Deep Classifier Mimicry without Data Access

Steven Braun[1] Martin Mundt[1,2] Kristian Kersting[1,2,3,4]

[1]Dep. of CS, TU Darmstadt [2]hessian.AI [3]DFKI [4]Centre for Cognitive Science, TU Darmstadt

Overview

Problem:

- **Access to pre-trained models** is common, allowing for knowledge distillation in downstream tasks.
- **But: Original training data** is often unavailable, challenging distillation methods that depend on it.

Research Question: *Can we distill knowledge from models without original training data access independent of their architecture?*

Solution: 🎂 CAKE

- **Model-agnostic knowledge distillation** procedure without the need for original training data.
- **Contrastively diffuses** synthetic samples along the decision boundary at different scales.

Knowledge Distillation

Goal: Teach a student to predict like a teacher model.

- **Teacher** f^T: Usually larger model, pre-trained on data, provides soft targets.
- **Student** f^S: Learns to imitate teacher's predictions.
- **Distillation Loss**: Combines true label loss with teacher-student output similarity.

$$\mathcal{L}_{\text{KD}} = \lambda_{\text{true}}\ \text{CE}(y, \boldsymbol{f}^S(\boldsymbol{x})) + \lambda_{\text{soft}}\ \text{CE}(\boldsymbol{f}^T(\boldsymbol{x}), \boldsymbol{f}^S(\boldsymbol{x}))$$

↪ requires original training data $\boldsymbol{x}$ access!

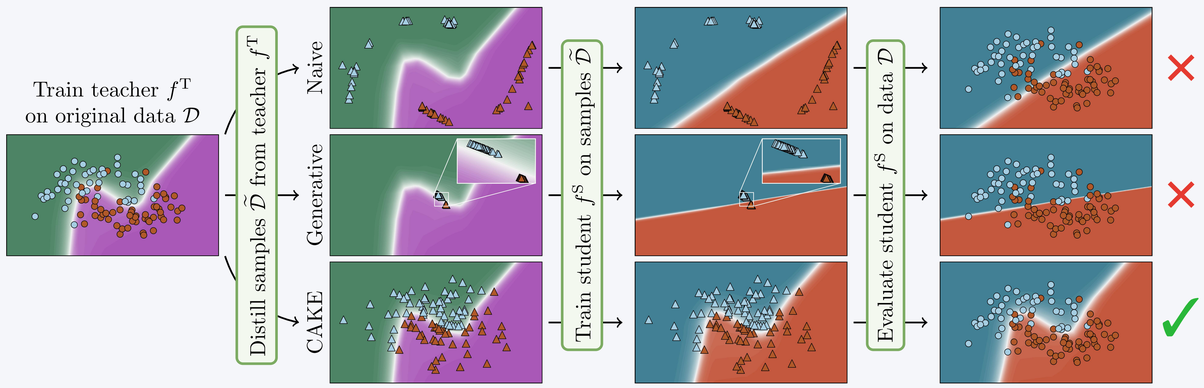
What if we only have access to the pre-trained model?

Paper

Code

Finding the Decision Boundary without Original Training Data

Single-instance KD data synthetization methods usually ignore inter-class relationships and are prone to collapse at locally optimal regions that are suboptimal for student model training.



Challenges

- **"Naive"** $\text{CE}(f^T(\boldsymbol{x}), y)$ pushes samples to the correct boundary sides but too far away
- **"Generative"** (e.g., GAN-based) is prone to collapsing its model parameters to single modes

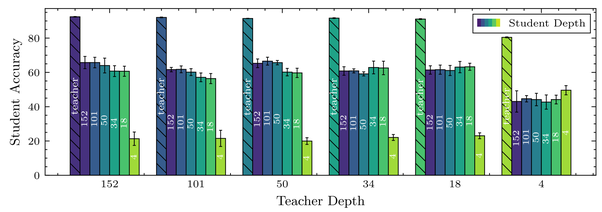
↪ we need to find a balance between pushing samples away from the border while keeping pairs from diff. classes close to each other and noisily scatter these pairs along the boundary at different scales.

CAKE: Contrastive Abductive Knowledge Extraction

- We generate M mini-batches $\tilde{D}_m$ of $\frac{N}{M}$ synthetic samples $(\boldsymbol{x}_i^{t=0}, y_i)$ from chosen priors $p(\boldsymbol{x}), p(y)$.
- For each $\tilde{D}_m$, we perform T gradient descent steps to minimize a weighted objective based on:
  - **Classification Loss**: $\mathcal{L}_{\text{cls}}(\boldsymbol{x}_i^t, y_i) = \text{CE}(y_i, p(\boldsymbol{f}^T(\boldsymbol{x}_i^t)))$
    - pushes samples $\boldsymbol{x}_i^t$ to the correct decision boundary sides according to the sampled class y_i
  - **Contrastive Loss**: $\mathcal{L}_{\text{contr}}\left(\boldsymbol{x}_i^t, \boldsymbol{x}_j^t\right) = \mathbb{1}\left[y_i \neq y_j\right]\ \|\boldsymbol{f}^T(\boldsymbol{x}_i^t) - \boldsymbol{f}^T\left(\boldsymbol{x}_j^t\right)\|_2^2$
    - pulls pairs of samples from diff. classes together, for C classes we get $C(C-1)$ forces/sample
  - **Domain Knowledge Loss** (e.g., Total Var. for images): $\mathcal{L}_{\text{TV}}(\boldsymbol{x}) = \sum_{j,k} \|\boldsymbol{x}_{j,k} - \boldsymbol{x}_{j-1,k}\| + \|\boldsymbol{x}_{j,k} - \boldsymbol{x}_{j,k-1}\|$
    - enables injection of meta-knowledge to constrain the space of relevant samples further
- **Noise injection**: to push samples *along* the decision boundary, we need to additionally disperse them
  Explicit (LAKE): Langevin Dynamics $\boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t + \nabla_x \mathcal{L}(\boldsymbol{x}_i^t)\eta(t) + \sqrt{2\eta(t)}\varepsilon_i^t$ with $\varepsilon_i^t \sim N(0, I)$.
  Implicit (CAKE): Stochasticity of SGD and step size schedule $\eta(t)$ scatters samples across the boundary.

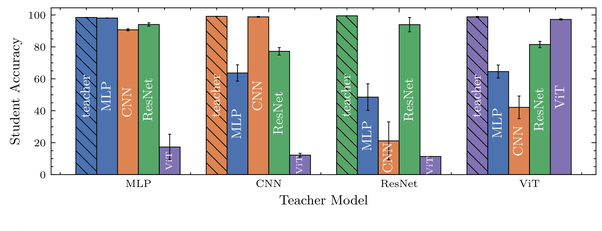
↪ CAKE operates data-free, requiring no access to original training data, and only needs the teacher model to be differentiable, making it completely model-agnostic.

CAKE across Scales



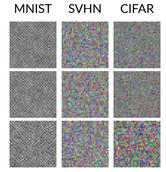
CAKE extracts knowledge and maintains high student accuracy across different model sizes of teachers and students.

CAKE across Models



CAKE enables knowledge transfer across model types with high student accuracy, esp. when matching model types.

Generated Samples



Synthesized samples capture decision boundaries without resembling real data. Possible future work: differential privacy, data utility and privacy trade-offs, robustness against adversarial attacks?

CAKE vs. Others

Method	DF	MA	Dataset	Teacher	Acc.	Student	Acc.
KD	✗	✓	MNIST	LeNet-5	99.3	LeNet-5-Half	98.8
			FMNIST	LeNet-5	90.8	LeNet-5-Half	89.7
			CIFAR-10	ResNet-34	95.6	ResNet-18	94.3
DAFL	✓	✗	MNIST	LeNet-5	97.9	LeNet-5-Half	97.6
			CIFAR-10	ResNet-34	93.7	ResNet18	90.4
DI	✓	✗	CIFAR-10	ResNet-34	95.4	ResNet-18	91.4
ADI	✓	✗	CIFAR-10	ResNet-34	95.4	ResNet-18	93.3
DD	✓	✓	CIFAR-10	ResNet-34	95.4	ResNet-18	30.0
			MNIST	LeNet-5	99.3	LeNet-5-Half	96.5
ZSDB3KD	✓	✓	FMNIST	LeNet-5	91.6	LeNet-5-Half	72.3
			CIFAR-10	AlexNet	79.3	AlexNet-Half	59.5
			MNIST	LeNet-5	99.3 ± 0.12	LeNet-5-Half	98.4 ± 0.18
			FMNIST	LeNet-5	91.0 ± 0.12	LeNet-5-Half	76.5 ± 1.01
			SVHN	LeNet-5	89.8 ± 0.38	LeNet-5-Half	62.9 ± 4.17
CAKE	✓	✓	SVHN	ViT-8	94.4 ± 0.13	ViT-4	83.7 ± 4.77
			SVHN	ResNet-34	96.1 ± 0.08	ResNet-18	94.2 ± 0.54
			CIFAR-10	ViT-8	73.2 ± 0.76	ViT-4	53.8 ± 5.63
			CIFAR-10	ResNet-34	91.8 ± 0.11	ResNet-18	78.9 ± 2.59

Contact: steven.braun@cs.tu-darmstadt.de