T\, (X+E)^T$$

Position embeddings

→ E_{pos} : $U_{pos} \in \mathbb{R}^d$ that is learnt

→ "sine embeddings"

$$E_{pos} = \begin{bmatrix} \sin c_1 \cdot pos \\ \cos c_1 \cdot pos \\ \sin c_2 \cdot pos \\ \cos c_2 \cdot pos \\ \vdots \end{bmatrix} \qquad c_i = \left(\frac{1}{10{,}000} \right)^{i/d}$$

for $pos = 1 \ldots T$

Interpretation of position embedding

For now, $W_Q, W_K \equiv$ identity

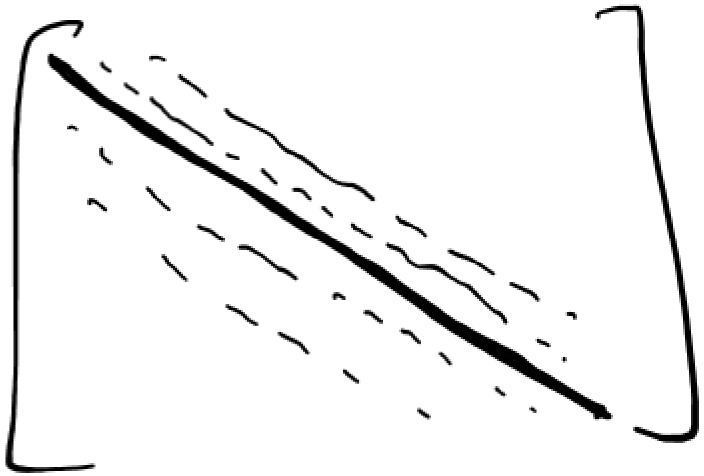
$$A(X) = (X+E)(X+E)^T = XX^T + \cancel{XE^T} + \cancel{EX^T} + EE^T$$

original attention term

can ignore

newly added

$EE^T =$



large at diagonal and slowly decay

Relative embeddings

- $A(x) = x x^T + E E^T$

  ↑ write the dot product directly

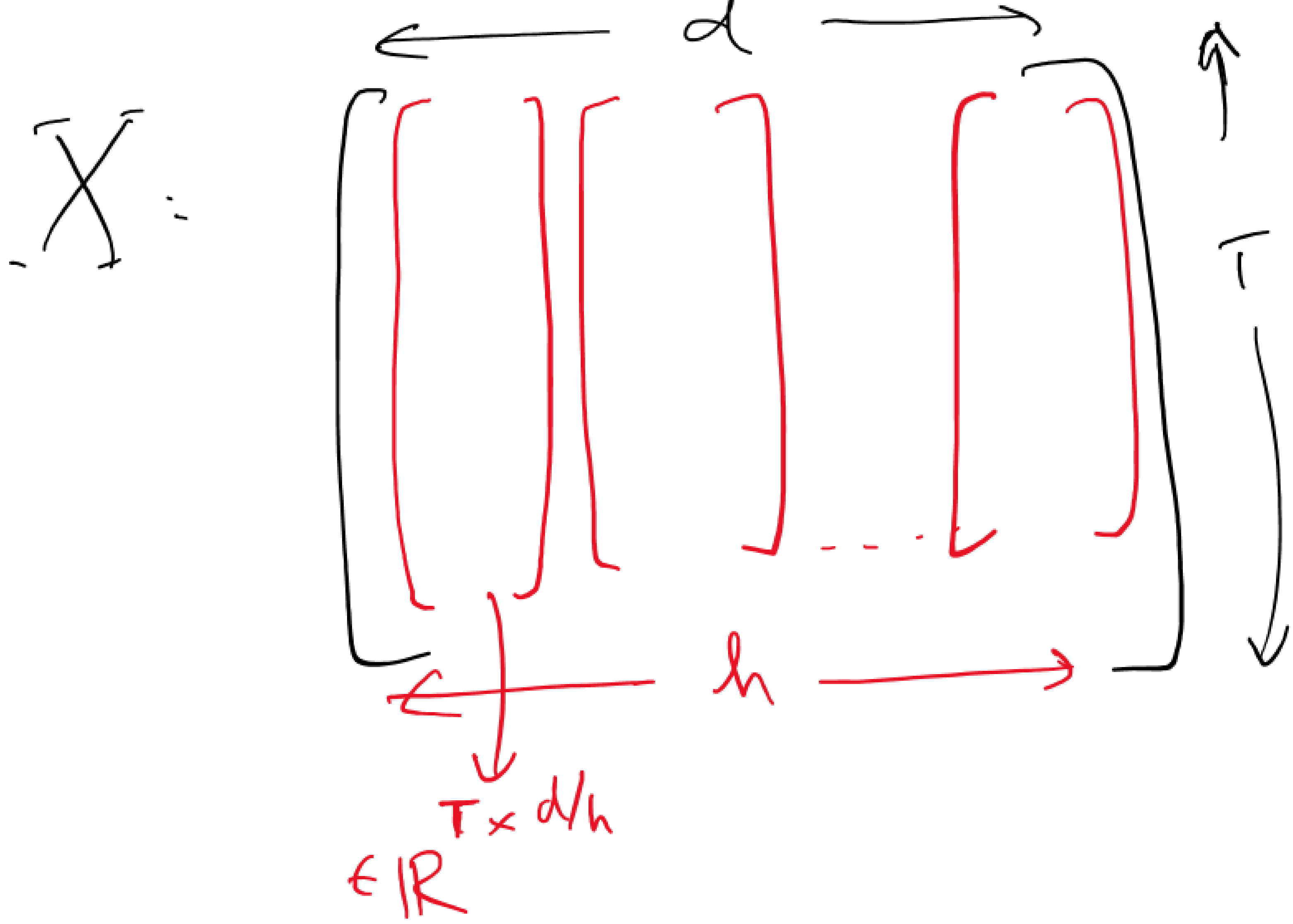
  $\rightarrow (E_i)^T (E_j) = \alpha_{ij} \rightarrow$ learnt & depends on $(j-i)$

- "rotary embeddings": E matrix that encodes relative position

  $\cancel{x} E$ (multiplied)

Back to self-attention

$Y = AX \qquad A = \mathrm{softmax}\left(\frac{X W_Q W_K^T X^T}{\sqrt{d}}\right)$



h diff blocks

Back to self-attention

$$Y = A X = \begin{bmatrix} \\ A \\ \\ \end{bmatrix} \begin{bmatrix} \underbrace{\begin{bmatrix} \\ \\ \\ \end{bmatrix}}_{\tilde{X}^{(1)}} & \underbrace{\begin{bmatrix} \\ \\ \\ \end{bmatrix}}_{\tilde{X}^{(2)}} & \cdots & \end{bmatrix}_{\tilde{X}^{(h)}}$$

$$\tilde{X}^{(1)} \quad \tilde{X}^{(2)} \quad \cdots \quad \tilde{X}^{(h)}$$

$Y = A\tilde{X}$ mixes the rows

Same mixing vs different mixing

$$\begin{bmatrix} \lambda & (1-\lambda) \\ \beta & (1-\beta) \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} \lambda a + (1-\lambda) c & \lambda b + (1-\lambda) d \\ \beta a + (1-\beta) c & \beta b + (1-\beta) d \end{bmatrix}$$

$$X = \begin{bmatrix} [\quad][\quad]\cdots\cdots \end{bmatrix}$$

h diff blocks

Piazza poll

Select the true statements under the standard self-attention formulation Y = A X:

(A) All blocks share the same linear combination or mixing

(B) Different blocks can have different mixing

we want this in practice

Multi-head attention

$K : W_K X$

$Q : W_Q X$

$$X = \begin{bmatrix} \underbrace{\begin{bmatrix} \quad \end{bmatrix}}_{h} & \cdots \end{bmatrix}$$

$$W_K, W_Q, W_V = \begin{bmatrix} \underbrace{\begin{bmatrix} \quad \end{bmatrix}}_{h} & \end{bmatrix}$$

$$Y = \begin{bmatrix} \tilde{A}^{(1)} \tilde{X}^{(1)}, & \tilde{A}^{(2)} \tilde{X}^{(2)} & \cdots \end{bmatrix}$$

can be different!

Multi-head attention

W_V: could be of higher dimensions

W_O: combines the different heads in some way

$$Y = \underbrace{A(x) \times W_V}_{\text{multi-head}} W_O$$

Transformer block

W_K, W_Q, W_V, W_O

$$\begin{cases} Y = \text{LayerNorm}\left(X + \text{M.SelfA}(X)\right) \\ Z = \text{LayerNorm}\left(\cancel{Y} + \sigma(\cancel{Y} W_1) W_2\right) \end{cases}$$

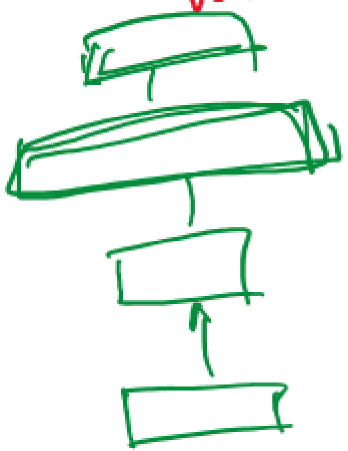
$Y \in \mathbb{R}^{T \times d_{model}}$

$W_1 \in \mathbb{R}^{d_{model} \times d_{ff}}$

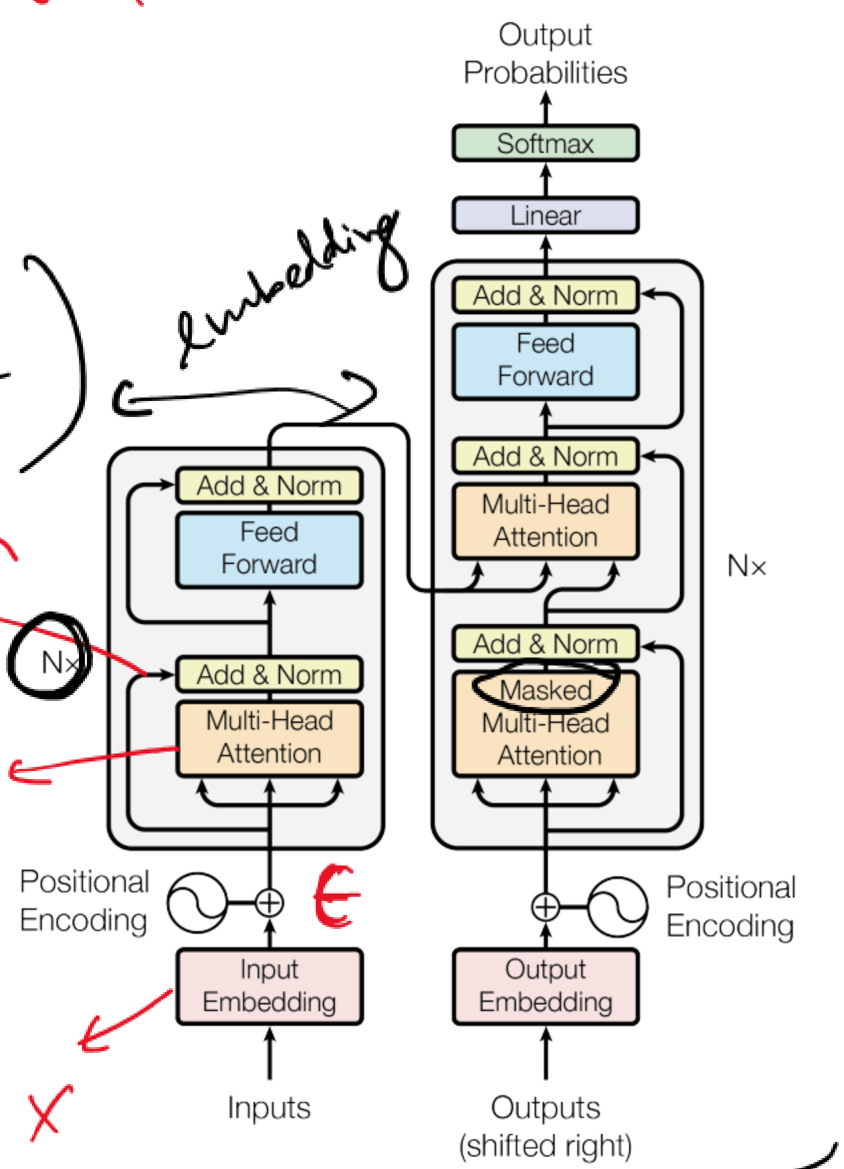
$W_2 \in \mathbb{R}^{d_{ff} \times d_{model}}$

residual & normalization

mixing



embedding



E

X

decoder part

GPT-3

- Dimension of hidden state: d_model=12288
- Dimension of the intermediate feed-forward layer: d_ff=4d_model
- Number of heads: n_heads=96
- Context length: L=2048

Masking

- A way to set certain attention scores to zero

A: completely determined by W_Q, W_K, W_V, E, X

$\rightarrow$

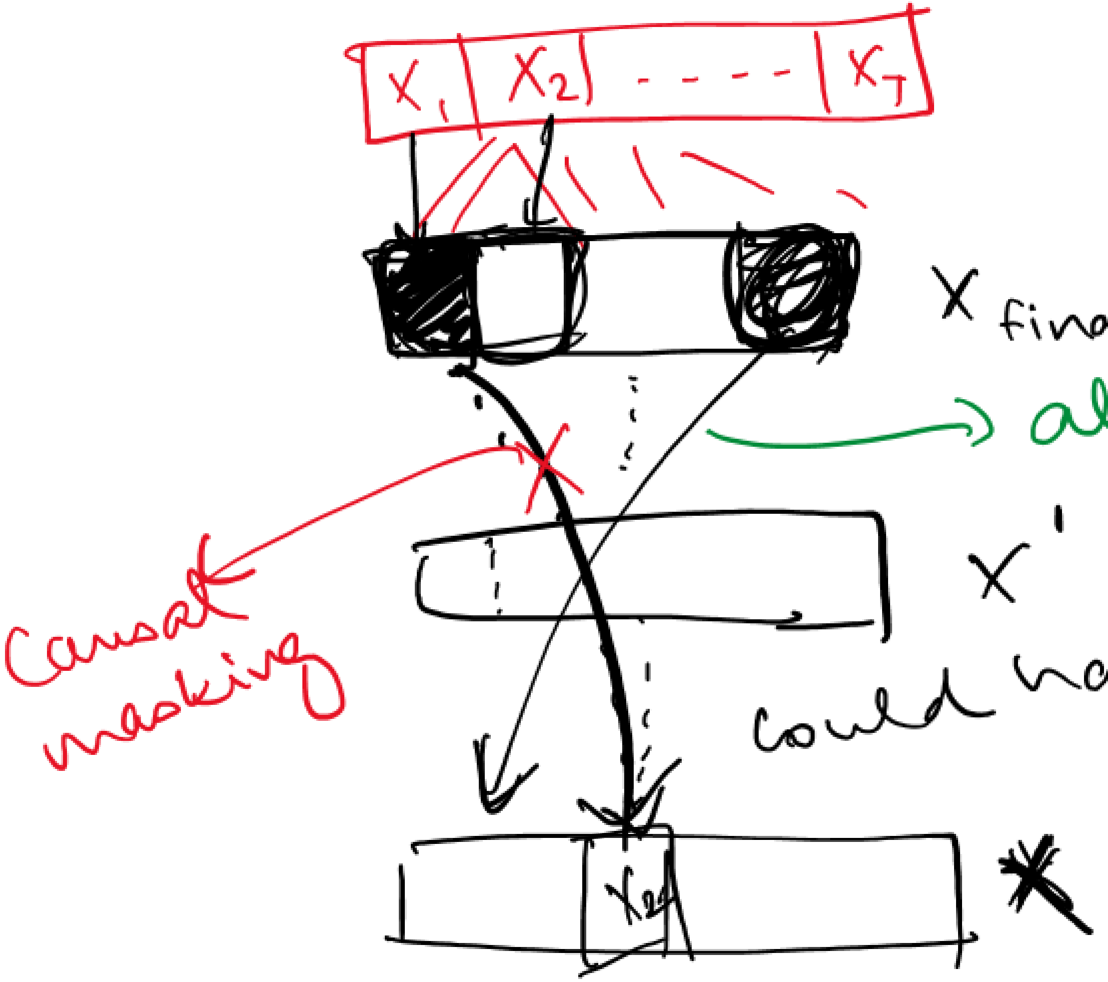
Masking

"Causal masking"

$P(x_i \mid x_1 \ldots {}_{-i})$

$Y = [x_1, x_2, \ldots, x_T]$

$X = [x_1 \ldots x_T]$



$\rightarrow$ contextual embedding

$X_{final} \in \mathbb{R}^{T \times d}$

$\rightarrow$ allowed

could have mixed information

$$A(x) \begin{bmatrix} \ddots & 0 \\ & \ddots \end{bmatrix}$$

$$A_{ij} = 0 \quad \text{if } j > i$$

$$A(x) = \text{softmax}\left(KQ^T + \begin{bmatrix} \ddots & -\infty \\ 0 & \ddots \end{bmatrix} \right)$$

Summary

- You know how to build GPT3!

train by minimizing loss via Adam

→ MLPs

→ Add Norm

→ Self Attention

multihead

masking

→ position embeddings