SKFAC: Training Neural Networks with Faster Kronecker-Factored Approximate Curvature

Appendix

Proof of Theorem 1

Proof Since $(\mathbf{\Omega}_\ell^{(\lambda)})^{-1} = -\frac{1}{\lambda}(-\mathbf{I}_n - \frac{\mathbf{A}^T}{\sqrt{\lambda M}}\frac{\mathbf{A}}{\sqrt{\lambda M}})^{-1}$, let $\mathbf{X} = -\mathbf{I}_n$ and $\mathbf{Y} = \frac{\mathbf{A}}{\sqrt{\lambda M}}$, where $\mathbf{A} \in \mathbb{R}^{n\times m}$. Formulate a matrix,

$$\begin{bmatrix} \mathbf{X} & \mathbf{Y}^T \\ \mathbf{Y} & \mathbf{I}_m \end{bmatrix}. \tag{1}$$

What's been apparent is that following formulations are hold.

$$\begin{bmatrix} \mathbf{X} - \mathbf{Y}^T\mathbf{Y} & \\ & \mathbf{I}_m \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & -\mathbf{Y}^T \\ & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{X} & \mathbf{Y}^T \\ \mathbf{Y} & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{I}_n & \\ -\mathbf{Y} & \mathbf{I}_m \end{bmatrix}, \tag{2}$$

$$\begin{bmatrix} \mathbf{X} & \\ & \mathbf{I}_m + \mathbf{Y}\mathbf{Y}^T \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \\ -\mathbf{Y} & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{X} & \mathbf{Y}^T \\ \mathbf{Y} & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{I}_n & -\mathbf{Y}^T \\ & \mathbf{I}_m \end{bmatrix}. \tag{3}$$

It is easy to verified that,

$$\begin{bmatrix} \mathbf{I}_n & -\mathbf{Y}^T \\ & \mathbf{I}_m \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I}_n & \mathbf{Y}^T \\ & \mathbf{I}_m \end{bmatrix}, \text{ and } \begin{bmatrix} \mathbf{I}_n & \\ -\mathbf{Y} & \mathbf{I}_m \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I}_n & \\ \mathbf{Y} & \mathbf{I}_m \end{bmatrix}. \tag{4}$$

Combining equation (3) with equation (4), we get the inverse of equation (1):

$$\begin{bmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{Y}^T & \mathbf{I}_m \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I}_n & -\mathbf{Y}^T \\ & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{X} & \\ & (\mathbf{I}_m + \mathbf{Y}\mathbf{Y}^T)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I}_n & \\ -\mathbf{Y} & \mathbf{I}_m \end{bmatrix}. \tag{5}$$

Then, the inverse of equation (2) can be obtained by

$$\begin{aligned} &\begin{bmatrix} (\mathbf{X} - \mathbf{Y}^T\mathbf{Y})^{-1} & \\ & \mathbf{I}_m \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \\ \mathbf{Y} & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{X} & \mathbf{Y}^T \\ \mathbf{Y} & \mathbf{I}_m \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{I}_n & \mathbf{Y}^T \\ & \mathbf{I}_m \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I}_n & \\ \mathbf{Y} & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{I}_n & -\mathbf{Y}^T \\ & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{X} & \\ & (\mathbf{I}_m + \mathbf{Y}\mathbf{Y}^T)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I}_n & \\ -\mathbf{Y} & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{I}_n & \mathbf{Y}^T \\ & \mathbf{I}_m \end{bmatrix} \\ &= diag(\mathbf{X} + \mathbf{Y}^T(\mathbf{I}_m + \mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Y}, \mathbf{I}_m). \end{aligned} \tag{6}$$

Thus, the inverse of $\mathbf{\Omega}_\ell^{(\lambda)}$ is given as follows,

$$\begin{aligned} (\mathbf{\Omega}_\ell^{(\lambda)})^{-1} &= -\frac{1}{\lambda}(-\mathbf{I}_n - \frac{\mathbf{A}^T}{\sqrt{\lambda M}}\frac{\mathbf{A}}{\sqrt{\lambda M}})^{-1} \\ &= \frac{1}{\lambda}\mathbf{I}_n - \frac{1}{\lambda M}\mathbf{A}^T(\mathbf{I}_m + \frac{1}{\lambda M}\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}. \end{aligned} \tag{7}$$

Via above steps, an analogous formulation of $(\mathbf{\Gamma}_\ell^{(\lambda)})^{-1}$ can be derived. □

There is another brief proof by introducing an existing theorem.

Theorem *If* $\mathbf{X}$ *and* $\mathbf{I}_m - \mathbf{V}\mathbf{X}^{-1}\mathbf{U}$ *are invertible for given* $\mathbf{X} \in \mathbb{R}^n, \mathbf{U} \in \mathbb{R}^{n\times m}, V \in \mathbb{R}^{m\times n}$*, then* $\mathbf{X} - \mathbf{U}\mathbf{V}$ *is invertible and following equation is hold,*

$$(\mathbf{X} - \mathbf{U}\mathbf{V})^{-1} = \mathbf{X}^{-1} + \mathbf{X}^{-1}\mathbf{U}(\mathbf{I}_m - \mathbf{V}\mathbf{X}^{-1}\mathbf{U})^{-1}\mathbf{V}\mathbf{X}^{-1}. \tag{8}$$

It is well-known Woodbury formulation (Woodbury 1950), and a detailed review is given in (Hager 1989).

We begin with rewriting $(\mathbf{\Omega}_\ell^{(\lambda)})^{-1} = \lambda\mathbf{I} + \frac{\mathbf{A}^T\mathbf{A}}{M}$ to obtain an analogous formulation as equation (8).

$$(\mathbf{\Omega}_\ell^{(\lambda)})^{-1} = -\lambda(-\mathbf{I}_n - \frac{\mathbf{A}}{\sqrt{\lambda M}}\frac{\mathbf{A}^T}{\sqrt{\lambda M}})^{-1}. \tag{9}$$

By direct substituting equation (10) into equation (8),

$$\mathbf{X} = -\lambda\mathbf{I}_n, \mathbf{U} = \frac{1}{\sqrt{\lambda M}}\mathbf{A}^T, \mathbf{V} = \frac{1}{\sqrt{\lambda M}}\mathbf{A}. \tag{10}$$

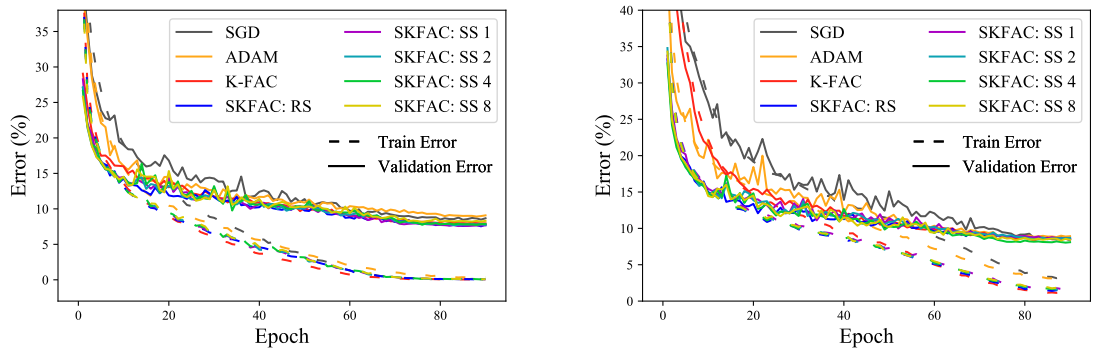
Then, following equation is obtained.

$$(\mathbf{\Omega}_\ell^{(\lambda)})^{-1} = \frac{1}{\lambda}\mathbf{I}_n - \frac{1}{\lambda M}\mathbf{A}(\mathbf{I}_m + \frac{1}{\lambda M}\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T. \tag{11}$$

An analogous formulation of the $(\mathbf{\Gamma}_\ell^{(\lambda)})^{-1}$ can be obtained by the above process. □

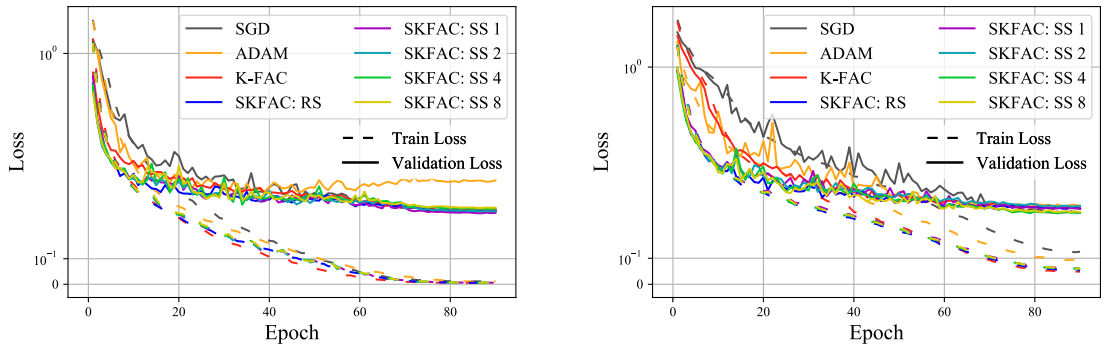
Additional Figures

Following figures demonstrate the convergence curves of error rates and loss function values with respect to epochs.



(a) VGG-11 (b) ResNet-34

Figure 1: Validate error for VGG-11 and ResNet-34 on Cifar-10 dataset.



(a) VGG-11 (b) ResNet-34

Figure 2: Training loss for VGG-11 and ResNet-34 on Cifar-10 dataset.

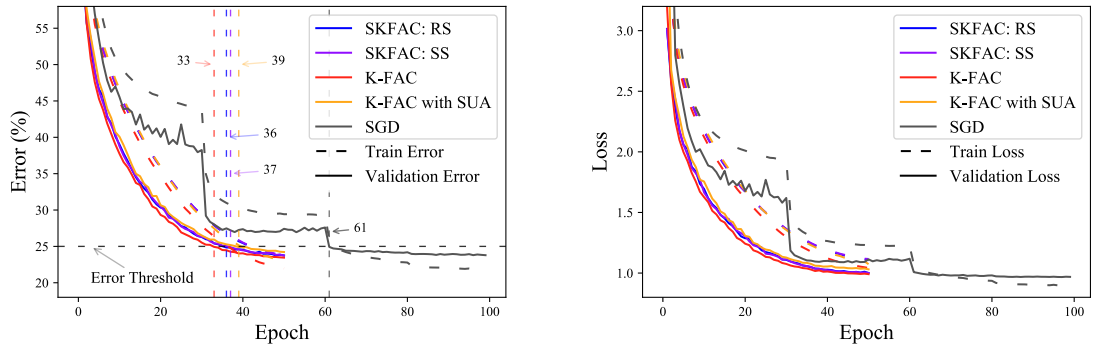


Figure 3: Training loss for ResNet50 on ImageNet-1000 dataset.

Additional experiments

To verify the time saving of training neural networks, we report the wall-clock time of running the algorithms on ResNet34 and VGG11. This method can accumulate the wall-clock training time of relatively small models on Cifar10 as shown in Table 1.

Table 1: Wall-clock time when reaching 90% validation accuracy

Alogorithm	Adam	KFAC	SGD	RS	SS_1	SS_2
ResNet34	62.45	50.22	52.58	49.45	50.32	**47.45**
VGG11	37.67	40.63	32.15	28.02	**24.27**	26.90

*(Ours) RS: SKFAC_ReduceSum SS_1: SKFAC_SpatialSampling_1 SS_2: SKFAC_SpatialSampling_2

Moreover, the approximate errors are measured in every epoch for showing the effectiveness of our proposed on approximating factors of Fisher information matrices. We show the approximate errors in Figure 4.

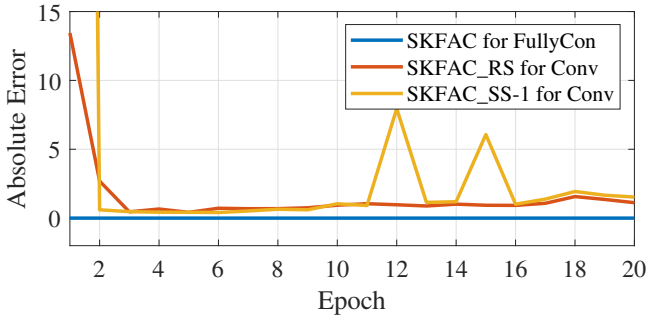


Figure 4: The absolute error between the exact inverse of Fisher information matrix and the approximation by the proposed method.

References

Hager, W. W. 1989. Updating the Inverse of a Matrix. *SIAM Review* 31(2): 221–239.

Woodbury, M. 1950. Inverting modified matrices. Memorandum rept. 42, Statistical Research Group, Princeton University, Princeton, NJ.