[Supplementary Materials] How to Train the Teacher Model for Effective Knowledge Distillation

Shayan Mohajer Hamidi*[1], Xizhen Deng*[2], Renhao Tan[1], Linfeng Ye[1], and Ahmed Hussein Salamah[1]

[1] University of Waterloo, Waterloo ON N2L 3G1, Canada
{smohajer,cameron.tan,l44ye,ahamsalamah}@uwaterloo.ca
[2] University of Michigan, Ann Arbor, MI 48109, USA xizhen@umich.edu
* Authors contributed equally

1 Proof of Theorem 1

We first prove the theorem for $\ell = \text{MSE}$ in Section 1.1, and then prove it for $\ell = \text{CE}$ in Section 1.2.

1.1 MSE loss

$$R(f_{\boldsymbol{\theta}}, \ell = \text{MSE}) = \mathbb{E}_{(\boldsymbol{x},y)}\left\{\sum_{i=1}^{C}\left(\mathbf{y}[i] - \boldsymbol{p}_{\boldsymbol{x}}[i]\right)^2\right\} \tag{1}$$

$$= \int \sum_{c=1}^{C}\left[\sum_{i=1}^{C}\left(\mathbf{y}[i] - \boldsymbol{p}_{\boldsymbol{x}}[i]\right)^2\right]\mathbb{P}(\boldsymbol{x}, c)d\boldsymbol{x} \tag{2}$$

$$= \int \left[\sum_{c=1}^{C}\sum_{i=1}^{C}\left(\mathbf{y}[i] - \boldsymbol{p}_{\boldsymbol{x}}[i]\right)^2\mathbb{P}(c|\boldsymbol{x})\right]\mathbb{P}(\boldsymbol{x})d\boldsymbol{x} \tag{3}$$

$$= \int \sum_{k=1}^{C}\left[\sum_{c=1}^{C}\sum_{i=1}^{C}\left(\mathbf{y}[i] - \boldsymbol{p}_{\boldsymbol{x}}[i]\right)^2\mathbb{P}(c|\boldsymbol{x})\right]\mathbb{P}(\boldsymbol{x}, k)d\boldsymbol{x} \tag{4}$$

$$= \mathbb{E}_{(\boldsymbol{x},y)}\left\{\sum_{c=1}^{C}\sum_{i=1}^{C}\left(\mathbf{y}[i] - \boldsymbol{p}_{\boldsymbol{x}}[i]\right)^2\mathbb{P}(c|\boldsymbol{x})\right\} \tag{5}$$

$$= \mathbb{E}_{(\boldsymbol{x},y)}\left\{\sum_{c=1}^{C}\sum_{i=1}^{C}\boldsymbol{p}_{\boldsymbol{x}}^2[i]\mathbb{P}(c|\boldsymbol{x}) - 2\boldsymbol{p}_{\boldsymbol{x}}[i]\mathbf{y}[i]\mathbb{P}(c|\boldsymbol{x}) + \mathbf{y}^2[i]\mathbb{P}(c|\boldsymbol{x})\right\} \tag{6}$$

$$= \mathbb{E}_{(\boldsymbol{x},y)}\left\{\sum_{i=1}^{C}\boldsymbol{p}_{\boldsymbol{x}}^2[i] - 2\boldsymbol{p}_{\boldsymbol{x}}[i]\sum_{c=1}^{C}\mathbf{y}[i]\mathbb{P}(c|\boldsymbol{x}) + \sum_{c=1}^{C}\mathbf{y}^2[i]\mathbb{P}(c|\boldsymbol{x})\right\} \tag{7}$$

$$= \mathbb{E}_{(\boldsymbol{x},y)}\left\{\sum_{i=1}^{C}\boldsymbol{p}_{\boldsymbol{x}}^2[i] - 2\boldsymbol{p}_{\boldsymbol{x}}[i]\mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\} + \mathbb{E}\Big\{\mathbf{y}^2[i] \;\Big|\; \boldsymbol{x}\Big\}\right\} \tag{8}$$

$$= \mathbb{E}_{(\boldsymbol{x},y)}\left\{\sum_{i=1}^{C}\boldsymbol{p}_{\boldsymbol{x}}^2[i] - 2\boldsymbol{p}_{\boldsymbol{x}}[i]\mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\}\right.$$

$$+\mathbb{E}\Big\{\mathbf{y}^2[i] \mid \boldsymbol{x}\Big\}+\mathbb{E}^2\Big\{\mathbf{y}[i] \mid \boldsymbol{x}\Big\}-\mathbb{E}^2\Big\{\mathbf{y}[i] \mid \boldsymbol{x}\Big\}\Big\} \tag{9}$$

$$=\mathbb{E}_{(\boldsymbol{x},y)}\Big\{\sum_{i=1}^{C}\Big(\boldsymbol{p}_{\boldsymbol{x}}[i]-\mathbb{E}\Big\{\mathbf{y}[i] \mid \boldsymbol{x}\Big\}\Big)^2\Big\}+\mathbb{E}_{(\boldsymbol{x},y)}\Big\{\sum_{i=1}^{C}\mathrm{Var}\Big\{\mathbf{y}[i] \mid \boldsymbol{x}\Big\}\Big\}, \tag{10}$$

where $\mathrm{Var}\Big\{\mathbf{y}[i] \mid \boldsymbol{x}\Big\}$ denotes the conditional variance of $\mathbf{y}[i]$. Note that the second term in Eq. (10) is independent of the DNN's parameters, and is an intrinsic characteristic of the underlying dataset. Hence, the minimum of $R(f_{\boldsymbol{\theta}}, \ell = \mathrm{MSE})$ is achieved when $\boldsymbol{p}_{\boldsymbol{x}}[i] = \mathbb{E}\Big\{\mathbf{y}[i] \mid \boldsymbol{x}\Big\}$. Also, note that

$$\mathbb{E}\Big\{\mathbf{y}[i] \mid \boldsymbol{x}\Big\}=\sum_{c=1}^{C}\mathbf{y}[c]\mathbb{P}(c|\boldsymbol{x})=\Pr(y|\boldsymbol{x}), \tag{11}$$

where the last equation is obtained since all the entries of $\mathbf{y}$ are zero except its y-th entry. Therefore, $\boldsymbol{p}_{\boldsymbol{x}} = [\Pr(y|\boldsymbol{x})]_{y\in[C]} = \boldsymbol{p}^*_{\boldsymbol{x}}$.

$$\begin{aligned}&\min_{\boldsymbol{\theta}} R(f_{\boldsymbol{\theta}}, \ell = \mathrm{MSE})\\ &\quad\equiv \min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},y)}\Big\{\sum_{i=1}^{C}\Big(\boldsymbol{p}_{\boldsymbol{x}}[i]-\mathbb{E}\Big\{\mathbf{y}[i] \mid \boldsymbol{x}\Big\}\Big)^2\Big\}.\end{aligned} \tag{12}$$

On the other hand,

$$\mathbb{E}_{(\boldsymbol{x},y)}\Big\{\sum_{i=1}^{C}\Big(\boldsymbol{p}_{\boldsymbol{x}}[i]-\mathbb{E}\Big\{\mathbf{y}[i] \mid \boldsymbol{x}\Big\}\Big)^2\Big\} \tag{13}$$

$$=\mathbb{E}_{(\boldsymbol{x},y)}\Big\{\sum_{i=1}^{C}\left(\boldsymbol{p}_{\boldsymbol{x}}[i]-\boldsymbol{p}^*_{\boldsymbol{x}}[i]\right)^2\Big\} \tag{14}$$

$$=\mathbb{E}_{(\boldsymbol{x},y)}\Big\{\mathrm{MSE}\left(\boldsymbol{p}_{\boldsymbol{x}}[i], \boldsymbol{p}^*_{\boldsymbol{x}}[i]\right)\Big\}. \tag{15}$$

Hence, the proof is concluded for MSE loss.

1.2 CE loss

The CE loss is as follows:

$$R(f_{\boldsymbol{\theta}}, \ell = \mathrm{CE}) = -\mathbb{E}_{(\boldsymbol{x},y)}\Big\{\sum_{i=1}^{C}\mathbf{y}[i]\log \boldsymbol{p}_{\boldsymbol{x}}[i]\Big\} \tag{16}$$

Similarly to the simplification we made for MSE loss, the expression in Eq. (16) could be simplified as

$$
\begin{aligned}
&R(f_{\boldsymbol{\theta}}, \ell = \mathrm{CE}) \\
&= -\mathbb{E}_{(\boldsymbol{x},y)}\left\{\sum_{i=1}^{C} \mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\} \log \boldsymbol{p}_{\boldsymbol{x}}[i]\right\} && (17)\\
&= -\mathbb{E}_{(\boldsymbol{x},y)}\left\{\sum_{i=1}^{C} \mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\} \log \boldsymbol{p}_{\boldsymbol{x}}[i]\right. \\
&\qquad\qquad - \mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\} \log \mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\} \\
&\qquad\qquad \left. + \mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\} \log \mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\}\right\} && (18)\\
&= -\mathbb{E}_{(\boldsymbol{x},y)}\left\{\sum_{i=1}^{C} \mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\} \log \frac{\boldsymbol{p}_{\boldsymbol{x}}[i]}{\mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\}}\right\} \\
&\qquad\qquad - \mathbb{E}_{(\boldsymbol{x},y)}\left\{\sum_{i=1}^{C} \mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\} \log \sum_{i=1}^{C} \mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\}\right\}. && (19)
\end{aligned}
$$

We note that the second term in the last equation is independent of the DNN. By taking the first derivative from the first term in Eq. (19), it is easy to show that the minimum is achieved for $\boldsymbol{p}_{\boldsymbol{x}}[i] = \mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\} = \Pr(y|\boldsymbol{x})$, or equivalently for $\boldsymbol{p}_{\boldsymbol{x}} = \boldsymbol{p}^*_{\boldsymbol{x}}$.

Note that one can confirm that the critical point of Eq. (19) is indeed a minimum point by calculating its second derivative.

$$
\begin{aligned}
&\min_{\boldsymbol{\theta}} R(f_{\boldsymbol{\theta}}, \ell = \mathrm{CE}) \\
&\qquad \equiv \min_{\boldsymbol{\theta}} -\mathbb{E}_{(\boldsymbol{x},y)}\left\{\sum_{i=1}^{C} \mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\} \log \frac{\boldsymbol{p}_{\boldsymbol{x}}[i]}{\mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\}}\right\}. && (20)
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
&\min_{\boldsymbol{\theta}} -\mathbb{E}_{(\boldsymbol{x},y)}\left\{\sum_{i=1}^{C} \mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\} \log \frac{\boldsymbol{p}_{\boldsymbol{x}}[i]}{\mathbb{E}\Big\{\mathbf{y}[i] \;\Big|\; \boldsymbol{x}\Big\}}\right\} && (21)\\
&\equiv \min_{\boldsymbol{\theta}} -\mathbb{E}_{(\boldsymbol{x},y)}\left\{\sum_{i=1}^{C} \boldsymbol{p}^*_{\boldsymbol{x}}[i] \log \frac{\boldsymbol{p}_{\boldsymbol{x}}[i]}{\boldsymbol{p}^*_{\boldsymbol{x}}[i]}\right\} && (22)\\
&\equiv \min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},y)}\Big\{\mathrm{CE}\left(\boldsymbol{p}^*_{\boldsymbol{x}}[i], \boldsymbol{p}_{\boldsymbol{x}}[i]\right)\Big\}, && (23)
\end{aligned}
$$

which concludes the proof for CE loss.

2 Hyper-parameters for KD Variants

In the paper, we incorporate several state-of-the-art KD variants. Specifically, we employ a linear combination of cross-entropy loss and knowledge distillation loss, as outlined below:

$$\ell = \ell_{\text{CE}} + \gamma \ell_{\text{Distill}} \tag{24}$$

For different KD variants, we directly adopted the γ values as those reported in their original papers. The γ values and additional details for each KD variant used in the paper are outlined below:

1. AT [10]: $\gamma = 1000$;
2. PKT [5]: $\gamma = 30000$;
3. SP [8]: $\gamma = 3000$;
4. CC [6]: $\gamma = 0.02$;
5. RKD [4]: $\gamma_1 = 25$ for distance, and $\gamma_2 = 50$ for angle;
6. VID [1]: $\gamma = 1$;
7. CRD [7]: $\gamma = 0.8$;
8. DKD [11]: Consistent with their official implementation, we select γ from the set $\{0.2, 1, 2, 4, 8\}$;
9. REVIEWKD [2]: Consistent with their official implementation, we select γ from the set $\{0.6, 1, 5, 8\}$;
10. HSAKD [9]: $\gamma = 1$.

For KD, we follow the original implementation in [3], set $\alpha = 0.9$, and $T = 4$.

For both zero-shot and few-shot, we follow the same training recipe as that discussed in this subsection.

3 Accuracy variances of Table 1 in the paper

Table 1: The variance in accuracy for the results in Table 1 in the main paper, where the teacher and student have the **same** architectures.

Teacher	ResNet-56	ResNet-110	ResNet-110	WRN-40-2	WRN-40-2	VGG-13
Student	ResNet-20	ResNet-20	ResNet-32	WRN-16-2	WRN-40-1	VGG-8
KD	0.10	0.20	0.13	0.12	0.34	0.22
AT	0.28	0.32	0.18	0.15	0.28	0.16
PKT	0.38	0.12	0.14	0.15	0.28	0.20
SP	0.07	0.26	0.18	0.27	0.16	0.14
CC	0.47	0.18	0.23	0.24	0.16	0.29
RKD	0.13	0.17	0.24	0.15	0.21	0.28
VID	0.25	0.09	0.13	0.25	0.28	0.10
CRD	0.13	0.17	0.08	0.15	0.11	0.05
REVIEWKD	0.26	0.30	0.43	0.16	0.14	0.18
DKD	0.14	0.15	0.15	0.27	0.29	0.24
HSAKD	0.23	0.14	0.16	0.14	0.21	0.22

Table 2: The variance in accuracy for the results in Table 1 in the main paper, where the teacher and student have **different** architectures.

Teacher	ResNet-50	ResNet-50	ResNet-32×4	ResNet-32×4	WRN-40-2	VGG-13
Student	MobileNetV2	VGG-8	ShuffleNetV1	ShuffleNetV2	ShuffleNetV1	MobileNetV2
KD	0.34	0.16	0.11	0.47	0.27	0.22
AT	0.21	0.94	0.17	0.13	0.25	0.26
PKT	0.18	0.17	0.27	0.18	0.14	0.35
SP	0.17	0.19	0.29	0.18	0.14	0.28
CC	0.42	0.20	0.17	0.17	0.19	0.39
RKD	0.49	0.27	0.18	0.29	0.31	0.10
VID	0.19	0.17	0.23	0.25	0.17	0.36
CRD	0.69	0.17	0.22	0.10	0.27	0.13
REVIEWKD	0.26	0.85	0.31	0.14	0.32	0.46
DKD	0.35	0.13	0.14	0.19	0.45	0.08
HSAKD	0.26	0.19	0.09	0.16	0.42	0.11

References

1. Ahn, S., Hu, S., Damianou, A., Lawrence, N., Dai, Z.: Variational information distillation for knowledge transfer pp. 9155–9163 (06 2019). https://doi.org/10.1109/CVPR.2019.00938
2. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5008–5017 (2021)
3. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
4. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation pp. 3967–3976 (2019)
5. Passalis, N., Tefas, A.: Learning deep representations with probabilistic knowledge transfer (2018)
6. Peng, B., Jin, X., li, D., Zhou, S., Wu, Y., Liu, J., Zhang, Z., Liu, Y.: Correlation congruence for knowledge distillation pp. 5006–5015 (10 2019). https://doi.org/10.1109/ICCV.2019.00511
7. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation (2020)
8. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 1365–1374 (2019), https://api.semanticscholar.org/CorpusID:198179476
9. Yang, C., An, Z., Cai, L., Xu, Y.: Hierarchical self-supervised augmented knowledge distillation. arXiv preprint arXiv:2107.13715 (2021)
10. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer (2017), https://arxiv.org/abs/1612.03928
11. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. arXiv preprint arXiv:2203.08679 (2022)