A. Liang
Transformer
March 21, 2023

Transformer Attention Derivative

Question 1. Attention Derivative W^O

Before starting our calculation, we should know what's attention is used in The Transformer (Vaswani et al.):

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

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, ..., \text{head}_\text{H})W^O \\ \text{where head}_\text{h} &= \text{Attention}(QW_h^Q, KW_h^K, VW_h^V) \end{aligned}$$

$$\begin{aligned} df(X) &= \text{tr}(\frac{\partial f(X)}{\partial X}^T d(\text{Concat}(\text{head}_1, ..., \text{head}_\text{H})W^O)) \\ &= \text{tr}(\frac{\partial f(X)}{\partial X}^T \text{Concat}(\text{head}_1, ..., \text{head}_\text{H})dW^O) \end{aligned}$$

$$\frac{\partial f(X)}{\partial W^O} = \text{Concat}(\text{head}_1, ..., \text{head}_\text{H})^T \frac{\partial f(X)}{\partial X}$$

Question 2. Attention Derivative W^v

$$\begin{aligned} df(X) &= \text{tr}\left(\frac{\partial f(X)}{\partial X}^T dX\right) \\ &= \text{tr}\left(\frac{\partial f(X)}{\partial X}^T d\text{Concat}(\text{head}_1, ..., \text{head}_\text{H})W^O\right) \\ &= \text{tr}\left(\frac{\partial f(X)}{\partial X}^T d(\text{softmax}(\frac{QW_1^Q(W_1^K)^TK^T}{\sqrt{d_k}})VW_1^V, ..., \text{softmax}(\frac{QW_H^Q(W_H^K)^TK^T}{\sqrt{d_k}})VW_H^V)W^O\right) \\ &= \text{tr}\left(W^O\frac{\partial f(X)}{\partial X}^T d(\text{softmax}(\frac{QW_1^Q(W_1^K)^TK^T}{\sqrt{d_k}})V, ..., \text{softmax}(\frac{QW_H^Q(W_H^K)^TK^T}{\sqrt{d_k}})V)\text{diag}(W_1^V, ..., W_H^V)\right) \end{aligned}$$

let

$$A = \left(\text{softmax}(\frac{QW_1^Q(W_1^K)^TK^T}{\sqrt{d_k}})V, ..., \text{softmax}(\frac{QW_H^Q(W_H^K)^TK^T}{\sqrt{d_k}})V\right)$$

$$W^v = \begin{bmatrix} W_1^V & & \\ & \ddots & \\ & & W_H^V \end{bmatrix}$$

We can get this formula:

$$
\begin{aligned}
df(X) &= \operatorname{tr}\left(W^O \frac{\partial f(X)}{\partial X}^T d(AW^v))\right) \\
&= \operatorname{tr}\left(W^O \frac{\partial f(X)}{\partial X}^T AdW^v)\right) \\
&= \operatorname{tr}\left((A^T \frac{\partial f(X)}{\partial X}(W^O)^T)^T dW^v)\right)
\end{aligned}
$$

So

$$
\frac{\partial f(X)}{\partial W^v} = A^T \frac{\partial f(X)}{\partial X}(W^O)^T
$$

Question 3. Attention Derivative W^K, W^Q

$$
\begin{aligned}
df(X) &= \operatorname{tr}\left(\frac{\partial f(X)}{\partial X}^T dX\right) \\
&= \operatorname{tr}\left(\frac{\partial f(X)}{\partial X}^T d\mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head}_\mathrm{H})W^O\right) \\
&= \operatorname{tr}\left(\frac{\partial f(X)}{\partial X}^T d(\mathrm{softmax}(\frac{QW_1^Q(W_1^K)^T K^T}{\sqrt{d_k}})VW_1^V, ..., \mathrm{softmax}(\frac{QW_H^Q(W_H^K)^T K^T}{\sqrt{d_k}})VW_H^V)W^O\right) \\
&= \operatorname{tr}\left(W^O \frac{\partial f(X)}{\partial X}^T d(\mathrm{softmax}(\frac{QW_1^Q(W_1^K)^T K^T}{\sqrt{d_k}}), ..., \mathrm{softmax}(\frac{QW_H^Q(W_H^K)^T K^T}{\sqrt{d_k}}))\mathrm{diag}(VW_1^V, ..., VW_H^V)\right)
\end{aligned}
$$

let

$$
\begin{aligned}
S &= (\mathrm{softmax}(A_1), ..., \mathrm{softmax}(A_H)) \\
&= \left(\mathrm{softmax}(\frac{QW_1^Q(W_1^K)^T K^T}{\sqrt{d_k}}), ..., \mathrm{softmax}(\frac{QW_H^Q(W_H^K)^T K^T}{\sqrt{d_k}})\right) \\
A &= (A_1, ..., A_H)
\end{aligned}
$$

$$
V' = \begin{bmatrix} VW_1^V & & \\ & \ddots & \\ & & VW_H^V \end{bmatrix}_{Hn \times d_{model}}
$$

$$
\begin{aligned}
df(X) &= \operatorname{tr}\left(W^O \frac{\partial f(X)}{\partial X}^T d(SV'))\right) \\
&= \operatorname{tr}\left(V' W^O \frac{\partial f(X)}{\partial X}^T dS)\right) \\
&= \operatorname{tr}\left(V' W^O \frac{\partial f(X)}{\partial X}^T d(\exp A \odot \Upsilon(\exp A\mathbb{I}))\right)
\end{aligned}
$$

where

$$\mathbb{I} = \begin{bmatrix} I & & \\ & \ddots & \\ & & I \end{bmatrix}_{Hn \times Hn}$$

$$I = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{n \times n}$$

$$\Upsilon(X_{m \times n}) = \begin{bmatrix} 1/x_{11} & \cdots & 1/x_{1n} \\ \vdots & \ddots & \vdots \\ 1/x_{m1} & \cdots & 1/x_{mn} \end{bmatrix}_{m \times n}$$

$$\begin{aligned}
df(X) &= \mathrm{tr}\left(V'W^O\frac{\partial f(X)}{\partial X}^T d(\exp A \odot \Upsilon(\exp A\mathbb{I}))\right) \\
&= \mathrm{tr}\left(V'W^O\frac{\partial f(X)}{\partial X}^T (d(\exp A) \odot \Upsilon(\exp A\mathbb{I}) + \exp A \odot d\Upsilon(\exp A\mathbb{I}))\right) \\
&= \mathrm{tr}\left(V'W^O\frac{\partial f(X)}{\partial X}^T (\exp A \odot \Upsilon(\exp A\mathbb{I}) \odot dA - \exp A \odot \Upsilon(\exp A\mathbb{I}) \odot \Upsilon(\exp A\mathbb{I}) \odot d(\exp A\mathbb{I}))\right) \\
&= \mathrm{tr}\left(V'W^O\frac{\partial f(X)}{\partial X}^T (S \odot dA - S \odot \Upsilon(\exp A\mathbb{I}) \odot d(\exp A\mathbb{I}))\right) \\
&= \mathrm{tr}\left(\left(\left(\frac{\partial f(X)}{\partial X}(W^O)^T(V')^T\right) \odot S\right)^T dA - \left(\left(\frac{\partial f(X)}{\partial X}(W^O)^T(V')^T\right) \odot S \odot \Upsilon(\exp A\mathbb{I})\right)^T (\exp A \odot dA)\mathbb{I}\right) \\
&= \mathrm{tr}\left(\left(\left(\frac{\partial f(X)}{\partial X}(W^O)^T(V')^T\right) \odot S\right)^T dA - \mathbb{I}\left(\left(\frac{\partial f(X)}{\partial X}(W^O)^T(V')^T\right) \odot S \odot \Upsilon(\exp A\mathbb{I})\right)^T (\exp A \odot dA)\right) \\
&= \mathrm{tr}\left(\left(\left(\frac{\partial f(X)}{\partial X}(W^O)^T(V')^T\right) \odot S\right)^T dA\right) \\
&- \mathrm{tr}\left(\left(\left(\left(\left(\frac{\partial f(X)}{\partial X}(W^O)^T(V')^T\right) \odot S \odot \Upsilon(\exp A\mathbb{I})\right)\mathbb{I}^T\right) \odot \exp A\right)^T dA\right)
\end{aligned}$$

So we can get

$$\frac{\partial f(X)}{\partial A} = \left(\frac{\partial f(X)}{\partial X}(W^O)^T(V')^T\right) \odot S - \left(\left(\left(\frac{\partial f(X)}{\partial X}(W^O)^T(V')^T\right) \odot S \odot \Upsilon(\exp A\mathbb{I})\right)\mathbb{I}^T\right) \odot \exp A$$

$$\begin{aligned}
df(X) &= \mathrm{tr}\left(\frac{\partial f(X)}{\partial A}^T dA\right) \\
&= \mathrm{tr}\left(\frac{\partial f(X)}{\partial A}^T dA\right) \\
&= \mathrm{tr}\left(\gamma\frac{\partial f(X)}{\partial A}^T d\left(QW_1^Q(W_1^K)^T K^T, ..., QW_H^Q(W_H^K)^T K^T\right)\right)
\end{aligned}$$

where

$$\gamma = \frac{1}{\sqrt{d_k}}$$

let

$$\mathbb{P}_h = \left(..., E^h_{n\times n}, ...\right)_{n\times nH}$$

E is the $n \times n$ identity matrix which is located in the h-th column, while other entries are zero matrices.

$$\begin{aligned} df(X) &= \operatorname{tr}\left(\gamma \frac{\partial f(X)}{\partial A}^T d\left(QW_h^Q (W_h^K)^T K^T \mathbb{P}_h\right)\right) \\ &= \operatorname{tr}\left(\gamma K^T \mathbb{P}_h \frac{\partial f(X)}{\partial A}^T Qd\left(W_h^Q (W_h^K)^T\right)\right) \\ &= \operatorname{tr}\left(\gamma (W_h^K)^T K^T \mathbb{P}_h \frac{\partial f(X)}{\partial A}^T Qd\left(W_h^Q\right)\right) \end{aligned}$$

$$\frac{\partial f(X)}{\partial W_h^Q} = \gamma Q^T \frac{\partial f(X)}{\partial A} \mathbb{P}_h^T K W_h^K$$

$$\begin{aligned} df(X) &= \operatorname{tr}\left(\gamma K^T \mathbb{P}_h \frac{\partial f(X)}{\partial A}^T Qd\left(W_h^Q (W_h^K)^T\right)\right) \\ &= \operatorname{tr}\left(\gamma K^T \mathbb{P}_h \frac{\partial f(X)}{\partial A}^T QW_h^Q d\left((W_h^K)^T\right)\right) \end{aligned}$$

$$\frac{\partial f(X)}{\partial W_h^K} = \gamma K^T \mathbb{P}_h \frac{\partial f(X)}{\partial A}^T QW_h^Q$$

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. URL `https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html`.

Tsinghua Shenzhen