On Dark Knowledge for Distilling Generators

On Dark Knowledge for Distilling Generators

Chi Hong[1], Robert Birke[3], Pin-Yu Chen[4], and Lydia Y. Chen[1,2] ✉

[1] Delft University of Technology, Delft, Netherlands
[2] University of Neuchatel, Neuchatel, Switzerland
[3] University of Torino, Turin, Italy
[4] IBM Research, New York, USA
c.hong@tudelft.nl, robert.birke@unito.it, pin-yu.chen@ibm.com,
lydiaychen@ieee.org

Abstract. Knowledge distillation has been applied on generative models, such as Variational Autoencoder (VAE) and Generative Adversarial Networks (GANs). To distill the knowledge, the synthetic outputs of a teacher generator are used to train a student model. While the dark knowledge, i.e., the probabilistic output, is well explored in distilling classifiers, little is known about the existence of an equivalent dark knowledge for generative models and its extractability. In this paper, we derive the first kind of empirical risk bound for distilling generative models from a Bayesian perspective. Through our analysis, we show the existence of the dark knowledge for generative models, i.e., Bayes probability distribution of a synthetic output from a given input, which achieves lower empirical risk bound than merely using the synthetic output of the generators. Furthermore, we propose a **D**ark **K**nowledge based **D**istillation , DKtill, which trains the student generator based on the (approximate) dark knowledge. Our extensive evaluation on distilling VAE, conditional GANs, and translation GANs on Facades and CelebA datasets show that the FID of student generators trained by DKtill combining dark knowledge are lower than student generators trained only by the synthetic outputs by up to 42.66%, and 78.99%, respectively.

Keywords: Knowledge distillation · Generators · Risk bounds.

1 Introduction

Generative models, such as Variational autoencoders (VAE) and generative adversarial networks (GANs), are increasingly applied to synthesize images. To practically deploy those models on edge devices, namely the generators, it is critical to distill/compress the models first due to the memory and computation constraints. Knowledge distillation via teacher and student models can effectively compress the machine learning models, especially for classification models [5, 2, 7]. The student model tries to imitate the (non-compressed) teacher model through the input and output pairs from the teacher model such that the same learning efficacy can be achieved via a smaller model. Dark knowledge [22] is shown particularly critical for the distillation quality of classifiers. Recently, some

Authors' preprint copy of Hong, C., Birke, R., Chen, PY., Chen, L.Y. (2024). On Dark Knowledge for Distilling Generators. In: Yang, DN., Xie, X., Tseng, V.S., Pei, J., Huang, JW., Lin, J.CW. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2024. Lecture Notes in Computer Science(), vol 14646. Springer, Singapore. https://doi.org/10.1007/978-981-97-2253-2_19

related studies [1, 20, 3, 20] focus on distilling the generator of `GANs`. This prior art empirically distills the teacher generator by directly leveraging the synthetic outputs. While the studies in [14, 21, 17] provide theoretical analysis for distilling classifiers, e.g., distillation bound, little is on the theoretical understanding of distilling generative models, for instance, the empirical risk bound and the existence of dark knowledge of the generators.

In this paper, we rethink the generator distillation from a theoretical Bayesian perspective. We hypothesize the existence the "dark knowledge", which is embedded in the synthetic output, e.g., image, but not directly observable. To such an end, we model the synthetic outputs of a generator as a conditional Bayes probability distribution, $\mathbb{P}(y|x)$, where y is the synthetic output and x is the input of the generator. We derive two types of empirical risks for distillation: regular empirical risk using only synthetic inputs, and Bayes empirical risk using the proposed conditional probability distribution. We show that the variance of the Bayes empirical risk is lower than the variance of regular empirical risk, demonstrating better generalization capability. We thus refer to the knowledge of $\mathbb{P}(y|x)$ as dark knowledge. Our analysis can be applied on both probabilistic generative models, e.g., `VAE`, and non-probabilistic models, e.g., GANs. To derive such Bayes empirical risk bound for GANs, we approximate the conditional probability distribution and quantify its impact in the distillation bound.

Motivated by the effectiveness of such dark knowledge, we further propose the **D**ark **K**nowledge **D**istillation, `DKtill`, algorithm to train a student generator model through the (approximate) dark knowledge. Specifically, we try to minimizing the difference between teacher and student outputs by controlling the tensor of the second last layer of their networks, which is an approximated conditional Bayes probability. We extensively evaluate `DKtill` on distilling `VAE`, conditional `GANs` and translational `GANs` for Facades, and `CelebA` datasets. Our results show that the student generators from `DKtill` achieve lower FID than the ones trained on only synthetic images by up to 78.99% and a slightly higher FID than the ones from the teacher model by 17.77%.

Our key contributions for this work can be summarized as follows. *i)* Having Theorem 1 and Proposition 1 as the distillation empirical risk analysis to show the effectiveness of dark knowledge in distilling generative models. *ii)* Deriving Proposition 2 as the approximate distillation empirical risk analysis to capture the impact of approximating dark knowledge in distilling the `GANs`. *i)* Proposing `DKtill` which is a dark knowledge based distillation algorithm for training student generative models. *iv)* Achieving higher distillation quality of `DKtill` on three different generative models and 3 different datasets, compared to distillation algorithms which do not use the proposed (approximated) dark knowledge.

2 Preliminary

We consider general generative models for synthesizing images, including non-probabilistic generators and probabilistic generators. In existing generator distillation works [1, 20, 3] the basic idea is as follows. First, they provide the same inputs x into a trained teacher model $\mathcal{T}$ and a student model f where x is the generator input. x has different formats in different generative models. For vanilla `GANs` or `VAE`, x is a noise vector. For conditional GANs, x is a noise and a conditional label. x can also be a picture in image translation `GANs`. Second, the distance between the two model outputs $\mathcal{T}(x)$ and $f(x)$ is minimized. In this way, the student f can mimic the input-output mapping of $\mathcal{T}$ and extract the knowledge from the teacher's synthetic image outputs $\{\mathcal{T}(x)\}$.

In classifier distillation [5], the final predicted labels of a teacher are not sufficient to train a student. The logits or the softmax outputs of the teacher model are needed to train the student so as to capture the underlying "dark knowledge", which is the probability the teacher assigns to "wrong" labels. Some basic settings and terms of generator distillation used in this paper are introduced in the following.

Synthetic images. In generator distillation, we input x into a trained teacher[5] $\mathcal{T}$ and obtain the corresponding synthetic output $y = \mathcal{T}(x)$. After N inferences, a training dataset $D = \{(x_n, y_n)\}_{n=1}^{N} \sim \mathbb{P}^N$ is obtained to train the student model f, where x_n is an input sample and $y_n = \mathcal{T}(x_n)$ is the corresponding synthetic image. To simplify the notation, without loss of generality, y_n is a single channel image and $y_n^{ij} \in \{0, ..., C\}$ is a pixel of the image, where C indicates the number of possible colors[6]. Let $i \in [1, ..., I]$ and $j \in [1, ..., J]$ be the pixel indexes of an image of width I and height J.

Distillation optimization. To distill the knowledge from the teacher model $\mathcal{T}$, a student generator which can be defined as $f : \mathcal{X} \to \mathbb{R}^{I \times J \times C}$ is trained to minimize the following risk:

$$R(f) = \mathop{\mathbb{E}}_{(x,y)\sim\mathbb{P}}[\ell(y, f(x))]. \tag{1}$$

According to the definition of f, the mapping result $f(x)$ is an $I \times J \times C$ tensor, and $f^{ij}(x) \in \mathbb{R}^C$ corresponds to a pixel. The vector $f^{ij}(x)$ represents the weights of taking different colors in the pixel. For each training pair (x, y) collected from $\mathcal{T}$, the loss of $f(x)$ takes the form $\ell(y, f(x)) = \sum_i \sum_j \ell^{ij}(y^{ij}, f^{ij}(x))$. The term $\ell^{ij}(y^{ij}, f^{ij}(x))$ is the loss value of predicting $f^{ij}(x)$ when the true pixel color is y^{ij}. To do distillation, the student needs to mimic the teacher's outputs. Therefore $\ell^{ij}(y^{ij}, f^{ij}(x))$ should be a loss that minimizes the distance between $y = \mathcal{T}(x)$ and $f(x)$, e.g., the softmax cross-entropy loss.

Distillation using only synthetic images. Having different definitions of $\ell(y, f(x))$ in $R(f)$ leads to different distillation methods. A common way in most existing works is to define $\ell(y, f(x))$ directly using the synthetic output images $\{y\}$ from the teacher. Given $y = \mathcal{T}(x)$, the basic idea for training the student f is to maximize the element $f^{ij,y^{ij}}(x)$ in the output probabilities $f^{ij}(x)$ of the student for each pixel i, j.

[5] We interchangeably use teacher or target model/generator.

[6] The analysis of this paper can be straightforwardly extended to three channel images.

3 Theoretical analysis of dark knowledge in distilling the generator

This section introduces the novel concept of dark knowledge in distilling generators and demonstrates that harvesting this dark knowledge can train a student generator f which generalizes better. In generator distillation, the goal is to transfer the knowledge of the teacher $\mathcal{T}$ as much as possible to a student f. We use the risk (generalization error) of the student f shown in Eq. 1 to evaluate the quality of the distilled student model. The risk describes the generalization ability of f on unseen new data sampled from $\mathbb{P}$ other than the observed training dataset $D = \{(x_n, y_n)\}_{n=1}^N \sim \mathbb{P}^N$. A lower risk value means having a better distilled f. In the following, we first introduce the concept of dark knowledge in generative models and then derive the distillation empirical risks.

3.1 Dark Knowledge of Generators

From a Bayesian viewpoint, each synthetic image y from the teacher is generated by a conditional distribution $\mathbb{P}(y|x)$ where the distribution $\mathbb{P}$ is potentially determined by the trained teacher generator $\mathcal{T}$. Obviously, the knowledge of $\mathbb{P}(y|x)$ cannot be fully extracted from the teacher's synthetic image output. A synthetic image y is only a sample drawn from the distribution $\mathbb{P}(y|x)$. Thus, in generator distillation we call the underlying distribution $\mathbb{P}(y|x)$ the "dark knowledge" of the teacher $\mathcal{T}$. The next section demonstrates that using the dark knowledge $\mathbb{P}(y|x)$ to define $\ell(y, f(x))$ so as to have a precise empirical estimation of $R(f)$ can facilitate the training of a student in generator distillation.

We consider two ways of distilling teacher generator: i) by solely the synthetic image outputs of $\mathcal{T}$, and ii) by using the underlying dark knowledge $\mathbb{P}(y|x)$ of $\mathcal{T}$. We analyze the generalization error of the student f under both distillation ways by comparing their (empirical) risks. According to the introduction in Section 2, in generator distillation, $R(f)$ should be minimized by the algorithm where Eq. 1 shows the general definition of $R(f)$. However, calculating the exact value of $R(f)$ is intractable because $\mathbb{P}$ is unknown or it has no explicit expression. Hence empirical risk definitions are required to approximate $R(f)$. The aforementioned two distillation ways leverage two different corresponding empirical risk definitions. These will be illustrated in detail in the following.

3.2 Distillation Empirical Risk

Distillation with sole synthetic images. First, we introduce the distillation empirical risk of using $\mathcal{T}$'s synthetic image outputs. In this way of distillation, the dataset $D = \{(x_n, y_n)\}_{n=1}^N$ collected from the teacher $\mathcal{T}$ is used to train the student f. We have the following distillation empirical risk definition to approximate $R(f)$:

$$\hat{R}(f; D) = \frac{1}{N} \sum_{n \in [N]} \sum_{i} \sum_{j} \ell^{ij}(y_n^{ij}, f^{ij}(x_n)) = \frac{1}{N} \sum_{n \in [N]} \sum_{i} \sum_{j} e_{y_n^{ij}}^{\top} \ell^{ij}(f^{ij}(x_n)), \tag{2}$$

where $e_{y_n^{ij}}$ is the one-hot vector encoding of the color value of y_n^{ij}, and $\ell^{ij}(f^{ij}(x)) = [\ell^{ij}(1, f^{ij}(x)), ..., \ell^{ij}(C, f^{ij}(x))] \in \mathbb{R}^C$ is a vector of loss values for each possible color of the pixel with index i, j. As shown in Eq. 2, this empirical risk definition only needs the synthetic image y_n from $\mathcal{T}$ to decide the value of $e_{y_n^{ij}}$. Then we can calculate $\hat{R}(f; D)$ with no additional information from $\mathcal{T}$.

Distillation with dark knowledge. Here we assume that besides the final synthetic image output y we also get the conditional probability $\mathbb{P}(y|x)$ (the dark knowledge) of a given input x from $\mathcal{T}$. Consequently, we can define another distillation empirical risk $\hat{R}_\alpha(f, D)$ to approximate $R(f)$:

$$\hat{R}_\alpha(f, D) = \frac{1}{N} \sum_{n \in [N]} \sum_i \sum_j p^{ij}\left(x_n\right)^\top \ell^{ij}((f^{ij}\left(x_n\right)) \tag{3}$$

where the conditional probability of pixel y^{ij} color is denoted by

$$p^{ij}(x) = [\mathbb{P}(y^{ij}|x)]_{y_{ij} \in [C]} \tag{4}$$

Connection to the empirical risk. $\hat{R}(f; D)$ (Eq. 2) can be seen as an approximation of $\hat{R}_\alpha(f, D)$ (Eq. 3). Considering that: $R(f) = \mathop{\mathbb{E}}_{(x,y)\sim\mathbb{P}}[\ell(y, f(x))] = \mathop{\mathbb{E}}_{x}[\mathop{\mathbb{E}}_{y|x}[\ell(y, f(x))]]$, we know that $\hat{R}_\alpha(f, D)$ is an empirical estimate of $\mathop{\mathbb{E}}_{x}[\mathop{\mathbb{E}}_{y|x}[\ell(y, f(x))]]$ over the random variable x. If we further use the one-hot encoding $e^\top_{y_n^{ij}}$ to approximate $p^{ij}(x_n)$ in Eq. 3, we have Eq. 2. To distinguish these two empirical risks, we call $\hat{R}(f; D)$ (Eq. 2) as the *distillation empirical risk with synthetic images* and $\hat{R}_\alpha(f, D)$ (Eq. 3) as the *distillation empirical risk with dark knowledge.*

3.3 Generalization of the student generator

Given a teacher $\mathcal{T}$, a student model f can be trained by optimizing the distillation empirical risk with synthetic images (Eq. 2) or the distillation empirical risk with dark knowledge (Eq. 3). In this section, we analyze the generalization ability of the student f under the two different distillation empirical risk bounds. In the following, we show that a student generator trained with Eq. 3 is expected to generalise better than trained under Eq. 2, which means using the dark knowledge can facilitate generator distillation. Although both Eq. 2 and Eq. 3 are unbiased estimates of $R(f)$, the variances of Eq. 2 and Eq. 3 over the observed training dataset D are different. This is formally demonstrated by Theorem 1 (proof in the supplementary).

Theorem 1. *Let D be a training dataset sampled from $\mathbb{P}^N$. $\mathbb{V}$ represents the variance of a random variable. For any fixed hypothesis $f : \mathcal{X} \to \mathbb{R}^{I \times J \times C}$, $\mathbb{V}_{D\sim\mathbb{P}^N}\left[\hat{R}_\alpha(f; D)\right] \leq \mathbb{V}_{D\sim\mathbb{P}^N}[\hat{R}(f; D)]$*

The two variances in Theorem 1 equal to each other when $p^{ij}(x)$ is concentrated on one single color and the probability on all other colors is zero. However, this case rarely happens. The benefit of having lower variance is that the student f generalises better as shown in the following. Applying Theorem 6 of [15], we have the following bound for the distillation empirical risk with dark knowledge.

Proposition 1. *Let $D \sim \mathbb{P}^N$ and $\mathcal{F}$ be a class of hypotheses $f : \mathcal{X} \to \mathbb{R}^{I\times J\times C}$ with induced class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ of $h(x) = \sum_i \sum_j p^{ij}(x)^\top \ell^{ij}((f^{ij}(x))$. Suppose $\mathcal{H}$ has uniform covering number $\mathcal{N}_\infty$. For any $\delta \in (0,1)$ and set $\mathcal{M}_N = \mathcal{N}_\infty(1/N, \mathcal{H}, 2N)$. Then with probability at least $1-\delta$ over D we have:*

$$R(f) \leq \hat{R}_\alpha(f;D) + \mathcal{O}\Big(\sqrt{\mathbb{V}_N(f)/N}\cdot\sqrt{\log\left(\mathcal{M}_N/\delta\right)}+ \log\left(\mathcal{M}_N/\delta\right)/N)$$

where $\mathbb{V}_N(f)$ is the empirical variance of $\{h(x_n)\}_{n=1}^N$.

Note that using the same procedure we can derive a similar bound for the distillation empirical risk with synthetic images (Eq. 2). Considering Theorem 1 and the two bounds, we know that the bound for Eq. 3 has lower variance penalty. Increasing the training dataset size N (number of queries on $\mathcal{T}$), Eq. 3 has a risk bound with a better rate of convergence. Thus, a student model trained by minimizing the distillation empirical risk with dark knowledge generalises better compared to one using the distillation empirical risk with synthetic images. We conclude that a student generator f trained by the dark knowledge $\mathbb{P}(y|x)$ is better than using sole synthetic image outputs $\{y\}$ from $\mathcal{T}$.

3.4 Impact of probability approximation

In the aforementioned analysis, we showed that having the dark knowledge $\mathbb{P}(y|x)$ from the teacher $\mathcal{T}$ allows for a more precise empirical approximation of $R(f)$. Thus it benefits the training of the student f in generator distillation. However, as mentioned before, doing distillation with dark knowledge requires the conditional probability $\mathbb{P}(y|x)$ for any given input x. In ideal cases, the intermediate layer output of $\mathcal{T}$, e.g., the second last layer output in `VAE`, can provide $\mathbb{P}(y|x)$. Then, it can be used to train the student model. Unfortunately, for some teacher models $\mathcal{T}$, e.g., `GANs`, the intermediate layer output cannot directly show $\mathbb{P}(y|x)$. We refer to the generators that can provide $\mathbb{P}(y|x)$ in the intermediate layer output as *probabilistic generators* and the generators that cannot show the probability as *non-probabilistic generators*. To distill the dark knowledge from non-probabilistic generators, it is required to approximate $\mathbb{P}(y|x)$. Let $\tilde{\mathbb{P}}(y|x)$ be an approximated probability of $\mathbb{P}(y|x)$. In this section, we study the impact on the student generalization error using such an approximated distribution to do distillation. Using the approximated probability to train the student f, referring to Eq. 3 we have the following distillation empirical risk:

$$\tilde{R}_\alpha(f,D) = \frac{1}{N}\sum_{n\in[N]}\sum_i\sum_j \tilde{p}^{ij}(x_n)^\top \ell^{ij}((f^{ij}(x_n)), \tag{5}$$

where $\tilde{p}^{ij}(x_n)$ is the approximation of $p^{ij}(x_n)$. According to Eq. 1 and Eq. 4 we can rewrite the population risk $R(f)$ as:

$$R(f) = \mathbb{E}_x[\mathbb{E}_{y|x}[\ell(y,f(x))]] = \mathbb{E}_x[\sum_i\sum_j p^{ij}(x)^\top \ell^{ij}(f^{ij}(x))]. \tag{6}$$

The following Proposition 2 (proof in the supplementary) reveals the connection between the approximation of $\mathbb{P}(y|x)$ and the generalization error of f.

Proposition 2. *If the loss ℓ is bounded, we train the student model by minimizing the empirical risk shown in Eq. 5. For any hypothesis $f : \mathcal{X} \to \mathbb{R}^{I \times J \times C}$,*

$$\begin{aligned} \mathbb{E}\left[(\tilde{R}_\alpha(f;D) - R(f))^2\right] \leq \frac{1}{N} \cdot \mathbb{V}[\sum_i \sum_j \tilde{p}^{ij}(x)^\top \ell^{ij}((f^{ij}(x))] + \\ \mathcal{O}(\mathbb{E}[\sum_i \sum_j \left\|\tilde{p}^{ij}(x) - p^{ij}(x)\right\|_2])^2. \end{aligned} \tag{7}$$

On the right side of Eq. (7), when N is big, the second term dominates the upper bound of the gap $\tilde{R}(f;D) - R(f)$. That means minimizing the distance between the approximated probability $\tilde{\mathbb{P}}(y|x)$ and the ground truth probability $\mathbb{P}(y|x)$ yields a tighter upper bound of the risk gap $\tilde{R}(f;D) - R(f)$. Hence, a tighter approximation leads to a better student model.

4 DKtill: Extracting Dark Knowledge for Training Student Generator

In the previous section, we theoretically demonstrate that using the underlying dark knowledge of a teacher $\mathcal{T}$ can improve the generator distillation and train a student f that generalizes better. However, to make use of the underlying dark knowledge in distillation, we first need to know how to extract and use the dark knowledge from a teacher generator $\mathcal{T}$. In this section, we propose two methods to extract the dark knowledge based on the class of the generator: one for probabilistic generators and one for non-probabilistic generators. For verifying the effectiveness of the extracted dark knowledge, we propose a generator distillation algorithm DKtill. When distilling $\mathcal{T}$, DKtill makes the student f to mimic the extracted (approximated) $\mathbb{P}(y|x)$ besides the synthetic images $y = \mathcal{T}(x)$ (see the supplementary for more implementation details).

4.1 Extracting from probabilistic generators

Probabilistic generators, e.g., variational auto-encoder (VAE) [10], commonly assume the existence of latent variables. More in detail, they assume that a synthetic image y is generated by some random process involving some latent random variables. In such generators, to do distillation with dark knowledge, the conditional probability $\mathbb{P}(y|x)$ can be calculated by some middle layer outputs of the teacher generator. In the following, we take VAE as an example for dark knowledge extraction in probabilistic generators. The training method of VAE is a kind of variational inference that uses $q(x|y)$ to approximate the intractable true posterior $p(x|y)$ and maximize the evidence lower bound. For VAE distillation, we focus on its trained decoder. In the decoder, the synthetic image is produced by the conditional distribution $\mathbb{P}(y|x) = \mathcal{N}(y; \mu, \sigma^2 I)$, where the distribution parameters μ and σ are the middle layer outputs of the decoder neural network. Thus, given any input x, we can obtain $\mathbb{P}(y|x)$ by getting the output value of μ and σ from the middle layers.

4.2 Extracting from non-probabilistic generators

Non-probabilistic generators, e.g., GANs, do not assume any probability relationship between output image y and latent random variables. y is directly produced by a neural network mapping $\mathcal{T}(x)$, where $\mathcal{T}$ is the teacher generator. Different from probabilistic generators, in non-probabilistic generators we cannot derive the knowledge of $\mathbb{P}(y|x)$ by some middle layer outputs of $\mathcal{T}$. In the following, we take GANs as an example to show extracting dark knowledge $\mathbb{P}(y|x)$ from non-probabilistic generators. Given a GANs teacher generator $\mathcal{T}$, we let g_α be the second last layer and g_β be the last layer of $\mathcal{T}$. Given an input x into the teacher generator, the intermediate output of g_α is represented by α. Then the corresponding synthetic image can be represented as $y = g_\beta(\alpha)$. To get the dark knowledge, we can apply a differentiable probabilistic network (e.g., MLP) g_γ to replace the original last layer g_β. The input of g_γ is g_α's output, α and the output of g_γ is a tensor that takes the shape corresponding to the underlying Bayes probabilities. The shape of $g_\gamma(\alpha)$ is decided by I, J and C where I and J are decided by the synthetic image size and C depends on the color range. The final synthetic image is now sampled from the tensor $g_\gamma(\alpha)$. If $\mathcal{T}$ is untrained, we can train it from scratch using this new architecture. Thus, after the training, we can easily get $\mathbb{P}(y|x)$ from $g_\gamma(\alpha)$. If the given $\mathcal{T}$ is pre-trained, we can just fine-tune the parameters of g_γ using the optimization objective, $\mathrm{argmin}_{g_\gamma} ||g_\beta(\alpha) - \hat{y}||_1$, where $\hat{y} \sim Discrete(g_\gamma(\alpha))$ is the synthetic image sampled from $g_\gamma(\alpha)$.

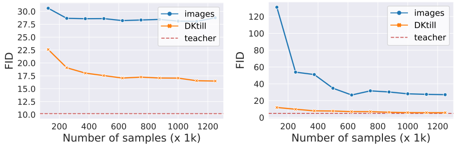
5 Empirical illustration

5.1 Setting

Datasets: we consider the datasets, Facades [6] and CelebA [13]. In the case of conditional generation, we focus on the gender class, i.e., male and female, in the CelebA dataset. We downsampled the CelebA images from originally 178 × 218 to 64 × 64 pixels. The input images size of Facades is 256 × 256.
Networks: the network structures of all generators are based on convolutional neural networks [10, 18]. For VAE, we compress a 27.2MB teacher into a 2.4MB student. As for conditional GANs, a 2.8MB student is distilled from a 14.4MB teacher. In the image translation experiment, Pix2pix is compressed from 217.8MB into 14.0MB.
Distillation process and baseline: for both DKtill and the baseline (abbreviated as "image" [4]), we train the student generators, by minimizing the distance of the generated images to the teacher from the same random inputs (details in Appendix). However, instead of only using generated images (baseline), DKtill also uses in parallel the information of dark knowledge (underlying distribution) for loss minimization. Note that although the baseline distillation is originally designed for Image Translation, we adopt it for VAE and conditional GANs.

Evaluation metrics: we use Fréchet Inception Distance (FID) to evaluate the quality of the images produced by the distilled student model. Lower values of FID indicate higher quality of generated data.

5.2 Distilling probabilistic generators



(a) VAE **(b)** conditional GANs

Fig. 1: FID of distilling VAE and conditional GANs on CelebA.

Here, we distill VAE (the teacher generator) using the baseline and DKtill. Fig. 1 shows the comparison results based on FID for VAE and conditional GANs (Fig. 1a). We also illustrate the FID of images generated by the teacher VAE, i.e., the horizontal line in Fig. 1a. The FID value decreases during the training, from 30.61 to 28.76, and 22.61 to 16.49, for baseline and DKtill respectively, demonstrating lower distance to ground truth images. When looking at the final FID (after 128K examples), DKtill is 42.66% lower than the baseline image and 62.14% higher than the teacher generator. Here FID for the teacher VAE is 10.17, which is much better than both students. This is within our expectation as the teacher model is 10X the size of the students. In general, we can see that DKtill, using the dark knowledge provided by the intermediate output of teacher VAE, can significantly improve the quality of distillation. As the number of examples increases, the student learns the teacher's mapping and distribution better. Using dark knowledge can achieve lower FID and thus it can distill better.

5.3 Distilling non-probabilistic generators

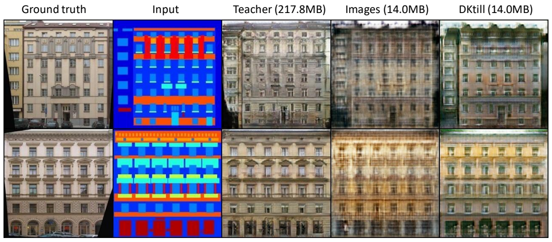


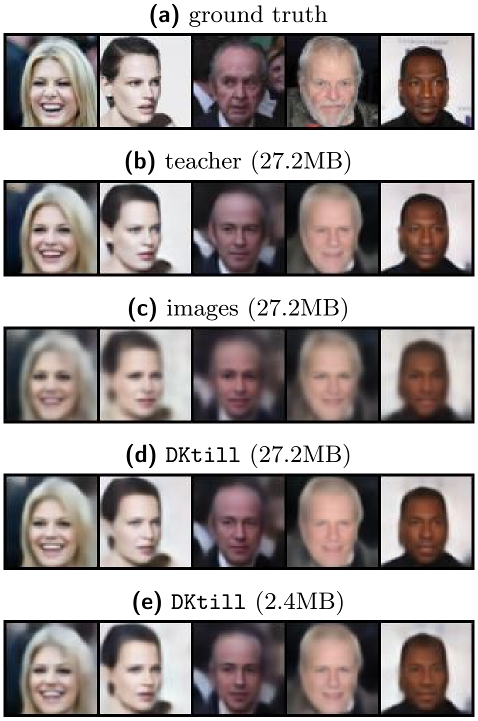
Fig. 2: Pix2pix image translation GANs on Facades.

Here we implement two tasks for non-probabilistic generators: the conditional GANs and image translation as the baseline [4]. Fig. 1b evaluates the distillation on conditional GANs.We can see that FID shares the same trend as distilling the probabilistic generator VAE. Thus, our proposed method DKtill is able to effectively extract knowledge from non-probabilistic teacher generators too, given the approximate nature of dark knowledge. The reason why FID of DKtill gets much closer to the teacher than Fig. 1a is that the teacher model is only 5X bigger in parameter size than the students. When looking at the final FID (after 128K examples), DKtill is 78.99% lower than the baseline image and only 17.77% higher than the teacher generator.

Let us zoom into the distillation results of the image translation task in Fig. 2. The goal is to train the translation GANs so that given input (as the "input" in visualization), the Pix2pix network approximately maps it into the ground truth image. From the results, we observe that even with 6% size of the teacher model (FID: 35), DKtill is able to distill the image translation generator at high quality (FID: 35) with dark knowledge, especially compared with the baseline (FID:

37.88). By distillation, the floating point operations per second (FLOPs) on one input image is reduced from $6.06G$ (Teacher) to $0.83G$ (Student).

5.4 Small generators through `DKtill`



(a) ground truth

(b) teacher (27.2MB)

(c) images (27.2MB)

(d) `DKtill` (27.2MB)

(e) `DKtill` (2.4MB)

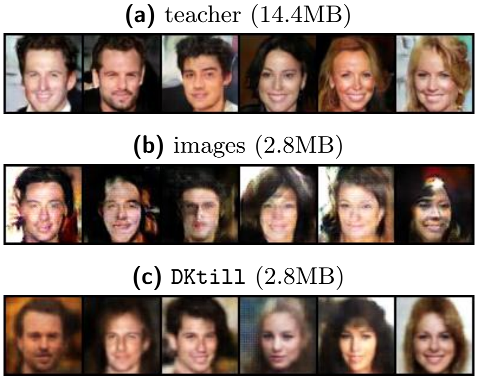
Fig. 3: `VAE` on `CelebA`.

In addition to achieving FID similar to the teacher model, we show another advantage of exploring dark knowledge, i.e., smaller student generator. We first solicit the synthetic images generated by the following model in Fig. 3: the ground truth, teacher `VAE`, 27.2MB student VAE trained by baseline, 27.2MB student `VAE` from `DKtill`, and 2.4MB student `VAE` from `DKtill`. Without any surprise, `DKtill` achieves better distillation quality when using bigger networks, comparing Fig. 3d and Fig. 3e. Another observation is that, smaller student (2.4MB) `VAE` trained by `DKtill`, achieve better image quality than baseline with 27.2MB, comparing Fig. 3c and Fig. 3e.

The conditional `GANs` result is also presented in Fig. 4 (on gender class, i.e., male and female for the left 3 and right 3 pictures). Note that here we do not show the corresponding real images of the synthetic ones since conditional `GANs` do not have a latent code encoder as `VAE`. Thus, having the latent code of a real image for reproducing some corresponding synthetic ones is difficult. Given the same small network with 2.8MB for `DKtill` and the baseline, `DKtill` shows better generated image quality than baseline. These results again prove the existence of dark knowledge for generators and its benefit for distilling generators.

6 Related work

Knowledge distillation [5, 2, 7, 12, 9], which transfers the knowledge from a big network to a small one, enables the light-weight deep learning. Generally, knowledge distillation is studied on classifiers for model compression, e.g, [5] compresses the knowledge of an ensemble into a single model which is much easier to deploy. Such techniques are used to train surrogate models [19, 11, 23], even without the necessity of knowing the target model parameters.



(a) teacher (14.4MB)

(b) images (2.8MB)

(c) DKtill (2.8MB)

Fig. 4: Conditional GANs on CelebA.

Besides, knowledge distillation has also been applied on GANs [1, 20, 3, 20]. KDGAN [20] simultaneously optimizes the distillation and adversarial losses between a classifier, a teacher, and a discriminator, to learn the true data distribution at the equilibrium. To further improve the performance, [3] include a student discriminator to measure the distances between real data, and the synthetic data generated by student and teacher generators. Recently, there are theoretical analyses on knowledge distillation for classifiers [14, 21, 17, 16, 22, 8]. On the one hand, [16] first provides the theoretical analysis of self-distillation, fitting a nonlinear function on Hilbert space and L2 regularization. This analysis sheds light on the relation between self-distillation rounds and (under-)over- fitting. On the other hand, based on neural tangent kernel, [8] provides a transfer risk bound for the linearized model of the wide neural networks, revealing the impact of soft (hard) labels for (im)perfect teachers according to the designed data inefficiency metric. Furthermore, [22] explores the bias-variance trade-off brought by distillation with soft labels. According to their analysis, novel weighted soft labels are inspired to help the network adaptively handle the trade-off. However, none of the existing work provides a generalization analysis on generators.

7 Conclusion

In this paper, we model the knowledge distillation for generative models from a Bayesian perspective, identifying dark knowledge and its influence on the generalization ability of student models, i.e., lower empirical risk. Furthermore, we propose a dark knowledge based distillation optimization, DKtill, to train student generators on both non-probability-based and probability-based generative models. Evaluation results on three datasets across different scenarios show that synthetic images from DKtill achieve lower FID by up to 78.99%, in contrast to using images only. DKtill also generates images of similar quality as the teacher model, using smaller and more compact generator networks.

Acknowledgements This work has been supported by the Spoke "FutureHPC & BigData" of the ICSC - Centro Nazionale di Ricerca in "High Performance Computing, Big Data and Quantum Computing", funded by EU - NextGenerationEU and the EPI project funded by EuroHPC JU under G.A. 101036168.

References

1. Aguinaldo, A., Chiang, P., Gain, A., Patil, A., Pearson, K., Feizi, S.: Compressing gans using knowledge distillation. CoRR **abs/1902.00159** (2019)
2. Chandrasekaran, V., Chaudhuri, K., Giacomelli, I., Jha, S., Yan, S.: Exploring connections between active learning and model extraction. In: USENIX Security
3. Chen, H., Wang, Y., Shu, H., Wen, C., Xu, C., Shi, B., Xu, C., Xu, C.: Distilling portable generative adversarial networks for image translation. In: AAAI (2020)
4. Chen, H., Wang, Y., Shu, H., Wen, C., Xu, C., Shi, B., Xu, C., Xu, C.: Distilling portable generative adversarial networks for image translation. In: AAAI (2020)
5. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. CoRR **abs/1503.02531** (2015)
6. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. pp. 1125–1134 (2017)
7. Jagielski, M., Carlini, N., Berthelot, D., Kurakin, A., Papernot, N.: High accuracy and high fidelity extraction of neural networks. In: USENIX Security (2020)
8. Ji, G., Zhu, Z.: Knowledge distillation in wide neural networks: Risk bound, data efficiency and imperfect teacher. In: NeurIPS 2020 (2020)
9. Kanwal, N., Eftestøl, T., Khoraminia, F., Zuiverloon, T.C., Engan, K.: Vision transformers for small histological datasets learned through knowledge distillation. In: PAKDD (2023)
10. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Bengio, Y., LeCun, Y. (eds.) ICLR (2014)
11. Krishna, K., Tomar, G.S., Parikh, A.P., Papernot, N., Iyyer, M.: Thieves on sesame street! model extraction of bert-based apis. In: ICLR (2020)
12. Liu, Z., Zhu, Y., Gao, Z., Sheng, X., Xu, L.: Itrievalkd: An iterative retrieval framework assisted with knowledge distillation for noisy text-to-image retrieval. In: PAKDD (2023)
13. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: (ICCV) (2015)
14. Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V.: Unifying distillation and privileged information. In: Bengio, Y., LeCun, Y. (eds.) ICLR (2016)
15. Maurer, A., Pontil, M.: Empirical bernstein bounds and sample variance penalization. COLT 2009 - The 22nd Conference on Learning Theory (2009)
16. Mobahi, H., Farajtabar, M., Bartlett, P.L.: Self-distillation amplifies regularization in hilbert space. In: NeurIPS (2020)
17. Phuong, M., Lampert, C.H.: Towards understanding knowledge distillation. CoRR **abs/2105.13093** (2021)
18. Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., Carin, L.: Variational autoencoder for deep learning of images, labels and captions. NIPS **29** (2016)
19. Truong, J., Maini, P., Walls, R.J., Papernot, N.: Data-free model extraction. In: CVPR (2021)
20. Wang, X., Zhang, R., Sun, Y., Qi, J.: KDGAN: knowledge distillation with generative adversarial networks. In: NeurIPS. pp. 783–794 (2018)
21. Zhang, Z., Sabuncu, M.R.: Self-distillation as instance-specific label smoothing. In: NeurIPS (2020)
22. Zhou, H., Song, L., Chen, J., Zhou, Y., Wang, G., Yuan, J., Zhang, Q.: Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. In: 9th International Conference on Learning Representations, ICLR 2021 (2021)
23. Zhou, M., Wu, J., Liu, Y., Liu, S., Zhu, C.: Dast: Data-free substitute training for adversarial attacks. In: CVPR. pp. 231–240. IEEE (2020)