Improving Ensemble Distillation With Weight Averaging and Diversifying Perturbation

Giung Nam[1] Hyungi Lee[1] Byeongho Heo[2] Juho Lee[1,3]

[1]KAIST, South Korea

[2]Naver, South Korea

[3]AITRICS, South Korea

Ensembles of Deep Neural Networks

Deep Ensemble [DE; Lakshminarayanan et al., 2017]

- Let $\mathcal{F} : \boldsymbol{x} \mapsto \boldsymbol{p}_{\text{Cat}}$ be a neural network for a K-way classification.
- DE ensembles the same model trained with different seeds,

$$\boldsymbol{p}_{\text{Cat}} \coloneqq \frac{1}{M} \sum_{m=1}^{M} \mathcal{F}_{\boldsymbol{\theta}_m}(\boldsymbol{x}), \quad \text{where } \boldsymbol{\theta}_m \xleftarrow[\text{SGD}]{\text{random init.}} \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{D}). \quad (1)$$

- The power of the ensemble comes from *the diversity* among ensemble members, and there have been several works to enhance it [Wenzel et al., 2020, Zaidi et al., 2021, Rame and Cord, 2021, D'Angelo and Fortuin, 2021].

Knowledge Distillation [KD; Hinton et al., 2015]

- The seminal work of Hinton et al. [2015]:

> Training an ensemble of models is a very simple way to take advantage of parallel computation and *the usual objection that an ensemble requires too much computation at test time can be dealt with by using distillation*.

- However, the diversities among ensemble teachers are removed by *mean* operation and hardly transferred to the student, i.e.,

$$\begin{aligned} \text{minimize} \quad & \mathcal{H}\left[\frac{1}{M}\sum_{m=1}^{M}\mathcal{T}_m(\boldsymbol{x}),\ \mathcal{S}_{\boldsymbol{\theta}}(\boldsymbol{x})\right], \\ \text{where} \quad & \mathcal{S}_{\boldsymbol{\theta}} \text{ is a single student network.} \end{aligned} \tag{2}$$

Ensemble Distillation

BatchEnsemble [BE; Wen et al., 2020]

- BE aims to address the computational and memory bottleneck of DE.
- To summarize (with a slight abuse of notation),

$$\begin{aligned} \text{DE:} \quad & \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M\}, \\ \text{BE:} \quad & \{\boldsymbol{\theta} \circ (\boldsymbol{r}_1 \boldsymbol{s}_1^\top), ..., \boldsymbol{\theta} \circ (\boldsymbol{r}_M \boldsymbol{s}_M^\top)\}. \end{aligned} \tag{3}$$

BatchEnsemble and one-to-one distillation [Mariet et al., 2021]

- Distilling DE teacher into BE student, i.e.,

$$\begin{aligned} \text{minimize} \quad & \sum_{m=1}^{M} \mathcal{H}\left[\mathcal{T}_m(\boldsymbol{x}),\ \mathcal{S}_m(\boldsymbol{x})\right], \\ \text{where} \quad & \mathcal{S}_m \text{ is the } m^{\text{th}} \text{ subnetwork of BE.} \end{aligned} \tag{4}$$

- As a result, BE mimics the DE at a lower cost. Nevertheless, in principle, BE still requires multiple forward passes for inference.

(#1) LatentBE : a weight avearged BE student

- In principle, BE requires multiple forward passes for inference,

$$\frac{1}{2}\left(\boldsymbol{p}_{\mathcal{S}_{\boldsymbol{\theta}_1}}(\boldsymbol{x}) + \boldsymbol{p}_{\mathcal{S}_{\boldsymbol{\theta}_2}}(\boldsymbol{x})\right), \quad \text{where} \begin{cases} \boldsymbol{\theta}_1 \leftarrow \boldsymbol{\theta} \circ (\boldsymbol{r}_1 \boldsymbol{s}_1^\top), \\ \boldsymbol{\theta}_2 \leftarrow \boldsymbol{\theta} \circ (\boldsymbol{r}_2 \boldsymbol{s}_2^\top). \end{cases} \tag{5}$$

- LatentBE only requires *a single forward pass* for inference,

$$\boldsymbol{p}_{\mathcal{S}_{\boldsymbol{\theta}}}(\boldsymbol{x}), \quad \text{where } \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} \circ (\boldsymbol{r}_1 \boldsymbol{s}_1^\top + \boldsymbol{r}_2 \boldsymbol{s}_2^\top). \tag{6}$$

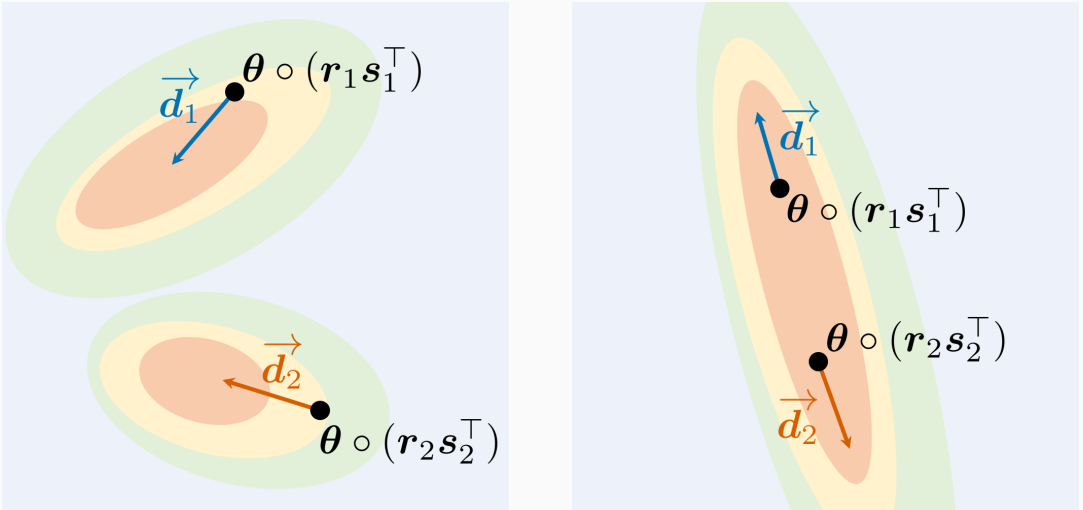

Figure 1: A schematic diagram depicting BE (left) and LatentBE (right).

(#2) A better perturbation strategy

- Nam et al. [2021] utilizes perturbation strategies, Output Diversified Sampling [ODS; Tashiro et al., 2020], to improve diversity transfer,

$$\boldsymbol{\varepsilon}_{\text{ODS}} \propto \nabla_{\boldsymbol{x}} \big(\boldsymbol{w}^{\top} \mathcal{T}_m(\boldsymbol{x}) \big), \tag{7}$$

  where $\boldsymbol{w}$ denotes a random guidance vector.

- We suggest perturbing input to the direction that *minimizes* the student diversity while the teacher diversity is *maximized*,

$$\boldsymbol{\varepsilon} \propto \nabla_{\boldsymbol{x}} \big(\operatorname{TDiv}(\boldsymbol{x}) - \operatorname{SDiv}(\boldsymbol{x}) \big), \tag{8}$$

  where $\operatorname{TDiv}(\boldsymbol{x})$ denotes the teacher diversity and $\operatorname{SDiv}(\boldsymbol{x})$ denotes the student diversity.

Results on Image Classification Tasks

Method	ACC (↑)	NLL (↓)	ECE (↓)	cNLL (↓)	cECE (↓)
DE-4 teacher	94.71	0.170	0.006	0.168	0.010
KD [Hinton et al., 2015]	93.70$\pm$0.10	0.270$\pm$0.004	0.042$\pm$0.001	0.201$\pm$0.003	0.011$\pm$0.002
AE-KD [Du et al., 2020]	93.67$\pm$0.05	0.278$\pm$0.002	0.042$\pm$0.000	0.205$\pm$0.001	0.010$\pm$0.001
Proxy-EnD^2 [Ryabinin et al., 2021]	93.67$\pm$0.04	0.270$\pm$0.005	0.042$\pm$0.001	0.200$\pm$0.002	0.011$\pm$0.001
KD + LatentBE (Ours)	**93.98**$\pm$0.20	0.263$\pm$0.003	0.041$\pm$0.002	0.194$\pm$0.002	0.011$\pm$0.002
+ ConfODS	93.95$\pm$0.12	0.223$\pm$0.007	0.032$\pm$0.001	0.186$\pm$0.004	**0.008**$\pm$0.001
+ TDiv-SDiv	93.95$\pm$0.01	**0.205**$\pm$0.006	**0.028**$\pm$0.001	**0.181**$\pm$0.005	**0.008**$\pm$0.002

Table 1: Results on CIFAR-10 with DE-4 teacher.

Method	ACC (↑)	NLL (↓)	ECE (↓)	cNLL (↓)	cECE (↓)
DE-4 teacher	81.37	0.706	0.031	0.700	0.018
KD [Hinton et al., 2015]	79.09$\pm$0.20	1.038$\pm$0.011	0.130$\pm$0.003	0.861$\pm$0.007	0.046$\pm$0.001
AE-KD [Du et al., 2020]	79.00$\pm$0.39	1.033$\pm$0.009	0.129$\pm$0.004	0.859$\pm$0.008	0.045$\pm$0.004
Proxy-EnD^2 [Ryabinin et al., 2021]	78.75$\pm$0.28	1.076$\pm$0.016	0.138$\pm$0.002	0.886$\pm$0.011	0.046$\pm$0.003
KD + LatentBE (Ours)	79.46$\pm$0.20	0.993$\pm$0.024	0.124$\pm$0.004	0.837$\pm$0.012	0.046$\pm$0.005
+ ConfODS	79.27$\pm$0.30	0.955$\pm$0.008	0.115$\pm$0.002	0.840$\pm$0.007	0.048$\pm$0.004
+ TDiv-SDiv	**80.02**$\pm$0.07	**0.792**$\pm$0.004	**0.067**$\pm$0.001	**0.772**$\pm$0.003	**0.041**$\pm$0.003

Table 2: Results on CIFAR-100 with DE-4 teacher.

Results on Image Classification Tasks

More experimental results are available in the paper:

- Additional experiments on CIFAR-10/100 further verify effectiveness of our approach; (1) robustness to common corruptions [Hendrycks and Dietterich, 2019], and (2) predictive uncertainty on out-of-distribution data.
- We also verify the scalability of our approach on large-scale datasets, including TinyImageNet and ImageNet-1k [Russakovsky et al., 2015].
- Analysis of training and testing runtime further clarifies the difference between BE and LatentBE.

References

Francesco D'Angelo and Vincent Fortuin. Repulsive deep ensembles are bayesian. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

Shangchen Du, Shan You, Xiaojie Li, Jianlong Wu, Fei Wang, Chen Qian, and Changshui Zhang. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations (ICLR)*, 2019.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *Deep Learning and Representation Learning Workshop, NIPS 2014*, 2015.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

Zelda E Mariet, Rodolphe Jenatton, Florian Wenzel, and Dustin Tran. Distilling ensembles improves uncertainty estimates. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021.

Giung Nam, Jongmin Yoon, Yoonho Lee, and Juho Lee. Diversity matters when learning from ensembles. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

Alexandre Rame and Matthieu Cord. Dice: Diversity in deep ensembles via conditional redundancy adversarial estimation. In *International Conference on Learning Representations (ICLR)*, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015.

Max Ryabinin, Andrey Malinin, and Mark Gales. Scaling ensemble distribution distillation to many classes with proxy targets. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

Y. Tashiro, Y. Song, and S. Ermon. Diversity can be transferred: Output diversification for white- and black-box attacks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

Sheheryar Zaidi, Arber Zela, Thomas Elsken, Chris C Holmes, Frank Hutter, and Yee Teh. Neural ensemble search for uncertainty estimation and dataset shift. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.