TF 実装に当たってのレポート

23266053 志田光

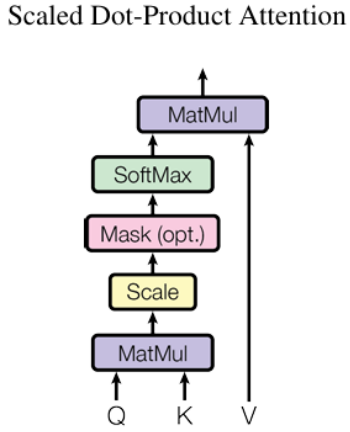
トランスフォーマー

日英コーパス

https://tatoeba.org/ja/downloads

逆伝播

scaled dot-product attention



行列積の微分

$$X \in \mathbb{R}^{2*3}, W \in \mathbb{R}^{3*4}, Y = XW$$

と設定し、

$$\frac{dY}{dW}$$

を計算する。この時、

$$\frac{dY}{dW} = \begin{pmatrix} \frac{\partial y_{11}}{\partial W} & \frac{\partial y_{12}}{\partial W} & \frac{\partial y_{13}}{\partial W} & \frac{\partial y_{14}}{\partial W} \\ \frac{\partial y_{21}}{\partial W} & \frac{\partial y_{22}}{\partial W} & \frac{\partial y_{23}}{\partial W} & \frac{\partial y_{24}}{\partial W} \end{pmatrix}$$

となり、さらに

$$\frac{\partial y_{11}}{\partial W} = \begin{pmatrix} \frac{\partial y_{11}}{\partial W_{11}} & \frac{\partial y_{11}}{\partial W_{12}} & \frac{\partial y_{11}}{\partial W_{13}} & \frac{\partial y_{11}}{\partial W_{14}} \\ \frac{\partial y_{11}}{\partial W_{21}} & \frac{\partial y_{11}}{\partial W_{22}} & \frac{\partial y_{11}}{\partial W_{23}} & \frac{\partial y_{11}}{\partial W_{24}} \\ \frac{\partial y_{11}}{\partial W_{31}} & \frac{\partial y_{11}}{\partial W_{32}} & \frac{\partial y_{11}}{\partial W_{33}} & \frac{\partial y_{11}}{\partial W_{34}} \end{pmatrix}$$

となる。よって、

$$\frac{dY}{dW} \in \mathbb{R}^{2*4*3*4}$$

となり、4 階テンソルとなることが分かる。

$$G = \frac{dL}{dY} \in \mathbb{R}^{2*4}$$

としたとき、

$$\frac{dL}{dW} = \frac{dL}{dY}\frac{dY}{dW}$$

で求められるはず。だが、これだと行列積が可能なサイズでない。よって、G を転置するしかない？

あだマール積なら完璧！ これを中身で見てみると、

$$\frac{dL}{dY} \circ \frac{dY}{dW} = \begin{pmatrix} \frac{\partial L}{\partial y_{11}} & \frac{\partial L}{\partial y_{12}} & \frac{\partial L}{\partial y_{13}} & \frac{\partial L}{\partial y_{14}} \\ \frac{\partial L}{\partial y_{21}} & \frac{\partial L}{\partial y_{22}} & \frac{\partial L}{\partial y_{23}} & \frac{\partial L}{\partial y_{24}} \end{pmatrix} \begin{pmatrix} \frac{\partial y_{11}}{\partial W} & \frac{\partial y_{12}}{\partial W} & \frac{\partial y_{13}}{\partial W} & \frac{\partial y_{14}}{\partial W} \\ \frac{\partial y_{21}}{\partial W} & \frac{\partial y_{22}}{\partial W} & \frac{\partial y_{23}}{\partial W} & \frac{\partial y_{24}}{\partial W} \end{pmatrix} = \begin{pmatrix} \frac{\partial L}{\partial y_{11}}\frac{\partial y_{11}}{\partial W} & \frac{\partial L}{\partial y_{12}}\frac{\partial y_{12}}{\partial W} & \frac{\partial L}{\partial y_{13}}\frac{\partial y_{13}}{\partial W} & \frac{\partial L}{\partial y_{14}}\frac{\partial y_{14}}{\partial W} \\ \frac{\partial L}{\partial y_{21}}\frac{\partial y_{21}}{\partial W} & \frac{\partial L}{\partial y_{22}}\frac{\partial y_{22}}{\partial W} & \frac{\partial L}{\partial y_{23}}\frac{\partial y_{23}}{\partial W} & \frac{\partial L}{\partial y_{24}}\frac{\partial y_{24}}{\partial W} \end{pmatrix}$$

合成関数の微分の定理（2 変数関数を 2 変数でそれぞれ偏微分）では同じ変数での偏微分結果をすべて足し合わせて求める。

つまり、

$$\begin{aligned} \frac{\partial L}{\partial w_{11}} = &\frac{\partial L}{\partial y_{11}}\frac{\partial y_{11}}{\partial w_{11}} + \frac{\partial L}{\partial y_{12}}\frac{\partial y_{12}}{\partial w_{11}} + \frac{\partial L}{\partial y_{13}}\frac{\partial y_{13}}{\partial w_{11}} + \frac{\partial L}{\partial y_{14}}\frac{\partial y_{14}}{\partial w_{11}} \\ &+ \frac{\partial L}{\partial y_{21}}\frac{\partial y_{21}}{\partial w_{11}} + \frac{\partial L}{\partial y_{22}}\frac{\partial y_{22}}{\partial w_{11}} + \frac{\partial L}{\partial y_{23}}\frac{\partial y_{23}}{\partial w_{11}} + \frac{\partial L}{\partial y_{24}}\frac{\partial y_{24}}{\partial w_{11}} \end{aligned}$$

このように、$\frac{dL}{dY} \circ \frac{dY}{dW}$の各要素である各 y 要素を W で偏微分した結果を足し合わせると、L を w の各要素で偏微分した結果が得られる。

すなわち、$G = \frac{dL}{dY}, K = \frac{dY}{dW}$とおくと、 $W \leftarrow W - \eta \sum_{i,j} (G \circ K)_{i,j}$ となる。

$$y_{i,j} = \sum_{k=1}^{3} (w_{k,i} x_{j,k})$$

$$\frac{dY}{dW} = \begin{pmatrix} \frac{\partial y_{11}}{\partial W} & \frac{\partial y_{12}}{\partial W} & \frac{\partial y_{13}}{\partial W} & \frac{\partial y_{14}}{\partial W} \\ \frac{\partial y_{21}}{\partial W} & \frac{\partial y_{22}}{\partial W} & \frac{\partial y_{23}}{\partial W} & \frac{\partial y_{24}}{\partial W} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} x_{11} & 0 & 0 & 0 \\ x_{12} & 0 & 0 & 0 \\ x_{13} & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & x_{11} & 0 & 0 \\ 0 & x_{12} & 0 & 0 \\ 0 & x_{13} & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & x_{11} & 0 \\ 0 & 0 & x_{12} & 0 \\ 0 & 0 & x_{13} & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & x_{11} \\ 0 & 0 & 0 & x_{12} \\ 0 & 0 & 0 & x_{13} \end{pmatrix} \\ \begin{pmatrix} x_{21} & 0 & 0 & 0 \\ x_{22} & 0 & 0 & 0 \\ x_{23} & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & x_{21} & 0 & 0 \\ 0 & x_{22} & 0 & 0 \\ 0 & x_{23} & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & x_{21} & 0 \\ 0 & 0 & x_{22} & 0 \\ 0 & 0 & x_{23} & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & x_{21} \\ 0 & 0 & 0 & x_{22} \\ 0 & 0 & 0 & x_{23} \end{pmatrix} \end{pmatrix}$$

この時、$\frac{dL}{dY}$とのあだマール積を求め、各要素の行列を足し合わせて 0 の成分を無視すると、

$$\frac{dL}{dY} \circ \frac{dY}{dW} = \begin{pmatrix} \frac{\partial L}{\partial y_{11}}x_{11} + \frac{\partial L}{\partial y_{21}}x_{21} & \frac{\partial L}{\partial y_{12}}x_{11} + \frac{\partial L}{\partial y_{22}}x_{21} & \cdots & \frac{\partial L}{\partial y_{14}}x_{11} + \frac{\partial L}{\partial y_{24}}x_{21} \\ \frac{\partial L}{\partial y_{11}}x_{12} + \frac{\partial L}{\partial y_{21}}x_{22} & \cdots & \ddots & \vdots \\ \frac{\partial L}{\partial y_{11}}x_{13} + \frac{\partial L}{\partial y_{21}}x_{23} & \frac{\partial L}{\partial y_{12}}x_{13} + \frac{\partial L}{\partial y_{22}}x_{23} & \cdots & \frac{\partial L}{\partial y_{14}}x_{13} + \frac{\partial L}{\partial y_{24}}x_{23} \end{pmatrix}$$

この行列は、$X^T G$の演算結果と同じことが分かる。すなわち、

$$\frac{dL}{dW} = \frac{dL}{dY} \circ \frac{dY}{dW} = X^T G$$

同じようにして、

$$\frac{dL}{dX} = \frac{dL}{dY} \circ \frac{dY}{dX} = GW^T$$