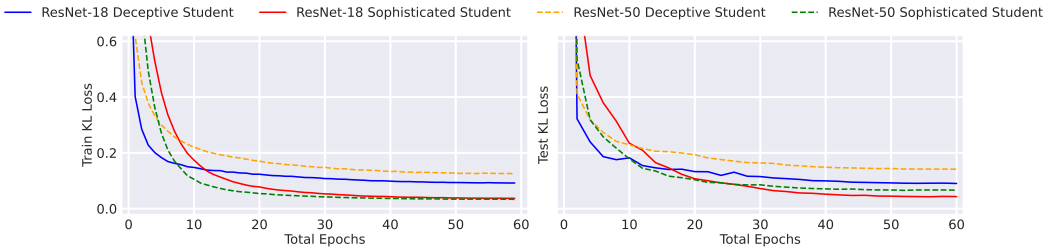

Figure 5: Training and test KL-divergence losses of student models in LoT using ResNet-18 and ResNet-50 on CIFAR-100 with different teacher models.

Table 8: Performance of LoT and Teacher-only on ImageNet-R and ImageNet-Sketch.

Dataset	Teacher	Student	Teacher-only / LoT
ImageNet-R	ViT-B/16	ViT-B/16	49.11 / 52.27
ImageNet-R	ViT-B/16	ViT-L/16	49.11 / 54.08
ImageNet-R	ViT-L/16	ViT-B/16	54.42 / 58.18
ImageNet-R	ViT-L/16	ViT-L/16	54.42 / 57.79
ImageNet-Sketch	ViT-B/16	ViT-B/16	38.85 / 41.46
ImageNet-Sketch	ViT-B/16	ViT-L/16	38.85 / 42.89
ImageNet-Sketch	ViT-L/16	ViT-B/16	43.83 / 47.61
ImageNet-Sketch	ViT-L/16	ViT-L/16	43.83 / 45.91

Table 9: Performance of LoT, BAN, ReviewKD, DKD on CIFAR100.

Method	Teacher	Student	Accuracy
Teacher-only	ResNet-50	N/A	84.09
BAN	ResNet-50	ResNet-50	84.73
ReviewKD	ResNet-50	ResNet-50	85.31
DKD	ResNet-50	ResNet-50	85.17
LoT	ResNet-50	ResNet-50	86.04
Teacher-only	ViT-B/16	N/A	91.57
BAN	ViT-B/16	ViT-B/16	92.44
ReviewKD	ViT-B/16	ViT-B/16	92.73
DKD	ViT-B/16	ViT-B/16	92.82
LoT	ViT-B/16	ViT-B/16	93.17

Table 10: Test/Validation perplexity of LoT and Teacher-only on the official test/validation datasets.

Dataset	Teacher	Student	Teacher-only (Valid)	Teacher-only (Test)	LoT (Valid)	LoT (Test)
PTB	LSTM	LSTM	86.02	82.75	73.98	71.72
PTB	AWD-LSTM	AWD-LSTM	60.62	58.69	55.07	53.31
Wikitext-103	Transformer-XL-B	Transformer-XL-B	24.68	23.72	22.24	21.65
Wikitext-103	Transformer-XL-L	Transformer-XL-L	18.65	18.50	16.41	16.47

Table 11: The performance of student models in LoT on lanugage modeling and image classification.

Task	Dataset	Teacher	Student	Teacher-only	LoT (Teacher)	LoT (Student)
Language Modeling	PTB	LSTM	LSTM	82.75	71.72	73.33
Language Modeling	WikiText-103	Transformer-XL-L	Transformer-XL-L	18.50	16.47	16.89
Image Classification	CIFAR100	ResNet-50	ResNet-18	84.09	85.77	83.24
Image Classification	CIFAR100	ResNet-50	ResNet-50	84.09	86.04	85.72
Image Classification	ImageNet-1K	ViT-B/16	ViT-B/16	91.57	93.17	92.95
Image Classification	ImageNet-1K	ViT-B/16	ViT-L/16	91.57	93.25	93.89

Table 12: Computational resources, memory usage, and training time of LoT and Teacher-only.

Dataset	Teacher Model / Student Model	Total Train Steps (teacher+student)	Computational Resources	CPU Usage (Teacher-only/LoT)	GPU Usage (Teacher-only/LoT)	Training Time (Teacher-only/LoT)	Performance (Teacher-only/LoT)
BeamRider	Standard Network / Standard Network	20M frames	1 NVIDIA A6000 48GB GPU	16 core / 16 core	0.8 GB / 0.9 GB	10 h / 10.1 h	3,651 score / 5,956 score (↑)
PTB	LSTM / LSTM	60 epochs	1 × NVIDIA A100 40GB GPU	1 core / 1 core	1.1 GB / 1.5 GB	0.6 h / 0.3 h	82.8 ppl / 71.7 ppl (↓)
WikiText-103	Transformer-XL-L / Transformer-XL-L	0.3M steps	4 × NVIDIA A100 40GB GPU	4 core / 4 core	4 × 21.4 GB / 4 × 33.2 GB	85.6 h / 67.7 h	18.5 ppl / 16.5 ppl (↓)
GSM8K	LLaMA-2 7B / LLaMA-2 7B	4 epochs	8 × NVIDIA A100 40GB GPU	8 core / 8 core	8 × 27.4 GB / 8 × 39.8 GB	8.1 h / 6.7 h	39.8 acc / 41.9 acc (↑)
CIFAR100	ResNet-50 / ResNet-18	60 epochs	1 × NVIDIA A100 40GB GPU	1 core / 1 core	13.6 GB / 16.7 GB	0.7 h / 0.5 h	84.1 acc / 85.8 acc (↑)
ImageNet-1K	ViT-L/16 / ViT-B/16	20K steps	4 × NVIDIA A100 40GB GPU	4 core / 4 core	4 × 17.5 GB / 4 × 23.1 GB	28.9 h / 18.7 h	85.2 acc / 86.0 acc (↑)

Table 13: Performance of using L2 loss for the LoT regularizer on CIFAR100.

Dataset	Teacher	Student	Teacher-only	LoT (KL-Divergence)	LoT (L2)
CIFAR100	ViT-B/16	ViT-B/16	91.57	93.17	92.77
CIFAR100	ViT-B/16	ViT-L/16	91.57	93.25	92.94
CIFAR100	ViT-L/16	ViT-B/16	93.44	94.29	94.12
CIFAR100	ViT-L/16	ViT-L/16	93.44	94.18	94.05