Learning from Teaching Regularization: Generalizable Correlations Should be Easy to Imitate

Can Jin[1] · Tong Che[2] · Hongwu Peng[3] · Yiyuan Li[4]

Dimitris N. Metaxas[1] · Marco Pavone[4]

[1]Rutgers University, [2]Nvidia Research,

[3]University of Connecticut, [4]University of North Carolina at Chapel Hill,

[5]Stanford University

Research Question

Among all possible models fitting the training data, which ones are inherently generalizable?

1. brute-force memorization
2. Overfitting

Motivation

- Cognitive Science: a common belief in cognitive science is that human intelligence development involves distilling information and filtering out extraneous details to discern ‘simple’ correlations among a few selected relevant abstract variables
- Emergent Language: more structured a language is, the more efficiently it can be transmitted to message receivers

Hypothesis

Generalizable correlations should be more easily imitable by learners compared to spurious correlations. Specifically, assume T_G and T_S are two teacher models that capture the generalizable correlation and spurious correlation from a dataset, respectively. We have student learners S_G and S_S that separately imitate T_G and T_S:

- From an effectiveness perspective, the final training and test losses of learner S_G after training are typically lower than those of learner S_S.
- From an efficiency perspective, during training, the test losses of learner S_G decrease more rapidly than those of S_S.

Hypothesis

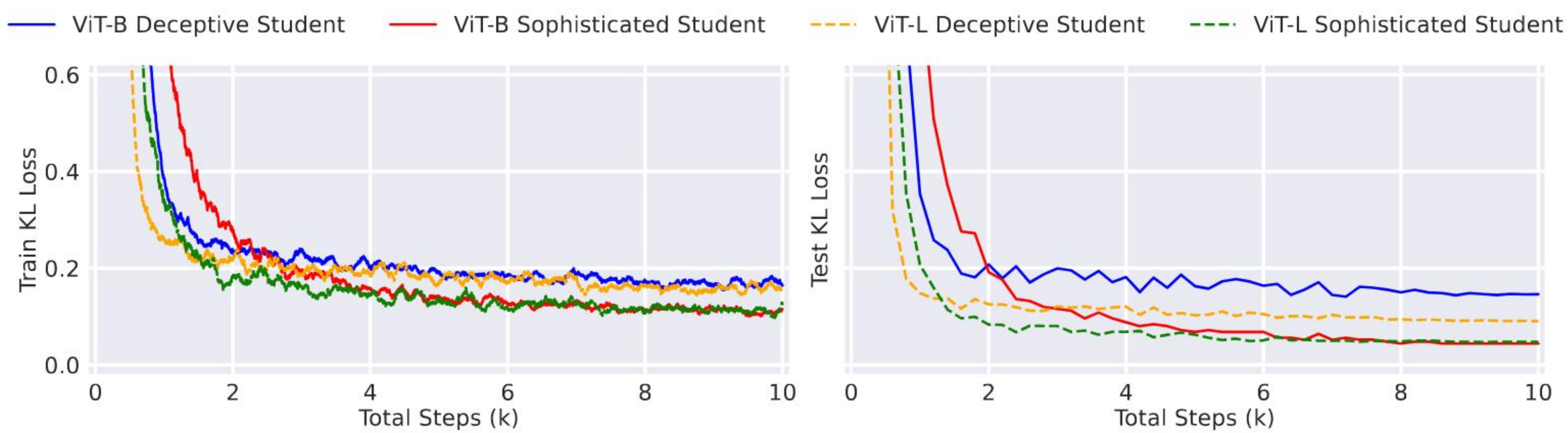


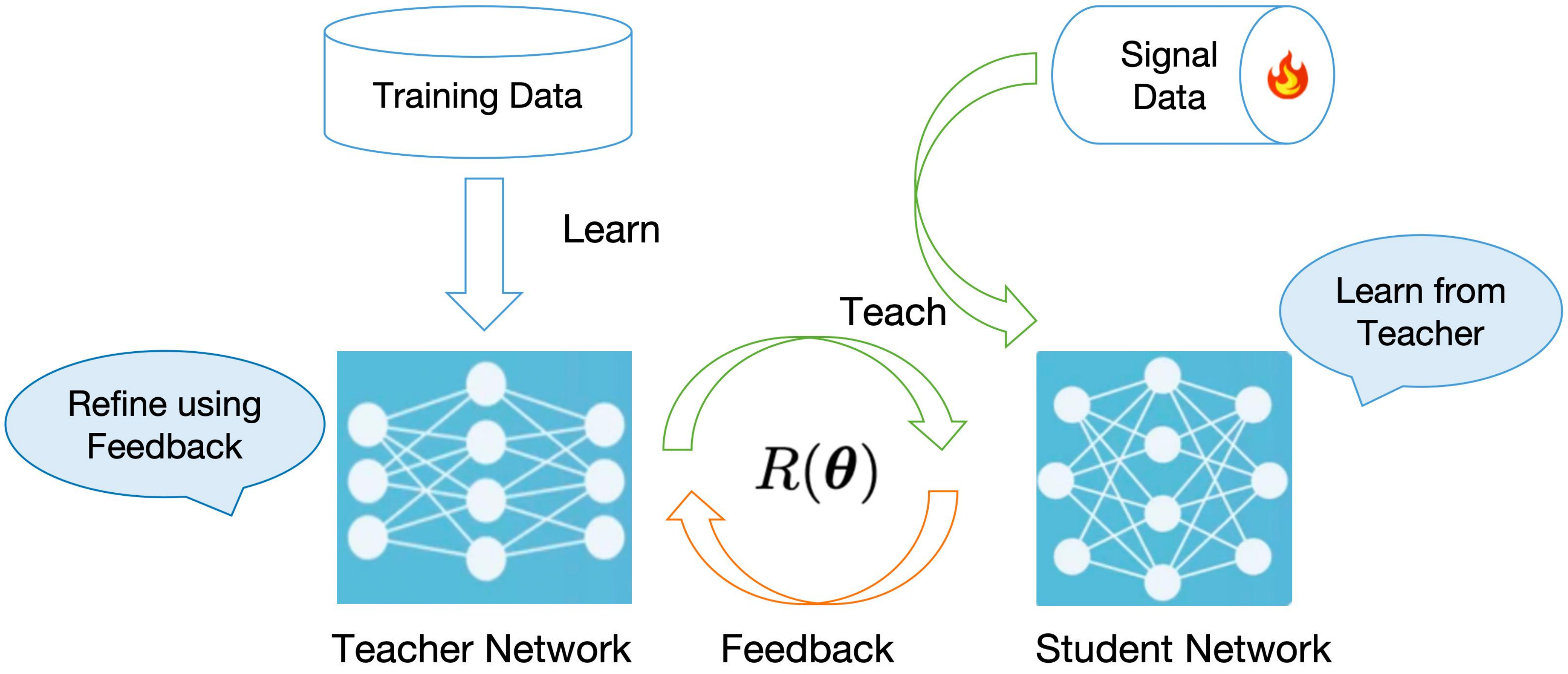
Figure 1: Training and test KL-divergence losses of student models in LoT using ViT-B/16 and ViT-L/16 on CIFAR-100 with different teacher models. The sophisticated students achieve lower losses than the deceptive students given the same computational budget.

LoT Regularizer

We define the Learning from Teaching (LoT) Regularizer to metric the teachability (imitability) of the teacher network.
By optimizing the regularizer, the teacher is optimized to be easier to imitate and, thus, possesses superior generalization compared to models without the LoT regularizer.

$$R(\boldsymbol{\theta}) = \frac{\alpha}{|\mathcal{D}_s|} \sum_{\mathbf{x} \in \mathcal{D}_s} \sum_{i=1}^{K} \lambda_i \mu_{t,s_i}(\mathbf{x})$$

Method Overview



Experiment Results

LoT can enhance the generalization on RL methods

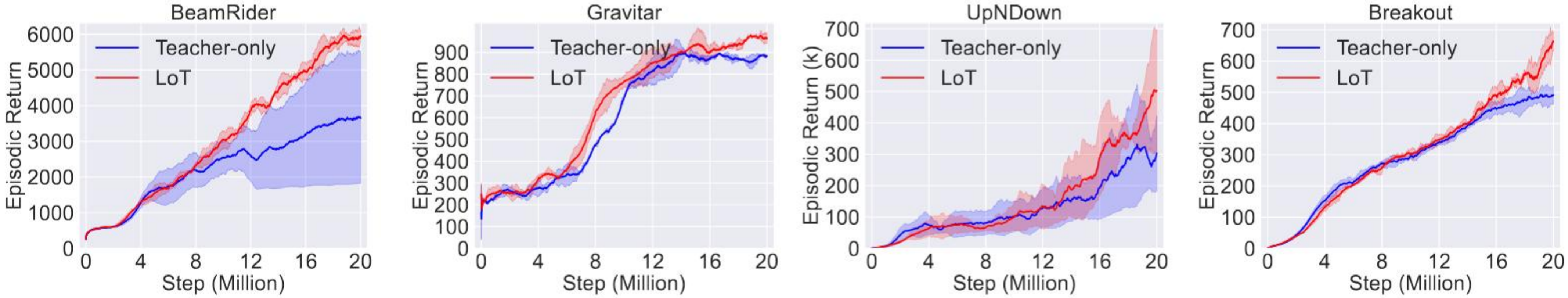


Figure 2: The episodic return of the teacher agent in LoT and the Teacher-only on four Atari games (averaged over ten runs). LoT demonstrates return gains over Teacher-only on all games.

Experiment Results

1. LoT can enhance the generalization on NLP tasks
2. LoT can enhance the generalization of LSTM and Transformers

Table 1: The test perplexity of the teacher model in LoT and the baseline on PTB and WikiText-103. Results are averaged over three runs. LoT achieves consistent perplexity reduction over different choices of architectures and benchmarks.

Dataset	Teacher	Student	Teacher #Param.	Teacher-only	LoT
PTB	LSTM	LSTM	20M	82.75 ± 0.36	$\mathbf{71.72} \pm 0.54$
	AWD-LSTM	AWD-LSTM	24M	58.69 ± 0.37	$\mathbf{53.31} \pm 0.56$
WikiText-103	Transformer-XL-B	Transformer-XL-B	151M	23.72 ± 0.41	$\mathbf{21.65} \pm 0.38$
	Transformer-XL-L	Transformer-XL-L	257M	18.50 ± 0.25	$\mathbf{16.47} \pm 0.23$

Table 2: The accuracy of the teacher model in LoT and the baseline on GSM8K and MATH. Results are averaged over three runs.

Setting	GSM8K	MATH
LLaMA-1 $7B_{+ICL}$	10.69 ± 0.87	2.84 ± 0.25
LLaMA-1 $7B_{+SFT}$	34.39 ± 1.28	4.78 ± 0.23
LLaMA-1 $7B_{+LoT}$	$\mathbf{36.42} \pm 1.46$	$\mathbf{5.39} \pm 0.28$
LLaMA-2 $7B_{+ICL}$	14.62 ± 0.96	2.46 ± 0.25
LLaMA-2 $7B_{+SFT}$	39.81 ± 1.34	5.79 ± 0.31
LLaMA-2 $7B_{+LoT}$	$\mathbf{41.87} \pm 1.62$	$\mathbf{6.28} \pm 0.22$

Experiment Results

1. LoT can enhance the generalization on CV tasks
2. Strong students can enhance the generalization of weak teachers
3. Weak students can futher enhance the generalization of strong teachers

Table 3: The test accuracy of the teacher model for various teacher-student model combinations in LoT and the baseline. Results are averaged over three runs. LoT consistently enhances test performance in all model choices and datasets.

Pretrained	Downstream	Teacher	Student	Image Size	Teacher/Student #Param.	Teacher-only	LoT
ImageNet-1K	CIFAR-100	ResNet-18	MobileNetV2	224^2	12M / 4M	81.14 ± 0.58	$\mathbf{82.78} \pm 0.36$
		ResNet-18	ResNet-18	224^2	12M / 12M	81.14 ± 0.58	$\mathbf{82.89} \pm 0.25$
		ResNet-18	ResNet-50	224^2	12M / 26M	81.14 ± 0.58	$\mathbf{83.13} \pm 0.26$
		ResNet-50	MobileNetV2	224^2	26M / 4M	84.09 ± 0.32	$\mathbf{85.38} \pm 0.44$
		ResNet-50	ResNet-18	224^2	26M / 12M	84.09 ± 0.32	$\mathbf{85.77} \pm 0.19$
		ResNet-50	ResNet-50	224^2	26M / 26M	84.09 ± 0.32	$\mathbf{86.04} \pm 0.38$
ImageNet-21K	CIFAR-100	ViT-B/16	ViT-B/16	384^2	86M / 86M	91.57 ± 0.31	$\mathbf{93.17} \pm 0.35$
		ViT-B/16	ViT-L/16	384^2	86M / 307M	91.57 ± 0.31	$\mathbf{93.25} \pm 0.44$
		ViT-L/16	ViT-B/16	384^2	307M / 86M	93.44 ± 0.28	$\mathbf{94.29} \pm 0.33$
		ViT-L/16	ViT-L/16	384^2	307M / 307M	93.44 ± 0.28	$\mathbf{94.18} \pm 0.26$
ImageNet-21K	ImageNet-1K	ViT-B/16	ViT-B/16	384^2	86M / 86M	83.97 ± 0.11	$\mathbf{84.54} \pm 0.15$
		ViT-B/16	ViT-L/16	384^2	86M / 307M	83.97 ± 0.11	$\mathbf{84.80} \pm 0.08$
		ViT-L/16	ViT-B/16	384^2	307M / 86M	85.15 ± 0.17	$\mathbf{85.92} \pm 0.09$
		ViT-L/16	ViT-L/16	384^2	307M / 307M	85.15 ± 0.17	$\mathbf{85.65} \pm 0.11$
		Swin-B	Swin-B	384^2	88M / 88M	86.37 ± 0.06	$\mathbf{86.68} \pm 0.15$
		Swin-B	Swin-L	384^2	88M / 197M	86.37 ± 0.06	$\mathbf{86.73} \pm 0.14$
		Swin-L	Swin-B	384^2	197M / 88M	87.27 ± 0.11	$\mathbf{87.64} \pm 0.12$
		Swin-L	Swin-L	384^2	197M / 197M	87.27 ± 0.11	$\mathbf{87.59} \pm 0.09$