Aligning Logits Generatively for Principled Black-Box Knowledge Distillation

Supplementary Material

In Appendix, we provide proof of theorems and more experimental results for MEKD. We also visualize the real and generated distributions of MEKD with DCGAN to verify the effectiveness of our method.

1. Proofs

The success of deep learning can be attributed to the discovery of intrinsic structures of data, which is defined as the manifold distribution hypothesis [13]. The data is concentrated on a manifold $\Sigma \in \mathbb{R}^n$, which is embedded in the image space $\mathcal{X}$, and data distribution can be abstracted as a probability distribution μ over the data manifold. The encoding-map $\varphi : \Sigma \to \Omega$ maps the data manifold Σ to the label manifold $\Omega \in \mathbb{R}^C$ in a label space $\mathcal{Y}$ which is also called latent space, while mapping the data distribution μ to latent distribution $\upsilon = \varphi_{\#}\mu$. Each sample x is mapped from the image space into the latent space, and its result $\varphi(x)$ is called a latent code. The decoding-map φ^{-1} remaps latent codes to the data manifold. Both φ and φ^{-1} are strongly nonlinear functions, which can be simulated with different neural networks [7, 8]. Meanwhile, the well-known Kolmogorov Theorem [2, 6] indicates that any multivariate continuous function can be represented as the sum of continuous real-valued functions with continuous one-dimensional outer and inner functions Φ_q and $\Psi_{q,p}$.

The teacher function $f_T \in \varphi$ can be considered as a kind of encoding map, and the generator function $f_G \in \varphi^{-1}$ can be considered as a kind of decoding map. Let $\mathcal{X} \in \mathbb{R}^n$ be the image space, where data x is sampled from. For a C-way classification task, let $\mathcal{Y} \in \mathbb{R}^C$ be the latent space, where $|\mathcal{Y}| = C$. Defining the model as a complex mapping function from the image distribution to the latent distribution, we can consider the teacher model as $f_T : \mathcal{X} \to \mathcal{Y}$ parameterized by $\theta_T \in \Theta_T$, whose outputs indicate the probabilities (*e.g.*, logits) of what category the samples belong to. The same for the student model $f_S : \mathcal{X} \to \mathcal{Y}$ parameterized by $\theta_S \in \Theta_S$.

Definition 1. (Function Equivalence) *Giving the student and teacher model f_S and f_T, for a data distribution $\mu \in \mathcal{X}$ in image space which is mapped to $\mathbb{P}_S \in \mathcal{Y}$ and $\mathbb{P}_T \in \mathcal{Y}$ in latent space. If the Wasserstein distance between $\mathbb{P}_S$ and $\mathbb{P}_T$ equals zero,*

$$W(\mathbb{P}_S, \mathbb{P}_T) = \inf_{\gamma \in \Pi(\mathbb{P}_S, \mathbb{P}_T)} \mathbb{E}_{(y_S, y_T) \sim \gamma} [\, \| y_S - y_T \| \,] = 0, \tag{1}$$

the student and teacher model are equivalent, i.e., *$f_S = f_T$, where $\Pi(\mathbb{P}_S, \mathbb{P}_T)$ is the set of all joint distributions $\gamma(y_S, y_T)$ whose marginals are $\mathbb{P}_S$ and $\mathbb{P}_T$, respectively.*

Definition 2. (Inverse Mapping) *Giving a prior distribution $p \in \mathbb{R}^C$, for a data distribution $\mu \in \mathbb{R}^n$, if the Wasserstein distance between generated distribution $\mu' = (f_G)_{\#}p$ and μ equals zero,*

$$W(\mu', \mu) = \inf_{\gamma \in \Pi(\mu', \mu)} \mathbb{E}_{(x', x) \sim \gamma} [\, \| x' - x \| \,] = 0, \tag{2}$$

then the generator $f_G : \mathbb{R}^C \to \mathbb{R}^n$ is the inverse mapping of the teacher function $f_T : \mathbb{R}^n \to \mathbb{R}^C$, denoted as $f_G = f_T^{-1}$, where $\Pi(\mu', \mu)$ is the set of all joint distributions $\gamma(x', x)$ whose marginals are respectively μ' and μ.

1.1. Proof of Theorem 1

Theorem 1. (Empirical Approximation) *For any $0 < \epsilon < 1/2$ and any integer $m > 4$, let $g : \mathbb{R}^C \to \mathbb{R}^n$ be the mapping function of generator G with $n \leq \frac{20 \log m}{\epsilon^2}$. For two sets $V_S = \{y_S : y_S \in \mathbb{P}_S\}$ and $V_T = \{y_T : y_T \in \mathbb{P}_T\}$, both of which have m points in $\mathbb{R}^C$, if the empirical Wasserstein distance between $g(V_S)$ and $g(V_T)$ equals zero,*

$$\hat{W}(g(V_S), g(V_T)) = \frac{1}{m} \sum_{i=1}^{m} \| g(y_S^i) - g(y_T^i) \| = 0, \tag{3}$$

then $W(\mathbb{P}_S, \mathbb{P}_T) = 0$.

Proof. According to Johnson-Lindenstrauss theorem, for $y_S \in V_S$ and $y_T \in V_T$, we have

$$\| y_S - y_T \| \leq (1 + \epsilon) \| g(y_S) - g(y_T) \|. \tag{4}$$

For set V_S and V_T, we can get the empirical Wasserstein distance between them:

$$\begin{aligned} \hat{W}(V_S, V_T) &= \frac{1}{m} \sum_{i=1}^{m} \| y_S^i - y_T^i \| \\ &\leq \frac{1}{m} \sum_{i=1}^{m} (1 + \epsilon) \| g(y_S^i) - g(y_T^i) \| \\ &= \frac{1 + \epsilon}{m} \sum_{i=1}^{m} \| g(y_S^i) - g(y_T^i) \| \\ &= (1 + \epsilon) \hat{W}(g(V_S), g(V_T)) = 0. \end{aligned} \tag{5}$$

Because the Wasserstein distance between $\mathbb{P}_S$ and $\mathbb{P}_T$ is the expectation of the empirical Wasserstein distance between V_S and V_T, i.e.,

$$W(\mathbb{P}_S, \mathbb{P}_T) = \mathbb{E}_{(V_S, V_T) \sim \Pi(\mathbb{P}_S, \mathbb{P}_T)} \left[\hat{W}(V_S, V_T) \right], \tag{6}$$

so we can get

$$W(\mathbb{P}_S, \mathbb{P}_T) \leq \hat{W}(V_S, V_T) = 0. \tag{7}$$

Since

$$W(\mathbb{P}_S, \mathbb{P}_T) = \inf_{\gamma \in \Pi(\mathbb{P}_S, \mathbb{P}_T)} \mathbb{E}_{(y_S, y_T) \sim \gamma} [\, \| y_S - y_T \| \,] \geq 0, \tag{8}$$

then the result $W(\mathbb{P}_S, \mathbb{P}_T) = 0$ is derived. □

1.2. Proof of Theorem 2

Theorem 2. (Optimization Direction) *Let $\mu \in \mathcal{X}$ be any distribution. f_S, f_T, f_G are the mapping functions of the student, teacher, and generator, respectively. f_S is parameterized by $\theta_S \in \Theta_S$. Then, when*

$$\min_{\theta_S \in \Theta_S} \mathbb{E}_{x \sim \mu} [\| f_G \circ f_S(x), f_G \circ f_T(x) \|] \to 0, \tag{9}$$

it holds that $f_S \to f_T$, and we have

$$\begin{aligned} \nabla_{\theta_S} \mathbb{E}_{x \sim \mu}[f_S(x)] &= \nabla_{\theta_S} W(\mathbb{P}_S, \mathbb{P}_T) \\ &= \mathbb{E}_{x \sim \mu}[\nabla_{\theta_S} \| f_G \circ f_S(x) - f_G \circ f_T(x) \|]. \end{aligned} \tag{10}$$

Proof. Let us define

$$V(f_S, \theta_S) = \mathbb{E}_{x \sim \mu} [\, \| f_S(x), f_T(x) \| \,], \tag{11}$$

$$V'(f_S, \theta_S) = \mathbb{E}_{x \sim \mu}[\, \| f_G \circ f_S(x), f_G \circ f_T(x) \| \,], \tag{12}$$

where f_S lies in $\mathcal{F}_S = \{f_S : \mathcal{X} \to \mathcal{Y}\}$ and $\theta_S \in \Theta_S$.

According to the Johnson-Lindenstrauss Lemma [4], for any $0 < \epsilon < 1/2$ and any integer $m > 4$, let $n = \frac{20 \log m}{\epsilon^2}$, then for any set S of m points in $\mathbb{R}^C$, the generator mapping function $f_G : \mathbb{R}^C \to \mathbb{R}^n$ for all $f_S(x), f_T(x) \in S$ holds that

$$\begin{aligned} &(1 - \epsilon) \, \| f_G \circ f_S(x), f_G \circ f_T(x) \| \\ &\quad \leq \| f_S(x), f_T(x) \| \\ &\quad\quad \leq (1 + \epsilon) \, \| f_G \circ f_S(x), f_G \circ f_T(x) \|. \end{aligned} \tag{13}$$

Using Squeeze Theorem [9], we know that the minimization of equation 11 and equation 12 converge to the same results, *i.e.*,

$$\inf V(f_S, \theta_S) = \inf V'(f_S, \theta_S). \tag{14}$$

We can rewrite the equation 1 using $x \sim \mu$:

$$\begin{aligned} W(\mathbb{P}_S, \mathbb{P}_T) &= \inf_{\gamma \in \prod(\mathbb{P}_S, \mathbb{P}_T)} \mathbb{E}_{(y_S, y_T) \sim \gamma} [\, \| y_S - y_T \| \,] \\ &= \inf_{\gamma \in \prod(f_S(\mu), f_T(\mu))} \mathbb{E}_{x \sim \mu} [\, \| f_S(x), f_T(x) \| \,] \\ &= \inf_{\gamma \in \prod(f_S(\mu), f_T(\mu))} V(f_S, \theta_S), \end{aligned} \tag{15}$$

where f_S and f_T map distribution μ to $\mathbb{P}_S$ and $\mathbb{P}_T$, respectively. So we can get

$$\inf \, V'(f_S, \theta_S) = \inf \, V(f_S, \theta_S) = W(\mathbb{P}_S, \mathbb{P}_T). \tag{16}$$

According to Def. 1, when $\inf V'(f_S, \theta_S) \to 0$, then $W(\mathbb{P}_S, \mathbb{P}_T) \to 0$, and we can derive that $f_S \to f_T$.

The rest of the proof will be dedicated to show that the optimal solution of $\min V'(f_S, \theta_S)$ leads to reduce the Wasserstein distance of $\mathbb{P}_S$ and $\mathbb{P}_T$, which drives f_S to approximate f_T.

We know by the Kantorovich-Rubinstein duality [14] that there is an $\tilde{f}_S \in \mathcal{F}_S$ that attains

$$\begin{aligned} &\inf \, \mathbb{E}_{x \sim \mu}[\, \| \tilde{f}_S(x), f_T(x) \| \,] \\ &\quad = \sup \, \mathbb{E}_{x \sim \mu}[\, \tilde{f}_S(x) \,] - \mathbb{E}_{x \sim \mu}[\, f_T(x) \,]. \end{aligned} \tag{17}$$

Let us define $\tilde{X}(\theta_S) = \{\tilde{f}_S \in \mathcal{F}_S : V(\tilde{f}_S, \theta_S) = W(\mathbb{P}_S, \mathbb{P}_T)\}$ which is non-empty. We know by a simple envelope theorem [10] that

$$\nabla_{\theta_S} W(\mathbb{P}_S, \mathbb{P}_T) = \nabla_{\theta_S} V(\tilde{f}_S, \theta_S), \tag{18}$$

for any $\tilde{f}_S \in \tilde{X}(\theta_S)$ when both terms are well-defined.

Let $\tilde{f}_S \in \tilde{X}(\theta_S)$, which we knows exists since $\tilde{X}(\theta_S)$ is non-empty for all θ_S. Then, we get

$$\begin{aligned} \nabla_{\theta_S} W(\mathbb{P}_S, \mathbb{P}_T) &= \nabla_{\theta_S} V(\tilde{f}_S, \theta_S) \\ &= \nabla_{\theta_S} \mathbb{E}_{x \sim \mu}[\, \| \tilde{f}_S(x), f_T(x) \| \,] \\ &= \nabla_{\theta_S} \mathbb{E}_{x \sim \mu}[\, \tilde{f}_S(x) \,] - \mathbb{E}_{x \sim \mu} [\, f_T(x) \,] \\ &= \nabla_{\theta_S} \mathbb{E}_{x \sim \mu}[\, \tilde{f}_S(x) \,]. \end{aligned} \tag{19}$$

In practice, we use empirical distance between generated images of the student and teacher as loss to update θ_S by back-propagation, *i.e.*,

$$\begin{aligned} \nabla_{\theta_S} \mathbb{E}_{x \sim \mu}[f_S(x)] &= \nabla_{\theta_S} W(\mathbb{P}_S, \mathbb{P}_T) \\ &= \nabla_{\theta_S} W((f_G)_{\#} \mathbb{P}_S, (f_G)_{\#} \mathbb{P}_T) \\ &= \nabla_{\theta_S} \mathbb{E}_{x \sim \mu}[\| f_G \circ f_S(x) - f_G \circ f_T(x) \|] \\ &= \mathbb{E}_{x \sim \mu}[\nabla_{\theta_S} \| f_G \circ f_S(x) - f_G \circ f_T(x) \|], \end{aligned} \tag{20}$$

when $W(\mathbb{P}_S, \mathbb{P}_T) \to 0$, the student function f_S converges to the teacher function f_T. □

1.3. Proof of Theorem 3

Theorem 3. (Generalization Bound) *Let $H \subseteq \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$ be a hypothesis set for C-way classification task. For any $0 < \epsilon < 1/2$ and a sample S of size $m > 4$ drawn according to μ, let $g : \mathbb{R}^C \to \mathbb{R}^n$ be a mapping function of generator*

G with $n \leq \frac{20 \log m}{\epsilon^2}$. Fix $\rho > 0$, for any $1 > \delta > 0$, with probability at least $1-\delta$, the following holds for all $h \in H$,

$$R(h) \leq \hat{R}_\rho(h) + \frac{2C^2}{\rho(1-\epsilon)}\sqrt{\frac{r^2\Lambda^2}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}. \quad (21)$$

For any $x \in \mathcal{X}$, the $\Lambda \geq 0$ and $(\sum_{y=1}^{C} \|h(x,y)\|^p)^{1/p} \leq \Lambda$ for any $p \geq 1$, and the $r > 0$ for $K(x,x) \leq r^2$ where kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite symmetric.

Proof. For the C-way classification task, a hypothesis $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ aims to get y with the minimum distance, i.e. $\arg\min_{y\in\mathcal{Y}} \|\overline{h}(x) - \overline{h}_y\|$ which is equivalent to $\arg\min_{y\in\mathcal{Y}}(1+\epsilon)\|g(\overline{h}(x)) - g(\overline{h}_y)\|$ by Johnson-Lindenstrauss theorem, as the result of x. We define the margin $\rho_h(x,y)$ of the hypothesis h as

$$\rho_h(x,y) = \|g(\overline{h}(x)) - g(\overline{h}_y)\| - \min_{y'\neq y} \|g(\overline{h}(x)) - g(\overline{h}_{y'})\|, \quad (22)$$

where $\overline{h}(x)$ is the vector of $h(x,y), y \in \mathcal{Y}$ and $\overline{h}_y$ use the mean of x which belong to class y as input. g is the mapping function of generator G.

For any $\rho < 0$, we can define the empirical margin loss of hypothesis h for multi-class classification as

$$\hat{R}_\rho(h) = \frac{1}{m}\sum_{i=1}^{m} \Phi_\rho(\rho_h(x_i, y_i)), \quad (23)$$

where Φ_ρ is the margin loss function

$$\Phi_\rho(x) = \begin{cases} 1 & 0 \leq x, \\ 1 - x/\rho & \rho \leq x \leq 0, \\ 0 & x \leq \rho. \end{cases} \quad (24)$$

Thus, empirical margin loss is upper bounded by

$$\hat{R}_\rho(h) \leq \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\rho_h(x_i,y_i)\geq\rho}. \quad (25)$$

Let $\tilde{H} = \{(x,y) \mapsto \rho_h(x,y) : h \in H\}$, consider the family of functions $\tilde{\mathcal{H}} = \{\Phi_\rho \circ r : r \in \tilde{H}\}$ derived from $\tilde{H}$, which take values in $[0,1]$. By Rademacher theorem, with the probability at least $1-\delta$, for all $h \in H$,

$$E[\Phi_\rho(\rho_h(x,y))] \leq \hat{R}_\rho(h) + 2\mathcal{R}_m(\Phi \circ \hat{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}. \quad (26)$$

Since $\mathbb{1}_{\mu\geq 0} \leq \Phi_\rho(\mu)$ for all $\mu \in \mathbb{R}$, the generalization error $R(h)$ is a lower bound on the left-hand side by Johnson-Lindenstrauss theorem, $R(h) = E\left[\mathbb{1}_{\|\overline{h}(x)-\overline{h}_y\| - \min_{y'\neq y}\|\overline{h}(x)-\overline{h}_{y'}\|\geq 0}\right] \leq E[\Phi_\rho(\rho_h(x,y))]$, and we get

$$R(h) \leq \hat{R}_\rho(h) + 2\mathcal{R}_m(\Phi \circ \hat{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}. \quad (27)$$

Let $\rho = -\rho$, because the $(1/\rho)$-Lipschitzness of Φ_p, so that $\mathcal{R}_m(\Phi_p \circ \tilde{H}) \leq \frac{1}{\rho}\mathcal{R}_m(\tilde{H})$. Here, $\mathcal{R}_m(\tilde{H})$ can be upper bounded as follows:

$$\begin{aligned}
\mathcal{R}_m(\tilde{H}) &= \frac{1}{m}\underset{S,\sigma}{E}[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i\rho_h(x_i,y_i)] \\
&= \frac{1}{m}\underset{S,\sigma}{E}[\sup_{h\in H}\sum_{i=1}^{m}\sum_{y\in\mathcal{Y}}\sigma_i\rho_h(x_i,y)\mathbb{1}_{y=y_i}] \\
&\leq \frac{1}{m}\sum_{y\in\mathcal{Y}}\underset{S,\sigma}{E}[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i\rho_h(x_i,y)\mathbb{1}_{y=y_i}] \\
&= \frac{1}{m}\sum_{y\in\mathcal{Y}}\underset{S,\sigma}{E}[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i\rho_h(x_i,y)(\frac{2(\mathbb{1}_{y=y_i})-1}{2}+\frac{1}{2})] \\
&\leq \frac{1}{2m}\sum_{y\in\mathcal{Y}}\underset{S,\sigma}{E}[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i(2(\mathbb{1}_{y=y_i})-1)\rho_h(x_i,y)] + \\
&\qquad \frac{1}{2m}\sum_{y\in\mathcal{Y}}\underset{S,\sigma}{E}[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i\rho_h(x_i,y)] \\
&= \frac{1}{m}\sum_{y\in\mathcal{Y}}\underset{S,\sigma}{E}[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i\rho_h(x_i,y)],
\end{aligned} \quad (28)$$

where $\sigma = (\sigma_1, \ldots, \sigma_m)^T$ with σ_i independent uniform random variables taking values in $\{-1,+1\}$, observing that σ_i and $-\sigma_i$ are distributed in the same way.

Let $\Pi_1(H)^{(C-1)} = \{\min\{h_1, \ldots, h_l\} : h_i \in \Pi_1(H), i \in [1, C-1]\}$. By Johnson-Lindenstrauss theo-

rem, we get

$$
\begin{aligned}
\mathcal{R}_m(\tilde{H}) \leq & \frac{1}{m}\sum_{y\in\mathcal{Y}} \underset{S,\sigma}{E}[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i(\|g(\overline{h}(x)) - g(\overline{h}_y)\| \\
& - \min_{y'\neq y}\|g(\overline{h}(x)) - g(\overline{h}_{y'})\|)] \\
\leq & \frac{1}{m}\sum_{y\in\mathcal{Y}} \underset{S,\sigma}{E}[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i\frac{1}{1-\epsilon}(\|\overline{h}(x_i) \\
& - \overline{h}_y\| - \min_{y'\neq y}\|\overline{h}(x_i) - \overline{h}_{y'}\|)] \\
\leq & \frac{1}{(1-\epsilon)m}\sum_{y\in\mathcal{Y}}[\underset{S,\sigma}{E}[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i\|\overline{h}(x_i) - \overline{h}_y\|] \\
& + \underset{S,\sigma}{E}[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i\min_{y'\neq y}\|\overline{h}(x_i) - \overline{h}_{y'}\|]] \\
\leq & \frac{1}{(1-\epsilon)m}\sum_{y\in\mathcal{Y}}[\underset{S,\sigma}{E}[\sup_{h\in\Pi_1(H)}\sum_{i=1}^{m}\sigma_i h(x_i)] \\
& + \underset{S,\sigma}{E}[\sup_{h\in\Pi_1(H)^{(C-1)}}\sum_{i=1}^{m}\sigma_i h(x_i)]] \\
\leq & \frac{C}{(1-\epsilon)m}[C\underset{S,\sigma}{E}[\sup_{h\in\Pi_1(H)}\sum_{i=1}^{m}\sigma_i h(x_i)]] \\
= & \frac{C^2}{1-\epsilon}[\frac{1}{m}\underset{S,\sigma}{E}[\sup_{h\in\Pi_1(H)}\sum_{i=1}^{m}\sigma_i h(x_i)]] \\
= & \frac{C^2}{1-\epsilon}\mathcal{R}_m(\Pi_1(H)).
\end{aligned}
\tag{29}
$$

Let $K : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a positive definite symmetric kernel and let $h(x,y) = \arg\max_{y\in\mathcal{Y}} w_y\cdot\Phi(x)$, where $\Phi : \mathcal{X}\to\mathbb{R}^n$ be a feature mapping associated to K. We denote W as $W = (w_1^\top,\ldots,w_C^\top)$. For any $p\geq 1$, the family of kernel-based hypotheses is

$$
H = \{h\in\mathcal{R}^{\mathcal{X}\times\mathcal{Y}} : h(x,y)\in\mathbb{R}^n, \|h\|_p\leq\Lambda\}, \tag{30}
$$

where $\|h\|_p = (\sum_{y=1}^{C}\|h(x,y)\|^p)^{1/p}$.

Observe that for all $l\in[1,C]$, we have $\|w_l\|\leq(\sum_{l=1}^{C}\|w_l\|^p)^{1/p} = \|W\|_p\leq\|h\|_p\leq\Lambda$.And for $i\neq j$, $E_\sigma[\sigma_i,\sigma_j] = 0$. The Radmacher complexity of the hypotheses set $\Pi_1(H)$ can be expressed and bounded as follows:

$$
\begin{aligned}
\mathcal{R}_m(\Pi_1(H)) &= \frac{1}{m}\underset{S,\sigma}{E}\left[\sup_{y\in\mathcal{Y},\|W\|\leq\Lambda}\left\langle w_y, \sum_{i=1}^{m}\sigma_i\Phi(x_i)\right\rangle\right] \\
&\leq \frac{1}{m}\underset{S,\sigma}{E}\left[\sup_{y\in\mathcal{Y},\|W\|\leq\Lambda}\|w_y\|\left\|\sum_{i=1}^{m}\sigma_i\Phi(x_i)\right\|\right] \\
&\leq \frac{\Lambda}{m}\underset{S,\sigma}{E}\left[\left\|\sum_{i=1}^{m}\sigma_i\Phi(x_i)\right\|\right] \\
&\leq \frac{\Lambda}{m}\left[\underset{S,\sigma}{E}\left[\left\|\sum_{i=1}^{m}\sigma_i\Phi(x_i)\right\|^2\right]\right]^{1/2} \\
&= \frac{\Lambda}{m}\left[\underset{S,\sigma}{E}\left[\sum_{i=1}^{m}\|\Phi(x_i)\|^2\right]\right]^{1/2} \\
&= \frac{\Lambda}{m}\left[\underset{S,\sigma}{E}\left[\sum_{i=1}^{m}K(x_i,x_i)\right]\right]^{1/2} \\
&\leq \frac{\Lambda\sqrt{mr^2}}{m} = \sqrt{\frac{r^2\Lambda^2}{m}},
\end{aligned}
\tag{31}
$$

which concludes the proof.

□

2. More Results

2.1. Complete Distillation Experiments

We conduct different teacher-student model pairs for distillation experiments, and use ResNet32 / ResNet56 / VGG13 / ResNet110 / ResNet50 / ResNeXt101 as teacher models and use ResNet8 / ResNet32 / VGG11 / MobileNet / ResNet34 / ResNeXt50 as student models. Distillation performance is tested on various datasets, such as MNIST, CIFAR-10, CIFAR-100, Tiny ImageNet, and ImageNet-1K, as top-1 classification accuracy is exploited as an evaluation metric. The experimental results are shown in Tab. 1, Tab. 2 and Tab. 3. For the training of teacher and student models, we adopt the same setting of hyperparameters, so as to verify the distillation effect of student models trained with different methods compared with the teacher model trained with vanilla supervised learning under the same conditions.

We also provide complete ablation results of different data sizes on CIFAR-10 and CIFAR-100, as shown in Tab. 4. We use an effective teacher-student pair of ResNet56 - MobileNet for experiments. The results show that B2KD methods are generally more robust than traditional KD methods for small data sizes, and they can utilize the information in available samples maximumly to model compression in extreme cases. In the comparison of all methods, MEKD achieves the best performance, which also validates the effectiveness and robustness of our proposed method.

Method	Data Size	MNIST				CIFAR-10			
Teacher	50K∼100K	ResNet32	VGG13	ResNet32	ResNet32	ResNet56	VGG13	ResNet56	ResNet56
		99.50	99.52	99.50	99.50	94.15	94.42	94.15	94.15
Student	50K∼100K	ResNet8	VGG11	VGG11	MobileNet	ResNet8	VGG11	VGG11	MobileNet
		99.24	99.41	99.41	99.18	87.74	91.81	91.81	90.04
KD [5]	50K∼100K	99.33	99.44	99.31	99.30	86.58	92.16	92.25	90.43
ML [1]	50K∼100K	99.49	99.40	99.44	99.40	87.89	91.58	91.91	91.19
AL [16]	50K∼100K	99.37	99.26	99.26	99.21	87.25	91.96	91.97	90.54
DKD [18]	50K∼100K	99.33	99.43	99.48	99.42	86.61	92.06	92.42	90.50
DAFL [3]	0K	96.42	97.00	96.14	97.85	60.67	65.41	66.03	69.59
KN [11]	10K	98.61	98.81	98.07	98.54	80.62	81.83	82.41	85.07
AM [15]	10K	99.33	99.47	99.50	99.42	74.89	77.25	74.26	73.65
DB3KD [17]	10K	98.94	99.16	98.91	98.91	78.47	83.72	85.84	81.67
MEKD (soft)	10K	99.40	99.43	99.36	99.25	85.36	86.11	87.27	86.85
MEKD (hard)	10K	99.40	99.45	99.28	99.27	84.45	86.16	87.25	86.53

Table 1. Top-1 classification accuracy (%) of the student model on MNIST and CIFAR-10.

Method	Data Size	CIFAR-100				Tiny ImageNet			
Teacher	50K∼100K	ResNet56	VGG13	ResNet56	ResNet56	ResNet110	VGG13	ResNet110	ResNet110
		72.06	74.68	72.06	72.06	60.71	59.89	60.71	60.71
Student	50K∼100K	ResNet8	VGG11	VGG11	MobileNet	ResNet32	VGG11	VGG11	MobileNet
		59.92	69.12	69.12	68.14	55.47	54.14	54.14	56.07
KD [5]	50K∼100K	53.31	70.88	67.97	71.86	54.14	54.40	49.63	57.85
ML [1]	50K∼100K	54.44	67.78	70.18	73.08	56.56	57.46	56.78	60.07
AL [16]	50K∼100K	58.36	69.92	71.13	71.33	46.02	46.26	45.60	51.29
DKD [18]	50K∼100K	54.28	67.32	70.10	72.38	55.99	55.88	56.52	59.43
DAFL [3]	0K	42.44	43.78	48.32	54.10	38.44	31.93	34.13	40.93
KN [11]	10K	48.75	57.83	55.64	58.49	48.92	46.99	45.05	50.22
AM [15]	10K	50.69	62.17	63.20	65.58	47.72	49.26	47.32	51.54
DB3KD [17]	10K	50.49	63.48	62.76	63.67	47.95	48.46	46.93	50.49
MEKD (soft)	10K	51.87	64.76	64.83	67.07	50.87	51.85	49.95	54.93
MEKD (hard)	10K	51.67	64.72	65.32	67.36	49.89	51.33	49.36	54.71

Table 2. Top-1 classification accuracy (%) of the student model on CIFAR-100 and Tiny ImageNet.

In all experiments, teacher and student models are trained for 350 epochs, except 12 epochs for MNIST. We use Nesterov SGD with momentum 0.9 and weight-decay 0.0005 for training and use a mini-batch size of 128 images on a single NVIDIA GeForce RTX 3090 GPU. The initial learning rate is 0.1, except 0.01 for MNIST, and we conduct a multi-step learning rate schedule which decreases the learning rate by 0.1 at the 116^{th} and 233^{th} epoch for the training of models, except no learning rate schedule is used for MNIST. For the training of student models, we follow the *unsupervised* setting and only use the soft or hard responses of teacher models for distillation. Note that for all experiments, we conduct *three* times experiments and report the mean accuracy.

For the training of DCGAN, we follow the hyperparameters' settings of the work [12]. DCGAN composes of a generator realized by transposed convolution layer and a discriminator realized by an ordinary convolution layer,

Dataset	T - S Pairs	Data Size	KD (soft)	ML (soft)	AL (soft)	DKD (soft)	KN (soft)	AM (soft)	DB3KD (hard)	MEKD (soft)	MEKD (hard)
ImageNet-1K	RN50 - RN34	100K	52.08	54.97	53.50	53.57	56.77	56.92	58.61	59.89	59.32
	RX101 - RX50	100K	54.90	56.58	50.88	55.31	57.43	55.64	59.90	61.21	60.54

Table 3. Top-1 classification accuracy (%) of the student model on ImageNet-1K. We use pretrained RN50 (76.13%) and RX101 (79.31%) as the teacher models, respectively. RN is ResNet and RX is ResNeXt

Dataset	T - S Pairs	Data Size	KD (soft)	ML (soft)	AL (soft)	DKD (soft)	KN (soft)	AM (soft)	DB3KD (hard)	MEKD (soft)	MEKD (hard)
CIFAR10	T: ResNet56	0.1K	16.74	17.78	12.97	20.66	27.67	48.31	43.05	49.04	47.12
	(94.15%)	1K	31.25	31.57	32.05	31.09	58.65	62.05	64.28	69.84	68.66
	S: MobileNet	10K	70.90	73.06	68.61	75.44	85.07	73.65	81.67	86.85	86.53
	(90.04%)	50K(full)	90.43	91.19	90.54	90.50	92.19	86.33	92.46	93.48	93.09
CIFAR100	T: ResNet56	0.1K	01.96	01.88	01.72	02.56	13.23	36.73	30.72	33.56	34.60
	(72.06%)	1K	10.36	10.06	09.62	10.81	35.80	52.09	50.14	53.84	54.52
	S: MobileNet	10K	44.32	48.08	40.57	47.24	58.49	65.58	63.67	67.07	67.36
	(68.14%)	50K(full)	71.86	73.08	71.33	72.38	70.85	71.77	73.36	73.84	73.27

Table 4. Ablation study of data size with top-1 classification accuracy (%) of the student model on CIFAR-10 and CIFAR-100.

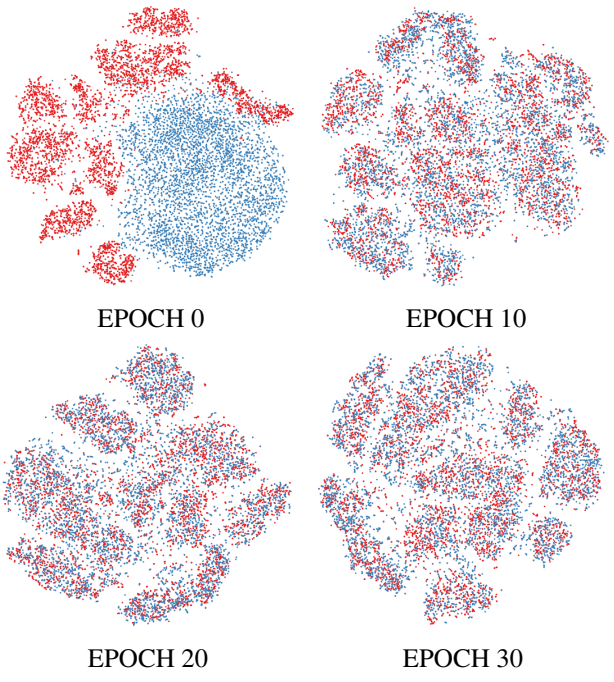

Figure 1. t-SNE visualization of synthetic (blue) and genuine (red) images of MEKD with DCGAN on MNIST.

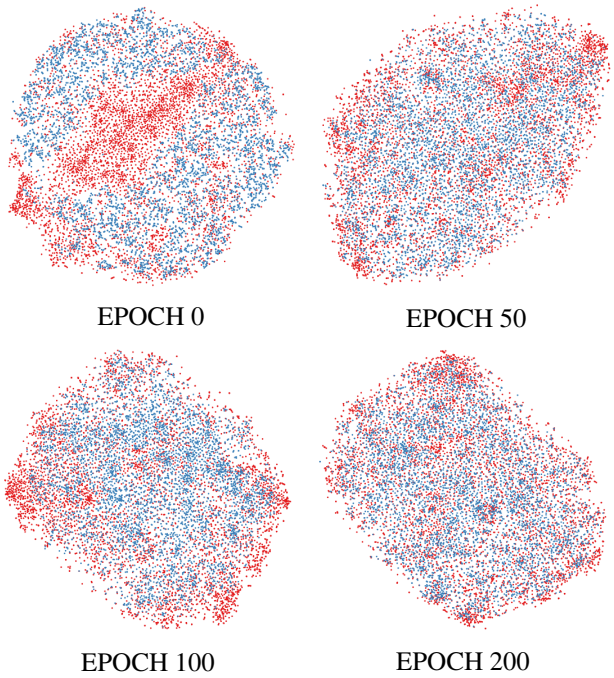

Figure 2. t-SNE visualization of synthetic (blue) and genuine (red) images of MEKD with DCGAN on CIFAR-10.

which greatly reduces the number of network parameters and improves the image generation effect. As an extension of our method, we believe that generative models of different architectures can also be used as emulators to learn the inverse mapping of the teacher function, by adding information maximization (IM) loss to alleviate the problem of mode collapse and achieve the purpose of deprivatization. This will be our research work in the future.

2.2. Visualization Results

We evaluate the training process of DCGAN in terms of whether the generated distribution is consistent with the real distribution, and visualize the synthetic and genuine images by t-SNE projection. As shown in Fig. 1 and Fig. 2, it can be observed that in the training process of DCGAN, the generated distribution is gradually closer to the real distribution. This verifies the effectiveness of using DCGAN as the emulator to learn the inverse mapping of the teacher function, and also proves that DCGAN can indeed alleviate the problem of mode collapse and generate images consistent with the distribution of real images. These synthetic images can not only effectively integrate various patterns in genuine images, but also serve as effective query samples to support the distillation of student models.

References

[1] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *Advances in Neural Information Processing Systems*, 27, 2014.

[2] Jürgen Braun and Michael Griebel. On a constructive proof of kolmogorov's superposition theorem. *Constructive Approximation*, 30(3):653–675, 2009.

[3] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3514–3522, 2019.

[4] Peter Frankl and Hiroshi Maehara. The johnson-lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory, Series B*, 44(3):355–362, 1988.

[5] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint:1503.02531*, 2(7), 2015.

[6] Mario Köppen. On the training of a kolmogorov network. In *International Conference on Artificial Neural Networks*, pages 474–479. Springer, 2002.

[7] Na Lei, Kehua Su, Li Cui, Shing-Tung Yau, and Xianfeng Gu. A geometric view of optimal transportation and generative model. *Computer Aided Geometric Design*, 68:1–21, 2019.

[8] Na Lei, Dongsheng An, Yang Guo, Kehua Su, Shixia Liu, Zhongxuan Luo, Shing-Tung Yau, and Xianfeng Gu. A geometric understanding of deep learning. *Engineering*, 6(3): 361–374, 2020.

[9] Adrian Stephen Lewis and RE Lucchetti. Nonsmooth duality, sandwich, and squeeze theorems. *SIAM Journal on Control and Optimization*, 38(2):613–626, 2000.

[10] Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

[11] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4954–4963, 2019.

[12] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint:1511.06434*, 2015.

[13] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[14] Cédric Villani. *Optimal transport: old and new*. Springer, 2009.

[15] Dongdong Wang, Yandong Li, Liqiang Wang, and Boqing Gong. Neural networks are more productive teachers than human raters: Active mixup for data-efficient knowledge distillation from a blackbox model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1498–1507, 2020.

[16] Yunhe Wang, Chang Xu, Chao Xu, and Dacheng Tao. Adversarial learning of portable student networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[17] Zi Wang. Zero-shot knowledge distillation from a decision-based black-box model. In *International Conference on Machine Learning*, pages 10675–10685. PMLR, 2021.

[18] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11953–11962, 2022.