On the local Hessian of back-propagation

Huishuai Zhang, Wei Chen, Tie-Yan Liu

Microsoft Research

Contact : huzhang@microsoft.com

1. Motivation

A new interpretation of Back-propagation: BP is the one-step gradient update solution of minimizing Back-matching losses

Rewrite the loss of neural network:

$$Q(W, z; \gamma) := \ell\left(y;\ F_B(W_B, z_{B-1})\right) + \sum_{b=1}^{B-1} \frac{\gamma}{2} \|z_b - F_b(W_b; z_{b-1})\|^2$$

Local back-matching loss for block b:

$$\ell_b^k(W_b, z_{b-1}) := \begin{cases} \ell\left(y^k; F_B(W_B, z_{B-1})\right), & \text{for } b = B \\ \frac{1}{2}\left\| z_b^{k+\frac{1}{2}} - F_b(W_b; z_{b-1}) \right\|^2, & \text{for } b = B-1, \dots, 1 \end{cases}$$

2. Local Hessian

Local Hessian is the Hessian of minimizing local back-matching loss

$$H_{vec(W_b)} = \frac{\partial^2 \ell_b(W_b, z_{b-1}^k)}{\partial vec(W_b)^2}, \qquad H_z = \frac{\partial^2 \ell_b(W_b^k, z_{b-1})}{\partial z_{b-1}^2}$$

Relation to global Hessian (Gauss-Newton)

- $G = \left(\frac{\partial F}{\partial W}\right)^T H_L \left(\frac{\partial F}{\partial W}\right)$, where $\frac{\partial F}{\partial W}$ is the Jacobian matrix.
- $G_b = \left(\frac{\partial z_b}{\partial W_b}\right)^T \cdot \left(\frac{\partial z_{b+1}}{\partial z_b}\right)^T \cdots \left(\frac{\partial F}{\partial z_B}\right)^T H_L \left(\frac{\partial F}{\partial z_B}\right) \cdots \left(\frac{\partial z_{b+1}}{\partial z_b}\right) \cdot \left(\frac{\partial z_b}{\partial W_b}\right)$
- G_b has the same eigenvalues with the product of local Hessians
- The local Hessian determines where the flow is blocked, and hence relates to the efficiency of BP.

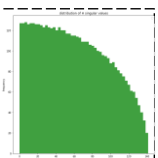
Case 1: Fully connected layer

$$z_b = F_b(W_b, z_{b-1}) = W_b \cdot z_{b-1}$$

Local Hessian

$$H_z = \left(W_b^k\right)^T W_b^k, \quad H_W = \textstyle\sum_{j=1}^m z_{b-1}^k[j]\left(z_{b-1}^k[j]\right)^T$$

For Gaussian initialization, H_z follows Marchenko-Pastur Law

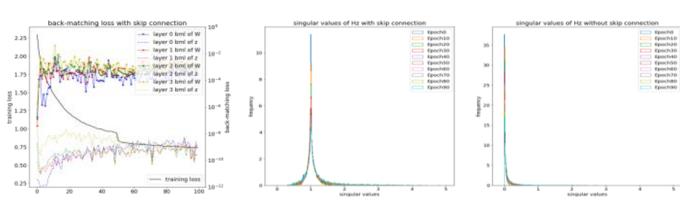


Case 2: Block with skip connection

$$z_b = F_b(W_b, z_{b-1}) = z_{b-1} + \phi_b(W_b, z_{b-1})$$

Local Hessian

$$H_z = \left(I + \frac{\partial \phi_b}{\partial z_{b-1}}\right)^T \left(I + \frac{\partial \phi_b}{\partial z_{b-1}}\right), H_W = \left(\frac{\partial \phi_b}{\partial vec(W_b)}\right)^T \left(\frac{\partial \phi_b}{\partial vec(W_b)}\right)$$



Remark: If a) $\sigma_{\max}\left(\frac{\partial \phi_b}{\partial z_{b-1}}\right) < 1 - s$, and b) $C\left(\frac{\partial \phi_b}{\partial z_{b-1}}\right) > \frac{1+s}{1-s}$, then

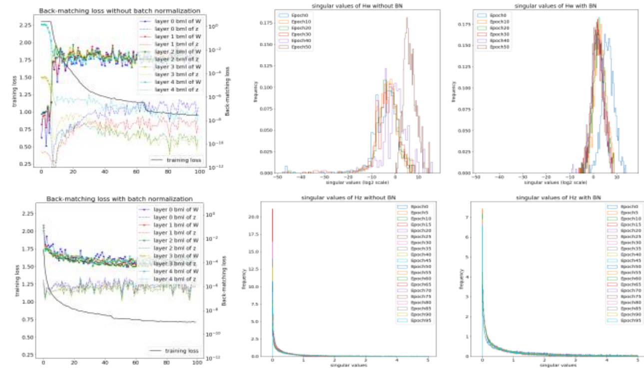
$$C\left(I + \frac{\partial \phi_b}{\partial z_{b-1}}\right) < C\left(\frac{\partial \phi_b}{\partial z_{b-1}}\right)$$

Case 3: Block with batch normalization

$$z_b = \mathrm{BN}(\tilde{z}_b) = (\tilde{z}_b - \mathbb{E}[\tilde{z}_b])/\sqrt{\mathrm{Var}[\tilde{z}_b]} \qquad \text{where } \tilde{z}_b = w_b^T z_{b-1}$$

Local Hessian

$$H_z \approx \sum_{i=1}^{n_b} \frac{w_b(i) w_b(i)^T}{\mathrm{Var}[w_b(i)^T z_{b-1}]}, \qquad H_w \approx \frac{\sum_{j=1}^m (z_{b-1}[j])(z_{b-1}[j])'}{m \cdot \mathrm{Var}[w_b(i)^T z_{b-1}]}$$



Remark 1: Scaling current layer does not affect H_z

Remark 2: Scaling layers bellow current layer does not affect H_w

Take-away : Local Hessian determines the efficiency of BP

5. Utilize local Hessian

Algorithm 2 Scale-amended SGD

Input: Gradient $\delta \boldsymbol{W}_b$ and scaling factor $m_{b,W}, m_{b,z}$; Initialize $m = 1$.

for $b = B, \dots, 1$ **do**

$$\delta' \boldsymbol{W}_b \leftarrow \delta \boldsymbol{W}_b / m / m_{b,W}$$
$$m \leftarrow m \cdot m_{b,z}$$

end for

1. Follow backward order
2. Scale the gradient by the local Hessian
3. Fully connected layer with BN,

$$m_{b,z} = \left\|W_b^T\right\|_{2,\mu}^2 / \|W_b\|_{2,\mu}^2 \quad \text{and} \quad m_{b,W} = 1/\|W_b\|_{2,\mu}^2$$

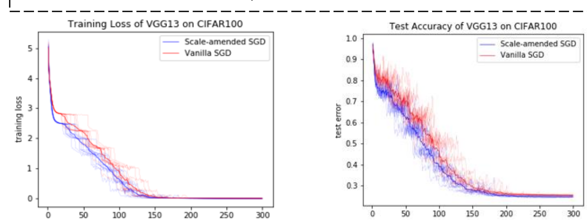


Table 1: Classification accuracies for CIFAR-10 and CIFAR-100.

	CIFAR10				CIFAR100			
	VGG11	VGG13	VGG16	VGG19	VGG11	VGG13	VGG16	VGG19
SGD	92.34	93.90	93.72	93.47	71.84	74.07	72.86	71.35
LARS	91.81	93.40	93.47	93.48	67.26	70.35	69.90	69.52
LSALR	92.58	93.68	93.35	93.46	71.14	73.74	73.14	70.76
OURS	92.45	**94.11**	**93.90**	**93.88**	**73.39**	**75.32**	**74.68**	**72.82**

NeurIPS-2019: The 32nd Conference on Neural Information Processing Systems