Peer Collaborative Learning for Online Knowledge Distillation

Guile Wu and Shaogang Gong
Queen Mary University of London
2021 AAAI

Du Shangchen
2021/03/17
Knowledge Distillation (KD)\textsuperscript{[1]}
Online KD

• self-distillation
• mutual/ collaborative learning
Online KD

- self-distillation / teacher-free distillation
  - self-distillation\(^2\)
  - born-again network\(^3\)

![Diagram](image.png)
Online KD

• self-distillation
• mutual/ collaborative learning
  • DML[4]
  • CL[5]
  • ONE[6]
  • OKDDip[7]
Problems

• collaborative learning and mutual learning fail to construct an online high-capacity teacher
• online ensembling ignores the collaboration among branches and its logit summation impedes the further optimisation of the ensemble teacher.
Methods

• a multi-branch network (each branch is a peer)
• assemble the features from peers with an additional classifier as the peer ensemble teacher
• employ the temporal mean model of each peer as the peer mean teacher
# Peer Ensemble Teacher

|                     | former work                                      | innovation          |
|---------------------|--------------------------------------------------|---------------------|
| augmentation        | applying random augmentation only *once*         | *m* times           |
| ensemble            | **logits**: logits from multiple networks / branches are usually summed | **features**: concatenate the features from peers and use an additional fully connected layer for classification |
| loss                | fixed weight                                     | weight ramp-up function to control the gradient magnitude. |
Peer Mean Teacher

- use temporal mean models of each peer as the peer mean teacher for peer collaborative distillation.

\[
\begin{align*}
\theta_{l,g}^t &= \phi(g) \cdot \theta_{l,g}^{t-1} + (1 - \phi(g)) \cdot \theta_{l,g} \\
\theta_{h,j,g}^t &= \phi(g) \cdot \theta_{h,j,g}^{t-1} + (1 - \phi(g)) \cdot \theta_{h,j,g} \\
\phi(g) &= \min(1 - \frac{1}{g}, \beta)
\end{align*}
\]

- \(g\) – epoch
- \(l\) – low level
- \(h\) – high level
- \(j\) – \(j\)-th classifier
- \(\beta\) – smoothing coefficient function
Problems

• collaborative learning and mutual learning fail to construct an online high-capacity teacher

• online ensembling ignores the collaboration among branches and its logit summation impedes the further optimisation of the ensemble teacher.

Peer Ensemble Teacher

Peer Mean Teacher
# Experiments

Table 1. Comparisons with the state-of-the-arts on CIFAR-10. Top-1 error rates (%).

| Network          | DML [28]  | CL [21]  | ONE [13] | FFL-S [10] | OKDDip [1] | Baseline | PCL(ours) |
|------------------|-----------|----------|----------|------------|------------|----------|----------|
| ResNet-32        | 6.06±0.07 | 5.98±0.28| 5.80±0.12| 5.99±0.11  | 5.83±0.15  | 6.74±0.15| 5.67±0.12|
| ResNet-110       | 5.47±0.25 | 4.81±0.11| 4.84±0.30| 5.28±0.06  | 4.86±0.10  | 5.01±0.10| 4.47±0.16|
| VGG-16           | 5.87±0.07 | 5.86±0.15| 5.86±0.23| 6.78±0.08  | 6.02±0.06  | 6.04±0.13| 5.26±0.02|
| DenseNet-40-12   | 6.41±0.26 | 6.95±0.25| 6.92±0.21| 6.72±0.16  | 7.36±0.22  | 6.81±0.02| 5.87±0.13|
| WRN-20-8         | 4.80±0.13 | 5.41±0.08| 5.30±0.14| 5.28±0.13  | 5.17±0.15  | 5.32±0.01| 4.58±0.04|
| ResNeXt-29-2×64d | 4.46±0.16 | 4.45±0.18| 4.27±0.10| 4.67±0.04  | 4.34±0.02  | 4.72±0.03| 3.93±0.09|

Table 2. Comparisons with the state-of-the-arts on CIFAR-100. Top-1 error rates (%).

| Network          | DML [28]  | CL [21]  | ONE [13] | FFL-S [10] | OKDDip [1] | Baseline | PCL(ours) |
|------------------|-----------|----------|----------|------------|------------|----------|----------|
| ResNet-32        | 26.32±0.14| 27.67±0.46| 26.21±0.41| 27.82±0.11  | 26.75±0.38  | 28.72±0.19| 25.86±0.16|
| ResNet-110       | 22.14±0.50| 21.17±0.58| 21.60±0.36| 22.78±0.41  | 21.46±0.26  | 23.79±0.57| 20.02±0.55|
| VGG-16           | 24.48±0.10| 25.67±0.08| 25.63±0.39| 29.13±0.99  | 25.32±0.05  | 25.68±0.19| 23.11±0.25|
| DenseNet-40-12   | 26.94±0.31| 28.55±0.34| 28.40±0.38| 28.75±0.35  | 28.77±0.14  | 28.97±0.15| 26.91±0.16|
| WRN-20-8         | 20.23±0.07| 20.60±0.12| 20.90±0.39| 21.78±0.14  | 21.17±0.06  | 21.97±0.40| 19.49±0.49|
| ResNeXt-29-2×64d | 18.94±0.01| 18.41±0.07| 18.60±0.25| 20.18±0.33  | 18.50±0.11  | 20.57±0.43| 17.38±0.23|
Ablation

• Comparison with Two-Stage Distillation

| Dataset    | Baseline | KD↑ | PCL    |
|------------|----------|-----|--------|
| CIFAR-10   | 6.74±0.15| 5.82±0.12 | 5.67±0.12 |
| CIFAR-100  | 28.72±0.19 | 26.23±0.21 | 25.86±0.16 |

• branch num

• augmentation
Reference

[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[2] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, “Be your own teacher: Improve the performance of convolutional neural networks via self distillation,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3713–3722.

[3] T. Furlanello, Z. C. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, “Born again neural networks,” *ICML*, 2018.

[4] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, “Deep mutual learning,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4320–4328.

[5] Guocong Song and Wei Chai. Collaborative learning for deep neural networks. In Advances in Neural Information Processing Systems, pages 1832–1841, 2018.

[6] X. Lan, X. Zhu, and S. Gong, “Knowledge distillation by on-the-fly native ensemble,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2018, pp. 7528–7538.

[7] D. Chen, J.-P. Mei, C. Wang, Y. Feng, and C. Chen, “Online knowledge distillation with diverse peers,” Association for the Advancement of Artificial Intelligence, 2020.