Live on the Hump: Self Knowledge Distillation via Virtual Teacher-Students Mutual Learning

Shuang Wang, Pengyi Hao*, Fuli Wu, Cong Bai*

Zhejiang University of Technology

ACM Multimedia 2024
Melbourne, Australia

Background

Self Knowledge Distillation:

- ***Peer Self KD*:** Sharing the same shallowed backbone to construct multiple peer auxiliary exits in the last segment of the backbone.
  - ✗ Multiple exits in the same location typically tend to cause homogenization.
- ***Hierarchical Self KD*:** Adding auxiliary exits hierarchically in the deep supervised manner.
  - ✗ Limited knowledge extracted from the deepest layer only.
  - ✗ A single auxiliary exit design ignores the impact of different network sizes.

Method

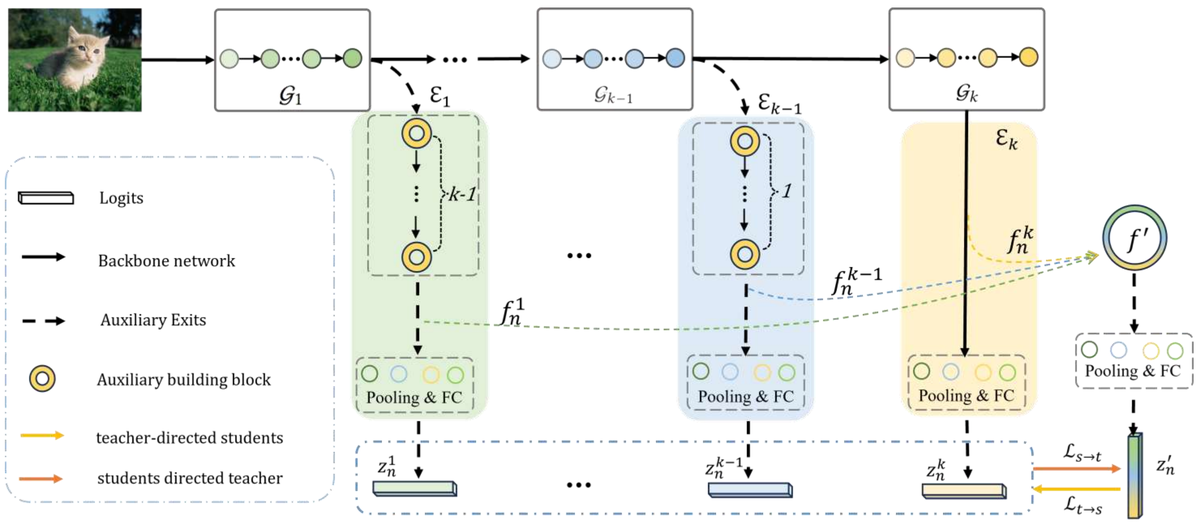

Figure 1: The overall architecture of LOTH. The dashed lines indicate auxiliary exits, which can be removed during the inference phase.

- LOTH constructs auxiliary exits hierarchically, which views each exit as a student and incorporates knowledge from auxiliary exits to build a knowledgeable virtual teacher.

α: Balance factor

Cross Entropy

$$\mathcal{L}_{t\to s} = \sum_{i=1}^{k}(2-\alpha^{k+1-i})\cdot\mathcal{L}_c(y_n, z_n^i) + \alpha^{k+1-i}\cdot\mathcal{L}_d(z_n^{'}, z_n^i)$$

KL divergence

Ensemble logit from each exits

$$\mathcal{L}_{s\to t} = \mathcal{L}_c(y_n, z_n^{'}) + \mathcal{L}_d(\bar{z}_n, z_n^{'})$$

Teacher's logit

- The bidirectional mutual learning between virtual teacher and students contributes significantly to the capabilities of multi-exits learning with few training overhead.

Results & Discussions

Table 1: Top-1 classification accuracy and parameter statistics of LOTH on CIFAR-100.

Networks	Baseline	Exit1		Exit2		Exit3		Exit4		Fusion
		Acc(%)	Param(M)	Acc(%)	Param(M)	Acc(%)	Param(M)	Acc(%)	Param(M)	
VGG16	73.54	76.31	6.64	76.46	7.43	76.87	10.58	**76.94** (↑ 3.40)	15.30	78.68
VGG19	73.34	75.23	6.64	75.73	8.02	**76.09**	13.53	76.06 (↑ 2.72)	20.61	77.83
ResNet18	77.65	78.47	3.82	79.53	4.17	80.90	5.58	**81.24** (↑ 3.59)	11.22	82.08
ResNet34	78.00	78.37	3.89	80.53	4.84	81.84	10.97	**81.96** (↑ 3.96)	21.33	82.77
MobileNetV1	73.40	76.99	2.22	77.40	2.23	78.62	3.30	**79.02** (↑ 5.62)	3.31	80.62
MobileNetV2	72.22	76.46	2.93	76.79	2.78	76.95	2.45	**77.51** (↑ 5.29)	2.35	80.13
ShuffleNetV1	71.39	74.31	1.91	75.31	1.89	**76.69**	1.85	76.35 (↑ 4.96)	1.01	79.06
ShuffleNetV2	71.85	73.04	1.51	74.82	1.52	75.93	1.68	**76.03** (↑ 4.18)	1.36	79.14

Table 1: Top-1 classification of ResNet18 on ImageNet.

Models	BYOT	ECSD	BEED	DTSKD	LOTH
Years	2019	2021	2022	2024	2024
Acc(%)	69.84	70.51	70.28	70.39	**70.74**

Table 2: Top-1 classification accuracy on CIFAR100.

Backbones	Baseline	Exit1	Exit2	Exit3	Exit4	Fusion
ResNet101	78.64	77.27	79.68	82.77	82.41	83.17
ResNet152	79.64	77.58	80.74	82.88	83.18	83.58

Table 2: Top-1 classification accuracy and parameter statistics of LOTH on Tiny-ImageNet.

Networks	Baseline	Exit1		Exit2		Exit3		Exit4		Fusion
		Acc(%)	Param(M)	Acc(%)	Param(M)	Acc(%)	Param(M)	Acc(%)	Param(M)	
VGG16	50.38	54.22	6.70	54.74	7.48	**55.02**	10.63	54.63 (↑ 4.25)	15.35	57.97
VGG19	48.34	54.45	6.70	54.88	8.07	**54.89**	13.58	54.33 (↑ 5.99)	20.66	57.76
ResNet18	57.20	59.34	3.87	61.25	4.22	61.49	5.63	**61.65** (↑ 4.45)	11.27	63.92
ResNet34	59.54	59.10	3.94	61.13	4.89	62.39	11.02	**63.35** (↑ 3.81)	21.38	64.96
MobileNetV1	52.64	56.44	2.33	57.00	2.33	58.48	3.41	**59.22** (↑ 6.58)	3.41	61.81
MobileNetV2	51.61	55.27	3.06	55.92	2.91	56.66	2.58	**57.04** (↑ 5.43)	2.48	59.98
ShuffleNetV1	51.25	53.80	2.01	55.56	1.98	**56.15**	1.94	55.68 (↑ 4.43)	1.11	58.47
ShuffleNetV2	51.84	49.91	1.61	52.45	1.62	55.61	1.78	**56.47** (↑ 4.63)	1.46	58.11

- *LOTH performs well on various network architectures and multi-datasets.*
- *LOTH exceeds the performance of almost all baselines at the shallowest exit.*
- *Virtual teacher (Fusion) has richer knowledge, which far exceeds other exits.*

Table 4: Top-1 accuracy and parameter statistics of LOTH VS. advanced SKDs with hierarchical exits in ResNet18 on Tiny-ImageNet.

Methods	Supervision	Exit1		Exit2		Exit3		Exit4		Fusion
		Acc(%)	Params(M)	Acc(%)	Params(M)	Acc(%)	Params(M)	Acc(%)	Params(M)	
BYOT	Deepest	44.83	2.91	53.26	3.47	57.80	5.63	58.97	11.27	61.32
ECSD	Ensemble-Avg	47.09	0.44	53.91	0.97	57.34	3.08	59.03	11.27	60.51
BEED	Ensemble-Weight	59.20	3.87	60.25	4.22	60.52	5.63	61.11	11.27	64.13
LOTH	Mutual Learning	**59.44**	3.87	**61.25**	4.22	**61.49**	5.63	**61.65**	11.27	63.92

- *The deepest knowledge alone fails to provide wealth of knowledge, which gets the worst performance.*
- *Mutual learning can enhance the learning capabilities at all exits.*

Table 5: Top-1 accuracy comparison of different fusion mechanism in MobileNetV1 on Tiny-ImageNet.

Fuse type	Exit1	Exit2	Exit3	Exit4	Fusion
Summation	54.65	55.33	56.70	56.78	59.45
Concatenation	56.79	57.49	58.40	58.75	60.93
Att-Sample	56.87	57.19	58.79	58.84	61.30
Ours	56.44	57.00	58.48	**59.22**	**61.81**

- *Our adaptive fusion strategy can mitigate feature semantic gaps with multi-exits, facilitating the distillation efficiency.*

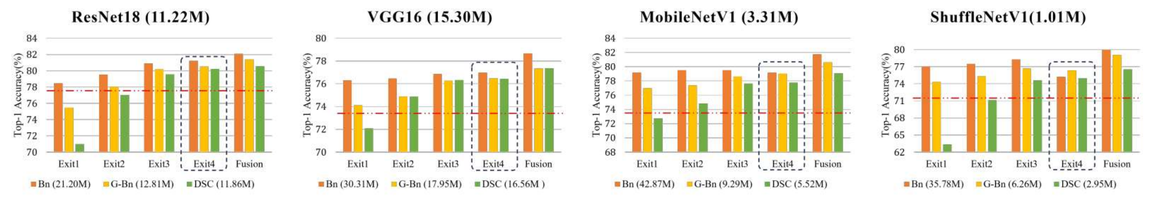

Figure 4: The effect of different auxiliary blocks for model performance on the CIFAR-100, where the red dotted line indicates the baseline. The bracketed values indicate the total trained parameters of multi-exits and the backbone networks.

- *Excessive auxiliary exit scales can divert training targets and reduce backbone network's performance(Exit4).*

Conclusions

- **A novel self knowledge distillation framework via virtual teacher-student mutual learning been proposed, which focus on exploiting the complementary knowledge of early exits to further enhance the effective ness of distillation.**
- **Efficient adaptive fusion strategy can mitigate the semantic gaps between multi-exits, resulting in a knowledgeable virtual teacher.**
- **The scale of auxiliary exits affects the performance of model, and our well-designed two auxiliary blocks can balance effectiveness and efficiency.**

References

[1] BYOT---Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the ICCV.*

[2] ECSD---Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. 2021. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 8 (2021).

[3] BEED---Hojung Lee and Jong-Seok Lee. 2022. Rethinking Online Knowledge Distillation with Multi-exits. In *Proceedings of the ACCV. 2289–2305* .

[4] DTSKD---Zheng Li, Xiang Li, Lingfeng Yang, Renjie Song, Jian Yang, and Zhigeng Pan. 2024. Dual teachers for self-knowledge distillation. *Pattern Recognition* (2024), 110422 .