Lipschitz Parameterization for Residual Network

Jiarui

August 10, 2023

1 Notation

We use $0_{p:q}$ to denote a $p \times q$ real matrix filled with 0, one of the size subscript is dropped when its clear from the context (e.g stacking with another matrix).

$S_{c,n}$ represents a permutation matrix such that $S_{c,n}x = y$, where $x = [1, 2, 3.., n, ...1, 2, 3, ...n]$ which contains c repetitions of $[1, 2, 3, ..., n]$, $y = [1 \cdot 1_n^T, ..., n \cdot 1_n^T]^T$.

F_n represents the 2D DFT matrix such that $\hat{X} = F_n X F_n$ where $X \in \mathbb{R}^{n \times n}$ and $\hat{X}$ is the 2D fourier transform of X.

2 Preliminaries

Theorem 2.1 *If exists a non-negative diagonal* Λ *such that*

$$\begin{bmatrix} \gamma I - H^T H & -H^T G - W^T \Lambda \\ -G^T H - \Lambda W & 2\Lambda - G^T G \end{bmatrix} \succcurlyeq 0$$

then $h(x) = Hx + G\sigma(Wx + b)$ *is* $\sqrt{\gamma}$*-Lipschitz*

Definition 2.1 (Cayley Transform for Square Matrix) *Define the Cayley Transform for an arbitrary square matrix Cayley:*$\mathbb{C}^{m \times m} \to \mathbb{C}^{m \times m}$ *as*

$$Cayley(B) = (I_m - B + B^*)(I_m + B - B^*)^{-1}$$

Note that if B is real, then $B - B^*$ is real and skew-symmetric, then $Cayley(B)$ is an orthogonal matrix. By Proposition A.7 in [1], we define the Cayley Transform for an tall matrix

Definition 2.2 *For a tall matrix* $W = \begin{bmatrix} U \\ V \end{bmatrix} \in \mathbb{R}^{(m+n) \times m}, U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times m}$*, define the Cayley Transform of* W *as*

$$Cayley(W) = Cayley(\begin{bmatrix} W & 0_{:,n} \end{bmatrix}) \begin{bmatrix} I_m \\ 0_{n,:} \end{bmatrix} = \begin{bmatrix} (I_m - U + U^* - V^*V)(I_m + U - U^* + V^*V)^{-1} \\ -2V(I_m + U - U^* + V^*V)^{-1} \end{bmatrix}$$

it can be easily proved that $Cayley(W)$ is orthogonal (i.e $Cayley(W)^T Cayley(W) = I_m$) if W is real.

3 Fully Connected Layer

Theorem 3.1 *Let* $W_A \in \mathbb{R}^{nout \times nout}, W_B \in \mathbb{R}^{nin \times nout}, \lambda \in \mathbb{R}^{nout}$,

$$\begin{bmatrix} G^T \\ H^T \end{bmatrix} = Cayley(\begin{bmatrix} W_A \\ W_B \end{bmatrix})$$

$$\Lambda = diag(0.5 + \exp(\lambda))$$

$$W = -\sqrt{\gamma}\Lambda^{-1} G^T H$$

then $h(x) = \sqrt{\gamma}Hx + G\sigma(Wx + b)$ *is* $\sqrt{\gamma}$*-Lipschitz.*

4 Convolution Layer

By Corollary A.1.1 in [1], if $C \in \mathbb{R}^{(cout \cdot n^2) \times (cin \cdot n^2)}$ represents a 2D circular convolution, then it can be diagonalized as $\mathcal{F}_{cout,n^2} C \mathcal{F}^*_{cin,n^2} = D$, where $\mathcal{F}_{c,n^2} = S_{c,n^2}(I_c \bigotimes (F_n \bigotimes F_n))$, D is a block-diagonal matrix with n^2 blocks each with size $cin \times cout$

We can derive a lemma similar to Lemma A.4 in [1]

Lemma 4.1 *Let C be a circular convolution matrix maps from cin channels to cout channels, then padding C with pn^2 columns of zeros on the right and qn^2 columns of zeros at the bottom is equivalent to padding each diagonal block of D with p columns of zeros on the right and q rows of zeros at the bottom.*

$$diag\big(\begin{bmatrix} D_1 & 0_{:,p} \\ 0_{q,:} & 0_{q:p} \end{bmatrix} \dots \begin{bmatrix} D_{n^2} & 0_{:p} \\ 0_{q:} & 0_{q:p} \end{bmatrix} \big) = \mathcal{F}_{cout+q,n^2} \begin{bmatrix} C & 0_{:pn^2} \\ 0_{qn^2:} & 0_{qn^2:pn^2} \end{bmatrix} \mathcal{F}^*_{cin+q,n^2}$$

Lemma 4.1 is illustrated in "Lemma4.1.jpg".

We can also derive a lemma similar to Lemma A.5 in [1]

Lemma 4.2 *Projecting out qn^2 rows and pn^2 columns of C is equivalent to projecting out q rows and p columns of each D_i*

$$\begin{bmatrix} I_{qn^2} & 0 \end{bmatrix} C \begin{bmatrix} I_{pn^2} \\ 0 \end{bmatrix} = \mathcal{F}^*_{q,n^2} diag\big(\begin{bmatrix} I_q & 0 \end{bmatrix} D_1 \begin{bmatrix} I_p \\ 0 \end{bmatrix}, ..., \begin{bmatrix} I_q & 0 \end{bmatrix} D_{n^2} \begin{bmatrix} I_p \\ 0 \end{bmatrix} \big) \mathcal{F}^*_{p,n^2}$$
$$\begin{bmatrix} 0 & I_{qn^2} \end{bmatrix} C \begin{bmatrix} I_{pn^2} \\ 0 \end{bmatrix} = \mathcal{F}^*_{q,n^2} diag\big(\begin{bmatrix} 0 & I_q \end{bmatrix} D_1 \begin{bmatrix} I_p \\ 0 \end{bmatrix}, ..., \begin{bmatrix} 0 & I_q \end{bmatrix} D_{n^2} \begin{bmatrix} I_p \\ 0 \end{bmatrix} \big) \mathcal{F}^*_{p,n^2}$$

Theorem 4.3 *Let $W_A \in \mathbb{R}^{cout \times cout \times n \times n}, W_A \in \mathbb{R}^{cin \times cout \times n \times n}$ be convolution weights, and $C_A \in \mathbb{R}^{cout \cdot n^2 \times cout \cdot n^2}, C_A \in \mathbb{R}^{cin \cdot n^2 \times cout \cdot n^2}$ be the convolution matrix induced by W_A and W_B, then*

$$\begin{bmatrix} I_{cout \cdot n^2} & 0_{:cin \cdot n^2} \end{bmatrix} Cayley(\begin{bmatrix} C_A & 0_{:cin \cdot n^2} \\ C_B & 0_{:cin \cdot n^2} \end{bmatrix}) \begin{bmatrix} I_{cout \cdot n^2} \\ 0_{cin \cdot n^2 :} \end{bmatrix} = \mathcal{F}^*_{cout,n^2} diag\Bigg(\begin{bmatrix} I_{cout} & 0_{:cin} \end{bmatrix} Cayley(\begin{bmatrix} D_1 & 0_{:cin} \end{bmatrix}) \begin{bmatrix} I_{cout} \\ 0_{cin:} \end{bmatrix}, ... \Bigg) \mathcal{F}_{cout,n^2}$$
$$\begin{bmatrix} 0_{:cout \cdot n^2} & I_{cin \cdot n^2} \end{bmatrix} Cayley(\begin{bmatrix} C_A & 0_{:cin \cdot n^2} \\ C_B & 0_{:cin \cdot n^2} \end{bmatrix}) \begin{bmatrix} I_{cout \cdot n^2} \\ 0_{cin \cdot n^2 :} \end{bmatrix} = \mathcal{F}^*_{cin,n^2} diag\Bigg(\begin{bmatrix} 0_{:cout} & I_{cin} \end{bmatrix} Cayley(\begin{bmatrix} D_1 & 0_{:cin} \end{bmatrix}) \begin{bmatrix} I_{cout} \\ 0_{cin:} \end{bmatrix}, ... \Bigg) \mathcal{F}_{cout,n^2}$$

where

$$\begin{bmatrix} C_A \\ C_B \end{bmatrix} = \mathcal{F}^*_{cin+cout,n^2} diag(D_1, ..., D_{n^2}) \mathcal{F}_{cout,n^2}$$

with each D_i has shape $(cin + cout) \times cout$

In a convolution layer, the convolution $Conv(X; W)$ can be written as $Cvec(X)$ where C is the convolution matrix induced by W. Thus we can write the lipschitz convolution residual layer in a similar structure as in the fully-connected layer

$$h(x) = \sqrt{\gamma} H vec(x) + G\sigma(W vec(x) + b)$$

where $W = -\sqrt{\gamma}\Lambda^{-1} G^T H$ and $\begin{bmatrix} G^T \\ H^T \end{bmatrix} = Cayley(\begin{bmatrix} C_A \\ C_B \end{bmatrix})$, C_A and C_B are convolution matrix induced by W_A and W_B. By theorem 4.3, we have

$$G^T = \mathcal{F}^*_{cout,n^2} diag\big(G_1^T, ..., G_{n^2}^T\big) \mathcal{F}_{cout,n^2}$$
$$H^T = \mathcal{F}^*_{cout,n^2} diag\big(H_1^T, ..., H_{n^2}^T\big) \mathcal{F}_{cout,n^2}$$

where $\begin{bmatrix} G_i^T \\ H_i^T \end{bmatrix} = Cayley(D_i)$, and the operations $Hvec(x)$, $Gvec(x)$, $G^T Hvec(x)$ can be efficiently computed with FFT, permutation and batch matrix multiplication as in [1]

References

[1] Asher Trockman and J Zico Kolter. Orthogonalizing convolutional layers with the cayley transform. *arXiv preprint arXiv:2104.07167*, 2021.