Supplementary Material for "Knowledge Transfer with Jacobian Matching"

Suraj Srinivas [1] **François Fleuret** [1]

1. Proof for Proposition 1

Proposition. *Consider the squared error cost function for matching soft targets of two neural networks with k-length targets ($\in \mathbb{R}^k$), given by $\ell(\mathcal{T}(\mathbf{x}), \mathcal{S}(\mathbf{x})) = \sum_{i=1}^{k}(\mathcal{T}^i(\mathbf{x}) - \mathcal{S}^i(\mathbf{x}))^2$, where $\mathbf{x} \in \mathbb{R}^D$ is an input data point. Let $\boldsymbol{\xi}$ ($\in \mathbb{R}^D$) $= \sigma\, \mathbf{z}$ be a scaled version of a unit normal random variable $\mathbf{z} \in \mathbb{R}^D$ with scaling factor $\sigma \in \mathbb{R}$. Then the following is locally true.*

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{i=1}^{k}\left(\mathcal{T}^i(\mathbf{x}+\boldsymbol{\xi}) - \mathcal{S}^i(\mathbf{x}+\boldsymbol{\xi})\right)^2\right] \\
&= \sum_{i=1}^{k}\left(\mathcal{T}^i(\mathbf{x}) - \mathcal{S}^i(\mathbf{x})\right)^2 + \sigma^2\sum_{i=1}^{k}\|\nabla_x\mathcal{T}^i(\mathbf{x}) - \nabla_x\mathcal{S}^i(\mathbf{x})\|_2^2 \\
&+ \mathcal{O}(\sigma^4)
\end{aligned}
$$

Proof. There exists σ and $\boldsymbol{\xi}$ small enough that first-order Taylor series expansion holds true

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{i=1}^{k}\left(\mathcal{T}^i(\mathbf{x}+\boldsymbol{\xi}) - \mathcal{S}^i(\mathbf{x}+\boldsymbol{\xi})\right)^2\right] \\
&= \mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{i=1}^{k}\left(\mathcal{T}^i(\mathbf{x}) + \boldsymbol{\xi}^T\nabla_x\mathcal{T}^i(\mathbf{x}) - \mathcal{S}^i(\mathbf{x}) - \boldsymbol{\xi}^T\nabla_x\mathcal{S}^i(\mathbf{x})\right)^2\right] \\
&+ \mathcal{O}(\sigma^4) \\
&= \sum_{i=1}^{k}\left(\mathcal{T}^i(\mathbf{x}) - \mathcal{S}^i(\mathbf{x})\right)^2 \\
&+ \mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{i=1}^{k}\left[\boldsymbol{\xi}^T\left(\nabla_x\mathcal{T}^i(\mathbf{x}) - \nabla_x\mathcal{S}^i(\mathbf{x})\right)\right]^2\right] + \mathcal{O}(\sigma^4) \qquad (1)
\end{aligned}
$$

To get equation 1, we use the fact that mean of $\boldsymbol{\xi}$ is zero. To complete the proof, we use the diagonal assumption on the covariance matrix of $\boldsymbol{\xi}$.

□

Proofs of other statements are similar. For proof for cross-entropy loss of Proposition 2, use a second order Taylor series expansion of $\log(\cdot)$ in the first step.

2. Proof for Proposition 3

Proposition. *From the notations in the main text, we have*

$$
\frac{1}{|\mathcal{D}_l|}\sum_{\mathbf{x}\sim\mathcal{D}_l}\ell(f(\mathbf{x}), g(\mathbf{x})) \le \max_{\mathbf{x}\sim\mathcal{D}_s}\ell(f(\mathbf{x}), g(\mathbf{x})) + \mathrm{K}\mathcal{H}_a(\mathcal{D}_l, \mathcal{D}_s)
$$

Proof. Let us denote $\rho(\mathbf{x}) = \ell(f(\mathbf{x}), g(\mathbf{x}))$ for convenience. Assume Lipschitz continuity for $\rho(\mathbf{x})$ with Lipschitz constant K, and distance metric $\psi_{\mathbf{x}}(\cdot,\cdot)$ in the input space -

$$
\begin{aligned}
&\|\rho(\mathbf{x}_1) - \rho(\mathbf{x}_2)\|_1 \le \mathrm{K}\psi_{\mathbf{x}}(\mathbf{x}_1, \mathbf{x}_2) \\
\implies &\rho(\mathbf{x}_1) \le \rho(\mathbf{x}_2) + \mathrm{K}\psi_{\mathbf{x}}(\mathbf{x}_1, \mathbf{x}_2)
\end{aligned}
$$

Assuming that $\rho(\mathbf{x}_1) \ge \rho(\mathbf{x}_2)$. Note that it holds even otherwise, but is trivial.

Now, for every datapoint $\mathbf{x}_l \in \mathcal{D}_l$, there exists a point $\mathbf{x}_s \in \mathcal{D}_s$ such that $\psi_{\mathbf{x}}(\mathbf{x}_s, \mathbf{x}_l)$ is the smallest among all points in $\mathcal{D}_s$. In other words, we look at the point in $\mathcal{D}_s$ closest to each point $\mathbf{x}_l$. Note that in this process only a subset of points $\mathrm{d}_s \subseteq \mathcal{D}_s$ are chosen, and individual points can be chosen multiple times. For these points, we can write

$$
\begin{aligned}
&\rho(\mathbf{x}_l) \le \rho(\mathbf{x}_s) + \mathrm{K}\psi_{\mathbf{x}}(\mathbf{x}_s, \mathbf{x}_l) \\
\implies &\frac{1}{|\mathcal{D}_l|}\sum_{\mathbf{x}_l\sim\mathcal{D}_l}\rho(\mathbf{x}_l) \le \frac{1}{|\mathcal{D}_l|}\sum_{\mathbf{x}_s \text{ closest to } \mathbf{x}_l}\rho(\mathbf{x}_s) \\
&+ \frac{1}{|\mathcal{D}_l|}\sum_{\mathbf{x}_s \text{ closest to } \mathbf{x}_l}\mathrm{K}\psi_{\mathbf{x}}(\mathbf{x}_s, \mathbf{x}_l)
\end{aligned}
$$

[1] Idiap Research Institute & EPFL, Switzerland. Correspondence to: Suraj Srinivas <suraj.srinivas@idiap.ch>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

We see that $\frac{1}{|\mathcal{D}_l|}\sum_{\mathbf{x}_s}\rho(\mathbf{x}_s) \leq \max_{\mathbf{x}\sim \mathrm{d}_s}\rho(\mathbf{x}) \leq \max_{\mathbf{x}\sim \mathcal{D}_s}\rho(\mathbf{x})$, which is a consequence of the fact that the max is greater than any convex combination of elements.

Also, we have $\psi_{\mathbf{x}}(\mathbf{x}_l, \mathbf{x}_s) \leq \mathcal{H}_a(\mathcal{D}_l, \mathcal{D}_s)$, which is the maximum distance between any two 'closest' points from $\mathcal{D}_l$ to $\mathcal{D}_s$ (by definition).

Applying these bounds, we have the final result.

□

2.1. Proof for Corollary

Corollary. *For any superset* $\mathrm{D}'_s \supseteq \mathcal{D}_s$ *of the target dataset,* $\mathcal{H}_a(\mathcal{D}_l, \mathcal{D}'_s) \leq \mathcal{H}_a(\mathcal{D}_l, \mathcal{D}_s)$

Proof. From the previous proof, we have $\rho(\mathbf{x}_l) \leq \rho(\mathbf{x}_s) + \mathrm{K}\psi_{\mathbf{x}}(\mathbf{x}_s, \mathbf{x}_l)$ for an individual point $\mathbf{x}_l$. Now if we have $\mathrm{D}'_s \supseteq \mathcal{D}_s$, then we have $\rho(\mathbf{x}_l) \leq \rho(\mathbf{x}'_s) + \mathrm{K}\psi_{\mathbf{x}}(\mathbf{x}'_s, \mathbf{x}_l)$, where $\mathbf{x}'_s$ is the new point closest to $\mathbf{x}_l$. It is clear that $\psi_{\mathbf{x}}(\mathbf{x}'_s, \mathbf{x}_l) \leq \psi_{\mathbf{x}}(\mathbf{x}_s, \mathbf{x}_l)$ for all $\mathbf{x}_l$. Hence it follows that $\mathcal{H}_a(\mathcal{D}_l, \mathcal{D}'_s) \leq \mathcal{H}_a(\mathcal{D}_l, \mathcal{D}_s)$.

□

3. Justification for Jacobian loss

We use the following loss term for Jacobian matching for transfer learning.

$$\text{Match Jacobians} = \left|\left| \frac{\nabla_x f(\mathbf{x})}{\|\nabla_x f(\mathbf{x})\|_2} - \frac{\nabla_x g(\mathbf{x})}{\|\nabla_x g(\mathbf{x})\|_2} \right|\right|_2^2 \tag{2}$$

We can show that the above loss term corresponds to adding a noise term $\boldsymbol{\xi}_f \propto \|\nabla_x f(\mathbf{x})\|_2^{-1}$ for $f(\mathbf{x})$ and $\boldsymbol{\xi}_g \propto \|\nabla_x g(\mathbf{x})\|_2^{-1}$ for $g(\mathbf{x})$ for the distillation loss. From the first order Taylor series expansion, we see that $g(x+\boldsymbol{\xi}) = g(x) + \boldsymbol{\xi}_g \nabla_x g(\mathbf{x})$. Thus for networks $f(\cdot)$ and $g(\cdot)$ with different Jacobian magnitudes, we expect different responses for the same noisy inputs. Specifically, we see that $\mathbb{E}_{\boldsymbol{\xi}_g}\|g(x+\boldsymbol{\xi}_g) - g(x)\|_2^2 = \sigma_g^2 \|\nabla_x g(\mathbf{x})\|_2^2 = \sigma^2 \frac{\|\nabla_x g(\mathbf{x})\|_2^2}{\|\nabla_x g(\mathbf{x})\|_2^2} = \sigma^2$ for a gaussian model with covariance matrix being σ times the identity.

4. Experimental details

4.1. VGG Network Architectures

The architecture for our networks follow the VGG design philosophy. Specifically, we have blocks with the following elements:

- 3×3 conv kernels with c channels of stride 1
- Batch Normalization
- ReLU

Whenever we use Max-pooling (M), we use stride 2 and window size 2.

The architecture for VGG-9 is - $[64 - M - 128 - M - 256 - 256 - M - 512 - 512 - M - 512 - 512 - M]$. Here, the number stands for the number of convolution channels, and M represents max-pooling. At the end of all the convolutional and max-pooling layers, we have a Global Average Pooling (GAP) layer, after which we have a fully connected layer leading up to the final classes. Similar architecture is used for the case of both CIFAR and MIT Scene experiments.

The architecture for VGG-4 is - $[64 - M - 128 - M - 512 - M]$.

4.2. Loss function

The loss function for distillation experiments use the following form.

$$\ell(\mathcal{S}, \mathcal{T}) = \alpha \times (\text{CE}) + \beta \times (\text{Match Activations}) + \gamma \times (\text{Match Jacobians})$$

In our experiments, α, β, γ are either set to 1 or 0. In other words, all regularization constants are 1.

Here, 'CE' refers to cross-entropy with ground truth labels. 'Match Activations' refers to squared error term over pre-softmax activations of the form $(y_s - y_t)^2$. 'Match Jacobians' refers to the same squared error term, but for Jacobians.

For the MIT Scene experiments, α, β, γ are either set to 10 or 0, depending on the specific method. To compute the Jacobian, we use average pooling over a $feature\ size/5$ window with a stride of 1. We match the Jacobian after the first residual block for resnet, and after the second maxpool for VGG. This corresponds to feature level "1" in the ablation experiments.

4.3. Optimization

For CIFAR100 experiments, we run optimization for 500 epochs. We use the Adam optimizer, with an initial learning rate of $1e-3$, and a single learning rate annealing (to $1e-4$) at 400 epochs. We used a batch size of 128.

For MIT Scenes, we used SGD with momentum of 0.9, for 75 epochs. The initial learning rate is $1e-3$, and it is reduced 10 times after 40 and 60 epochs. We used batch size 8. This is because the Jacobian computation is very memory intensive.