Data-Free Distillation

Ernest Nasyrov, Nikita Okhotnikov, Yuri Sapronov,
Vladimir Solodkin

1 Introduction

Knowledge distillation (KD) is a powerful technique for transferring knowledge from a large, complex "teacher" model to a smaller, more efficient "student" model. Traditionally, this process relies on access to the original training data. However, in many real-world scenarios, such as when dealing with sensitive information or legacy systems, the original data may be unavailable or inaccessible. Data-free distillation addresses this challenge by synthesizing "surrogate" data directly from the teacher model, enabling effective knowledge transfer even in the absence of the original dataset.

Why Data-Free Distillation?

- *Data Privacy:* Medical, financial, or user-generated data often cannot be shared due to legal or ethical constraints.
- *Legacy Systems:* Original training data might be lost or corrupted over time.
- *Resource Efficiency:* Generating synthetic data tailored to the teacher's expertise avoids costly data collection.

The core idea is to reverse-engineer representative samples from the teacher model's internal representations (e.g., feature statistics, spectral patterns, or adversarial examples) and use these to train the student.

2 Theoretical Bachground

Let T be a pre-trained teacher model with parameters θ_T, and S a student model with parameters θ_S. Traditional KD minimizes the Kullback-Leibler

(KL) divergence between the teacher's and student's output distributions over a dataset $\mathcal{D} = \{x_i\}_{i=1}^{N}$:

$$\mathcal{L}_{\text{KD}} = \mathbb{E}_{x \sim \mathcal{D}} \left[\text{KL} \left(T(x) \,\|\, S(x) \right) \right]$$

Here, $T(x)$ and $S(x)$ are probability distributions of the teacher and student, respectively. This approach requires access to $\mathcal{D}$, which becomes problematic when data is unavailable.

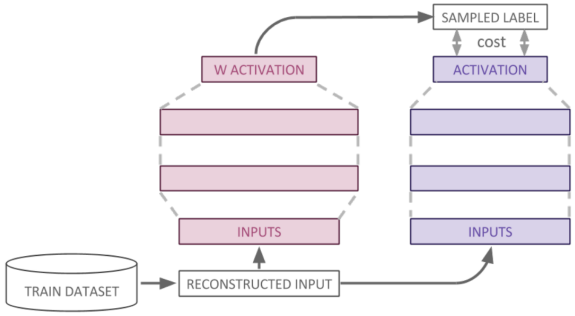


Figure 1: The workflow of classical KD

Transition to Data-Free Knowledge Distillation (DFKD) In our case, when $\mathcal{D}$ is inaccessible, DFKD synthesizes a surrogate dataset $\mathcal{D}' = \{x'_i\}_{i=1}^{M}$ directly from T. Therefore, the objective generalizes to:

$$\mathcal{L}_{\text{DFKD}} = \mathbb{E}_{x' \sim \mathcal{D}'} \left[\text{KL} \left(T(x') \,\|\, S(x') \right) \right]$$

The key challenge here lies in designing high quality surrogate dataset $\mathcal{D}'$ that is able to capture the inherent knowledge of T.

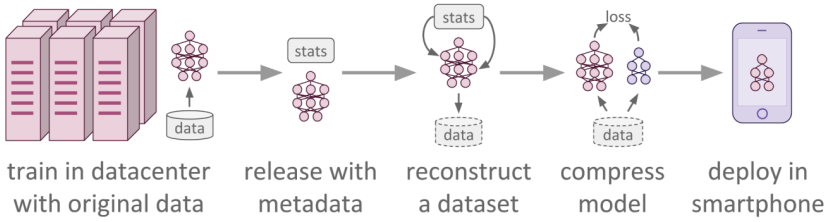


Figure 2: The workflow of DFKD

3 Suggested DKDF Algorithms

In our work, we investigate several DKDF algorithms, which are briefly described below:

1. Statistical Feature Matching: Let $\phi_T^l(x)$ and $\phi_S^l(x)$ denote the activations at layer l of T and S. We synthesize $\mathcal{D}'$ by aligning batch normalization statistics:

$$\min_{x'} \sum_l \left(\|\mu_T^l - \mu_S^l(x')\|_2^2 + \|\sigma_T^l - \sigma_S^l(x')\|_2^2 \right),$$

where μ_T^l, σ_T^l are the teacher's precomputed mean and standard deviation.

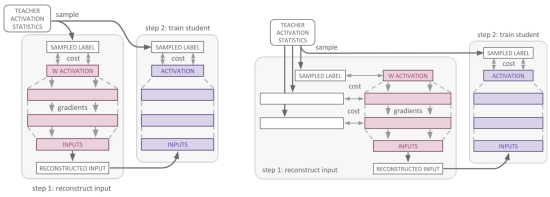
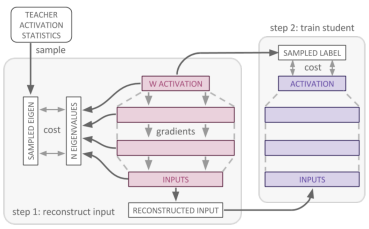


Figure 3: The workflow of Statistical Feature Matching

2. Spectral Feature Matching:

The idea is similar to the first one, but now, we extend to frequency domain by minimizing the Frobenius norm between Fourier transforms of activations:

$$\min_{x'} \sum_l \|\mathcal{F}(\phi_T^l) - \mathcal{F}(\phi_S^l(x'))\|_F^2$$



(d) All-Layers Spectral Activation Record

Figure 4: The workflow of Spectral Feature Matching

3. Adversarial Distillation:

   The third approach introduces a generator G that produces $x' = G(z)$ from noise z. The student S and G engage in a minimax game:

$$\min_{S} \max_{G} \mathbb{E}_z \left[\mathrm{KL} \left(T(G(z)) \,\|\, S(G(z)) \right) \right]$$

   Here G learns to generate samples where T and S disagree, while S adapts to mimic T on these adversarial examples.

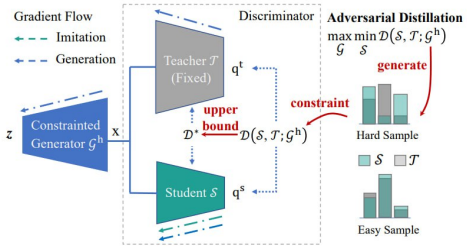


Figure 5: The workflow of Adversarial Distillation

4. Deep Inversion:

   The key idea here is to unify various existing techniques. One synthesizes x' by optimizing the following:

$$\min_{x'} \underbrace{\|\phi_T(x') - \phi_S(x')\|_2^2}_{\text{Feature matching}} + \lambda_1 \underbrace{\|\nabla_{x'} \phi_T(x')\|_2^2}_{\text{Smoothness}} + \lambda_2 \underbrace{\mathrm{TV}(x')}_{\text{Image prior}}$$

   The Total Variation (TV) distance regularization ensures spatial coherence in generated samples.

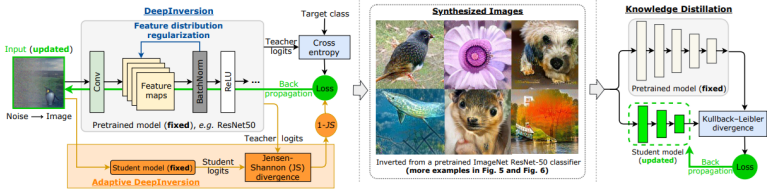


Figure 6: The workflow of Deep Inversion

Although there is a significant discrepancy in these approaches, all DFKD methods share a unified goal: approximate the original data distribution $p_{\mathcal{D}}(x)$

using information implicit in T: while classical KD minimizes $\mathrm{KL}(p_T \| p_S)$ over $p_{\mathcal{D}}$, DFKD minimizes it over a learned surrogate distribution $p_{\mathcal{D}'}$, constrained by:

$$p_{\mathcal{D}'}(x') \approx p_{\mathcal{D}}(x) \quad \text{s.t.} \quad \mathcal{D}' \sim \mathrm{Info}(T),$$

where $\mathrm{Info}(T)$ represents information on the original data extracted from T (statistics, adversarial examples, or inverted features). In other words, we can view DFKD as a generalized expectation-maximization process, where $\mathcal{D}'$ acts as latent variables reconstructed from T's behavior.

4 Our Framework

We now present the unified framework that allows one to use various specific DFKD algorithms to distill the knowledge from any source without having access to the initial data. Our implementation is a library that allows, given a teacher model (and possibly some statistics), to train a distilled student model out-of-the-box. The user can utilize any of the above-described algorithms as well as the classical KD approach if the original data is given.