Sinkhorn Distance Minimization for Knowledge Distillation

Xiao Cui[1], Yulei Qin[2]*, Yuting Gao[2], Enwei Zhang[2], Zihan Xu[2], Tong Wu[2],
Ke Li[2], Xing Sun[2], Wengang Zhou[1] and Houqiang Li[1†]

[1]University of Science and Technology of China, [2]Tencent YouTu Lab
cuixiao@mail.ustc.edu.cn, {zhwg,lihq}@ustc.edu.cn,
{yuleiqin, yutinggao, miyozhang, ianxxu, townswu, tristanli, winfredsun}@tencent.com

Abstract

Knowledge distillation (KD) has been widely adopted to compress large language models (LLMs). Existing KD methods investigate various divergence measures including the Kullback-Leibler (KL), reverse Kullback-Leibler (RKL), and Jensen-Shannon (JS) divergences. However, due to limitations inherent in their assumptions and definitions, these measures fail to deliver effective supervision when few distribution overlap exists between the teacher and the student. In this paper, we show that the aforementioned KL, RKL, and JS divergences respectively suffer from issues of mode-averaging, mode-collapsing, and mode-underestimation, which deteriorates logits-based KD for diverse NLP tasks. We propose the Sinkhorn Knowledge Distillation (SinKD) that exploits the Sinkhorn distance to ensure a nuanced and precise assessment of the disparity between teacher and student distributions. Besides, profit by properties of the Sinkhorn metric, we can get rid of sample-wise KD that restricts the perception of divergence in each teacher-student sample pair. Instead, we propose a batch-wise reformulation to capture geometric intricacies of distributions across samples in the high-dimensional space. Comprehensive evaluation on GLUE and SuperGLUE, in terms of comparability, validity, and generalizability, highlights our superiority over state-of-the-art methods on all kinds of LLMs with encoder-only, encoder-decoder, and decoder-only architectures.

Keywords: Knowledge distillation, Wasserstein distance, Sinkhorn distance

1. Introduction

Large language models (LLMs) such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), T0 (Sanh et al., 2021), and GPT (Radford et al., 2019; Brown et al., 2020) have set state-of-the-art (SOTA) records on various tasks in the field of natural language processing (NLP). On one hand, the scaling laws of LLMs undoubtedly stimulate the development of models with billions of parameters. On the other hand, the surge of model size makes it unaffordable for LLMs to be deployed under resource-constrained environments. Consequently, knowledge distillation (KD), emerging as a cost-efficient approach, has attracted attention from researchers to distill smaller models which maintain highly competitive performance.

One kind of the most representative KD methods is logits-based KD, where the divergence between the distributions of the predicted logits from teacher and student models is measured and minimized for knowledge transfer. The key to effective logits-based KD is exactly the proper measurement of such divergence. Existing studies have experimented with Kullback-Leibler (KL) divergence (Hinton et al., 2015), reverse Kullback-Leibler (RKL) divergence (Tu et al., 2020; Gu et al., 2023b), and Jensen-Shannon (JS) divergence (Wen et al., 2023; Yin et al., 2020; Fang
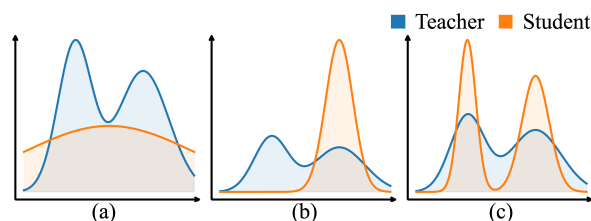

Figure 1: Limitations of existing divergence measures for the student to match the teacher in logits-based distillation. (a) Mode-averaging by Kullback-Leibler divergence. (b) Mode-collapsing by reverse Kullback-Leibler divergence. (c) Mode-underestimation by Jensen-Shannon divergence.

et al., 2021). All these measures can be viewed as variants of the f-divergence measures, which are notoriously limited in quantification of distributions that lack substantial intersections (Arjovsky et al., 2017). Moreover, as illustrated in Fig. 1, each measure has its own drawbacks. KL distillation results in a **mode-averaging** issue (Kim and Rush, 2016; Kim et al., 2021), causing the student to learn an excessively smooth distribution that encompasses the entire support of the teacher distribution. RKL distillation leads to **mode-collapsing** (Arjovsky and Bottou, 2017; Wen et al., 2023), where the student overly focuses on one of the highly probable regions of the teacher distribution and ignores the remaining one. JS distillation gives rise to **mode-underestimation** (Nowozin et al., 2016; Yu et al.,

---

2020) where the student underestimates the probability of rare events due to insufficient penalty.

Another challenge of performing sample-wise KD on LLMs is that for discriminative tasks, the low-dimensional categorical outputs from the teacher provide limited insights into their underlying distributions in the high-dimensional hidden space. One intuitive solution is to bring in a batch of samples to collectively grasp the distribution differences. Nevertheless, existing divergence measures can only independently deal with each sample for logit-by-logit matching because they are not distance metrics and cannot locate the paired teacher and student logits from the same sample in a batch for overall distance minimization.

To address these challenges, we propose **Sin**khorn **K**nowledge **D**istillation, termed as SinKD, for distillation of LLMs[1]. In consideration of generalizability, we tackle logits-based KD in the present study, which would benefit a broad range of applications. Our SinKD employs the Sinkhorn distance (Cuturi, 2013), a variant of the Wasserstein distance (Vallender, 1974a; Frogner et al., 2015a), as divergence measure. The Wasserstein distance quantifies the dissimilarity between two distributions by calculating the minimum cost required to transform one distribution into the other. Compared with traditional divergence measures, it is more sensible as a cost function for distillation. Furthermore, it is differentiable almost everywhere, enabling easy optimization. Despite these advantages, the Wasserstein distance itself is difficult to be computed analytically. Its associated computational cost is prohibitively high for distilling LLMs. Under such circumstance, we propose to use Sinkhorn distance as an efficient approximation, which not only retains all the benefits of Wasserstein distance but also greatly mitigates its cost issue.

A straightforward application of Sinkhorn distance on sample-wise logits matching, though feasible, cannot take full advantage of its perception of structural differences in distributions. Fortunately, Sinkhorn distance is a symmetric metric and its derivation from the optimal transport (OT) imposes explicit constraints on **matching correctness**. It means that given a batch of logits outputs from the teacher and the student respectively as sets A and B, the minimization of the overall Sinkhorn distance between A and B enforces a precise element-wise matching between the two outputs coming from the same sample in a batch. Such properties allow it to work beyond sample-wise distillation. As a result, we propose the batch-wise reformulation. In this way, we can capture geometric structures of the intricate and implicit distributions even through low-dimensional observations. We do not introduce additional network layer or modify output formats specific to NLP tasks.

Extensive experiments are conducted in view of 1) **comparability**, 2) **validity**, and 3) **generalizability**. For comparability, we test SinKD with BERT on the GLUE benchmark (Wang et al., 2018) and it consistently outperforms the SOTA KD methods. For validity, we provide a comprehensive analysis on ablation studies and hyperparameters, Our findings may benefit practitioners on how to adopt SinKD in their own work. For generalizability, we test SinKD on the Super-GLUE benchmark (Wang et al., 2019) with LLMs of various architectures, ranging from the encoder-decoder T0 (Sanh et al., 2021) to the decoder-only GPT-Neo (Black et al., 2021) transformers. Our SinKD showcases robustness across model choices while previous studies merely investigate KD techniques on the encoder-only BERT.

In summary, our contributions are:

- We propose SinKD, a knowledge distillation approach that employs the Sinkhorn distance for divergence measurement. It not only addresses limitations of KL, RKL, and JS divergences under extreme distribution scenarios, but also circumvents the computation burden of Wasserstein distance for distillation.

- We unearth the properties of Sinkhorn distance and reformulate SinKD into batch-wise OT, extending its applicability in NLP tasks.

- Extensive experiments in terms of comparability, validity, and generalizability demonstrate the superiority of SinKD over SOTA methods. We offer practical guidelines of distilling various LLMs for real-world applications.

2. Related Work

2.1. Knowledge Distillation

Knowledge distillation is initially introduced by (Buciluǎ et al., 2006) where an ensemble of models act as the teacher to train a single student model, and now frequently referred to as a model compression technique. Existing KD methods can be simply classified into two categories: 1) logits-based KD and 2) representation-based KD. The logits-based KD is popularized by (Hinton et al., 2015). They force the student to match the predictions of the teacher as soft targets via cross-entropy loss, which is equivalent to minimize the KL divergence between teacher and student probabilities. (Kim and Rush, 2016) bring logits-based KD into generative language models and propose sequence KD. (Sanh et al., 2019) and (Turc

---

[1]Codes and models are available at `https://github.com/2018cx/SinKD`.

et al., 2019) apply KD on BERT for smaller models with minor degradation. (Tu et al., 2020) propose ENGINE to use the reverse KL for distillation of a non-autoregressive translation model. For representation-based KD, the hidden, intermediate representations of input tokens have been utilized as the matching targets of the student (Jiao et al., 2020; Sanh et al., 2019; Wu et al., 2023). There also exist methods that can adapt to either logits-based or representation-based KD (Zhou et al., 2022; Zhang et al., 2023).

In this paper, we primarily focus on logits-based KD and investigate the fundamental problem: *how to transfer label-supplementary knowledge from the teacher to the student with an effective and reliable divergence measure*. Previous studies exploit KL divergence (Hinton et al., 2015), RKL divergence (Tu et al., 2020; Gu et al., 2023b), JS divergence (Wen et al., 2023; Yin et al., 2020; Fang et al., 2021), and sophisticated distance measures (Sun et al., 2019; Jin et al., 2019; Park et al., 2019, 2021a) for distillation. However, these methods do not consistently capture subtle distribution differences and tend to take "shortcuts" in student imitating the teacher, which motivates our exploration of an alternative divergence measure.

2.2. Sinkhorn Distance

We first introduce the Wasserstein distance as a foundation for the Sinkhorn distance. It is a dissimilarity metric derived by the mass transportation theory of two probability measures. Since the Wasserstein distance takes into account the underlying geometry of the distribution space (Frogner et al., 2015a; Vallender, 1974a,b; Villani and Villani, 2009), it enjoys high popularity in generative adversarial networks (Arjovsky et al., 2017; Gulrajani et al., 2017; Peyré et al., 2019) and unsupervised learning (Gu et al., 2023a; Chen et al., 2022; He et al., 2022). However, the Wasserstein distance is too costly to be computed and its efficient approximation is a prerequisite for distillation. The Sinkhorn distance stems from it and incorporates an extra entropy regularization term to make the OT tractable. It is informally defined by the minimum transport cost of an entropy-regularized OT plan (Cuturi, 2013), and has been successful in classification (Frogner et al., 2015b; Liu et al., 2023), machine translation (Li et al., 2023), domain adaptation (Courty et al., 2017; Nguyen and Luu, 2022), and generative modeling (Genevay et al., 2018; Kammammettu and Li, 2023).

For distillation of LLMs, especially under discriminative tasks, the vanilla sample-wise SinKD cannot make the best use of its desirable properties in perceiving structural differences between distributions. On the contrary, we propose the batch-wise SinKD to make up the insufficient knowledge revealed from the low-dimensional outputs of the teacher, improving its generalization over tasks.

3. Methodology

In this section, we first review classic divergence measures and analyze their drawbacks. Then, we present details of SinKD with an OT framework.

3.1. Problem Statement

Given a sample $\mathbf{x}_i$ and its ground-truth label $\mathbf{y}_i \in \mathbb{R}^d$ in the training set, the output logits with softmax activation σ_τ from the teacher f_T and the student f_S are respectively $\mathbf{t}_i \in \mathbb{R}^d$ and $\mathbf{s}_i \in \mathbb{R}^d$:

$$\mathbf{t}_i = \sigma_\tau(f_T(\mathbf{x}_i)), \quad \mathbf{s}_i = \sigma_\tau(f_S(\mathbf{x}_i)), \quad (1)$$

where τ is the temperature and d is the dimension of the output logits. The objective of KD is to minimize the measured divergence $J(\mathbf{t}_i, \mathbf{s}_i)$.

3.2. Classic Divergence Measures

KL Divergence It quantifies the amount of information lost when $\mathbf{s}_i$ approximates $\mathbf{t}_i$ as:

$$J_{\mathsf{KL}}(\mathbf{t}_i, \mathbf{s}_i) \approx \sum_{j=1}^{d}(-\mathbf{t}_{i(j)} \log \mathbf{s}_{i(j)} + \mathbf{t}_{i(j)} \log \mathbf{t}_{i(j)}).$$

$$(2)$$

Here, j denotes the index of an element in a vector. Despite its popularity, KL divergence suffers from three limitations. First, it is asymmetric with $J_{\mathsf{KL}}(\mathbf{t}_i, \mathbf{s}_i) \neq J_{\mathsf{KL}}(\mathbf{s}_i, \mathbf{t}_i)$, which introduces inconsistencies due to its violation of the property as a distance metric. Second, the student model optimized by the KL loss attempts to average the teacher's multimodal distribution, ending up with an underfitting of these modes. This is known as the mode-averaging problem. Consequently, the student fails to capture all crucial patterns of data and ultimately impacts performance. Third, the KL divergence corresponds to a non-smooth function, posing challenges to model optimization.

RKL Divergence It addresses the issue of mode-averaging associated with $J_{\mathsf{KL}}(\mathbf{t}_i, \mathbf{s}_i)$:

$$J_{\mathsf{RKL}}(\mathbf{t}_i, \mathbf{s}_i) \approx \sum_{j=1}^{d}(\mathbf{s}_{i(j)} \log \mathbf{s}_{i(j)} - \mathbf{s}_{i(j)} \log \mathbf{t}_{i(j)}).$$

$$(3)$$

However, it shares the inherent asymmetry with KL which leads to inconsistencies in capturing differences. Furthermore, the student optimized by a RKL loss tends to pay attention only to highly likely events of the teacher's distribution, which is known as mode-collapsing. Accordingly, if the teacher assigns zero-probability to an event, the student is
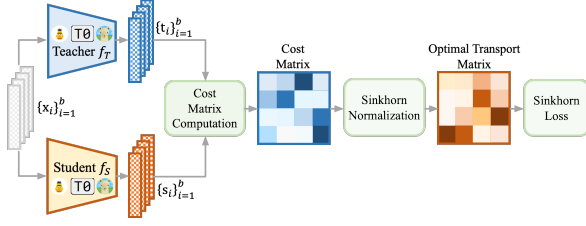
Figure 2: Illustration of our SinKD pipeline.

compelled to do the same. This "zero-forcing" effect could be problematic as the student lacks the capacity to track the complete distribution of the teacher, resulting in suboptimal performance.

JS Divergence   It combines both KL and RKL by:

$$J_{\text{JS}}(\mathbf{t}_i, \mathbf{s}_i) \approx \frac{1}{2} \sum_{j=1}^{d} (-\mathbf{s}_{i(j)} \log \mathbf{m}_{i(j)} + \mathbf{s}_{i(j)} \log \mathbf{s}_{i(j)}$$
$$- \mathbf{t}_{i(j)} \log \mathbf{m}_{i(j)} + \mathbf{t}_{i(j)} \log \mathbf{t}_{i(j)}), \tag{4}$$

where $\mathbf{m}_i = \frac{1}{2}(\mathbf{t}_i + \mathbf{s}_i)$. While the JS divergence overcomes the asymmetry shortcoming of the KL divergence, it is still subject to non-smoothness that makes it challenging to optimize. Moreover, the student may excessively underestimate the probability of rare events as the JS loss does not penalize adequately for matching low probability regions. There also exists a risk of gradient vanishing when $J_{\text{JS}}(\mathbf{t}_i, \mathbf{s}_i)$ degenerates as a constant on distributions with few or even no overlap.

3.3. Sinkhorn Distance

Sinkhorn distance is based on the relaxed formulation of an OT plan with entropy regularization. It considers the minimum cost of mass transmission in converting one probability into the other. Specifically, we first define the Wasserstein distance below. It involves the set of a transportation polytope $U(\mathbf{t}_i, \mathbf{s}_i)$, which consists of all matrices of $\mathbf{P} \in \mathbb{R}_{+}^{d \times d}$ that satisfy the following constraints:

$$U(\mathbf{t}_i, \mathbf{s}_i) = \{\mathbf{P} \in \mathbb{R}_{+}^{d \times d} | \mathbf{P}\mathbf{1}_d = \mathbf{s}_i, \mathbf{P}^\mathsf{T}\mathbf{1}_d = \mathbf{t}_i\}, \tag{5}$$

where $\mathbf{1}_d \in \mathbb{R}^d$ is a vector of ones. Given a cost matrix $\mathbf{D} \in \mathbb{R}^{d \times d}$, the Wasserstein distance is:

$$J_{\text{WD}}(\mathbf{t}_i, \mathbf{s}_i) = \min_{\mathbf{P} \in U(\mathbf{t}_i, \mathbf{s}_i)} \langle \mathbf{P}, \mathbf{D} \rangle = \sum_{m,n} \mathbf{P}_{m,n} \mathbf{D}_{m,n}, \tag{6}$$

where $\mathbf{D}_{m,n}$ is usually the absolute difference between the m-th and n-th elements of $\mathbf{t}_i$ and $\mathbf{s}_i$:

$$\mathbf{D}_{m,n} = |\mathbf{t}_{i(m)} - \mathbf{s}_{i(n)}|. \tag{7}$$

To circumvent the substantial computation entailed by solving such an OT problem, Sinkhorn distance

is proposed as a fast approximation to the Wasserstein distance for a constrained optimization (Cuturi, 2013). It is defined as the inner product between the OT plan $\mathbf{P}^\lambda$ and the cost matrix $\mathbf{D}$:

$$J_{\text{SD}}(\mathbf{t}_i, \mathbf{s}_i) = \langle \mathbf{P}^\lambda, \mathbf{D} \rangle, \tag{8}$$

where $\lambda > 0$ is the weight for entropy regularization. The OT plan $\mathbf{P}^\lambda$ is obtained by minimizing:

$$\mathbf{P}^\lambda = \underset{\mathbf{P} \in U(\mathbf{t}_i, \mathbf{s}_i)}{\operatorname{argmin}} \langle \mathbf{P}, \mathbf{D} \rangle - \lambda h(\mathbf{P}), \tag{9}$$

where $h(\mathbf{P})$ is the entropy of the matrix $\mathbf{P}$. The entropy term encourages the transport plan to be more spread out for easier optimization. The vanilla solution to $\mathbf{P}^\lambda$ by sample-wise Sinkhorn normalization (Cuturi, 2013) is performed between $\mathbf{t}_i$ and $\mathbf{s}_i$ in a manner of iterative updates:

$$\left(\mathbf{u}^t, \mathbf{v}^t\right) \leftarrow \left(\mathbf{t}_i \oslash \left(\mathbf{K}^\mathsf{T}\mathbf{v}^{t-1}\right), \mathbf{s}_i \oslash \left(\mathbf{K}\mathbf{u}^{t-1}\right)\right), \tag{10}$$

where $\oslash$ indicates element-wise division and t denotes the iteration time. Two vectors $\mathbf{u} \in \mathbb{R}^d, \mathbf{v} \in \mathbb{R}^d$ are non-negative. The kernel matrix $\mathbf{K} \in \mathbb{R}^{d \times d}$ is constructed by applying the Gaussian kernel on $\mathbf{D}$ with the weight λ for entropy regularization:

$$\mathbf{K} = \exp(-\frac{\mathbf{D}}{\lambda}). \tag{11}$$

Finally, $\mathbf{P}^\lambda$ is defined as:

$$\mathbf{P}^\lambda = \operatorname{diag}\left(\mathbf{v}^t\right) \mathbf{K} \operatorname{diag}\left(\mathbf{u}^t\right). \tag{12}$$

3.4. Batch-wise Reformulation

In view of properties of the Sinkhorn distance metric, we can get rid of the sample-wise KD that only works on each teacher-student sample pair, and instead perform KD on groups of teacher and student samples. A batch of b samples all participate in divergence measures with their overall output logits $\mathbf{t} \in \mathbb{R}^{b \times d}$ and $\mathbf{s} \in \mathbb{R}^{b \times d}$ respectively from the teacher and the student. It thereby increases the dimension of the "observational" space via batch-wise reformulation especially when $d \ll b$ holds.

Cost Matrix Computation   We employ the ℓ_p-norm to measure the pairwise differences between the i-th and j-th samples in a batch for the entry $\mathbf{D}_{i,j}$ of the "batchified" cost matrix $\mathbf{D} \in \mathbb{R}^{b \times b}$:

$$\mathbf{D}_{i,j} = \|\mathbf{t}_i - \mathbf{s}_j\|_p. \tag{13}$$

Sinkhorn Normalization   Before we propose the batch-wise Sinkhorn normalization, we reformulate the sample-wise solution to $\mathbf{P}^\lambda$ (Eq. 10) into a equivalent vector-form with iterations only on $\mathbf{K}$:

$$\widehat{\mathbf{K}}^t \leftarrow \operatorname{diag}\left(\mathbf{K}^{t-1}\mathbf{1}_d \oslash \mathbf{s}_i\right)^{-1} \mathbf{K}^{t-1},$$
$$\mathbf{K}^t \leftarrow \widehat{\mathbf{K}}^t \operatorname{diag}\left(\left(\widehat{\mathbf{K}}^t\right)^\mathsf{T} \mathbf{1}_d \oslash \mathbf{t}_i\right)^{-1}, \tag{14}$$

where $\mathbf{K}^0 = \mathbf{K} \in \mathbb{R}^{d \times d}$ is defined in Eq. 11. For distillation beyond the d-dimensional space, we propose a more compact solution in the matrix-form for batch-wise normalization with $\mathbf{K} \in \mathbb{R}^{b \times b}$:

$$
\begin{aligned}
\widehat{\mathbf{K}}^t &\leftarrow \operatorname{diag}\left(\mathbf{K}^{t-1}\mathbf{1}_b \oslash \mathbf{w}_s\right)^{-1}\mathbf{K}^{t-1}, \\
\mathbf{K}^t &\leftarrow \widehat{\mathbf{K}}^t\operatorname{diag}\left(\left(\widehat{\mathbf{K}}^t\right)^{\mathsf{T}}\mathbf{1}_b \oslash \mathbf{w}_t\right)^{-1},
\end{aligned}
\tag{15}
$$

where $\mathbf{w}_s$ and $\mathbf{w}_t$ respectively represent the weights of each element involved in the batch-wise KD from the student and teacher. Without loss of generality, we assume uniform distributions with $\mathbf{w}_s = \mathbf{w}_t = \frac{1}{b}\mathbf{1}_b$. Given such conditions, updates on $\mathbf{K}^t$ (Eq. 15) can be further simiplified as:

$$
\begin{aligned}
\widehat{\mathbf{K}}^t &\leftarrow \mathbf{K}^{t-1} \oslash \left(\mathbf{K}^{t-1}\mathbf{1}_b\mathbf{1}_b^{\top}\right), \\
\mathbf{K}^t &\leftarrow \widehat{\mathbf{K}}^t \oslash \left(\mathbf{1}_b\mathbf{1}_b^{\top}\widehat{\mathbf{K}}^t\right).
\end{aligned}
\tag{16}
$$

Out of simplicity, irrelevant constants are excluded from the equations above. With a pre-determined number of iterations T, the OT matrix is derived:

$$
\mathbf{P}^\lambda = \mathbf{K}^T
\tag{17}
$$

Sinkhorn Loss The batch-wise SinKD loss is:

$$
\mathcal{L}_{\mathsf{SD}} = J_{\mathsf{SD}}(\mathbf{t}, \mathbf{s}) = \langle \mathbf{P}^\lambda, \mathbf{D} \rangle = \sum_{i,j}\mathbf{K}_{i,j}^T\mathbf{D}_{i,j}
\tag{18}
$$

We illustrate the entire pipeline in Fig. 2.

Total Losses For each batch of b samples, we use the cross-entropy loss $\mathcal{L}_{\mathsf{CE}}$, the KL loss $\mathcal{L}_{\mathsf{KL}}$, and the Sinkhorn loss $\mathcal{L}_{\mathsf{SD}}$ for distillation:

$$
\begin{aligned}
\mathcal{L} = \sum_{i=1}^{b}&[(1-\alpha)\mathcal{L}_{\mathsf{CE}}(\mathbf{y}_i, \mathbf{s}_i) \\
&+ \alpha\mathcal{L}_{\mathsf{KL}}(\mathbf{t}_i, \mathbf{s}_i)] + \beta\mathcal{L}_{\mathsf{SD}},
\end{aligned}
\tag{19}
$$

where α and β are weights, and $\mathcal{L}_{\mathsf{KL}}(\mathbf{t}_i, \mathbf{s}_i) \approx \mathcal{L}_{\mathsf{CE}}(\mathbf{t}_i, \mathbf{s}_i)$ given that the second term in $J_{\mathsf{KL}}(\mathbf{t}_i, \mathbf{s}_i)$ can be viewed as a constant in distillation.

Alternative D Apart from Eq. 13, we can further take into account all the d-dimensional logits of b samples by flattenning $\mathbf{t}$ and $\mathbf{s}$ for a $\mathbf{D} \in \mathbb{R}^{bd \times bd}$:

$$
\mathbf{D}_{im,jn} = |\mathbf{t}_{i(m)} - \mathbf{s}_{j(n)}|.
\tag{20}
$$

Accordingly, the sinkhorn normalization is performed on $\mathbf{K} \in \mathbb{R}^{bd \times bd}$ with $\mathbf{w}_s = \mathbf{w}_t = \frac{1}{bd}\mathbf{1}_{bd}$. In this case, SinKD takes a broader perspective of the batch distributions with a multiplied dimension of bd, significantly exceeding the sample-wise KD.

4. Experimental Settings

4.1. Datasets

We evaluate our method on seven tasks of the GLUE benchmark (Wang et al., 2018), including CoLA (Warstadt et al., 2019), SST-2 (Socher et al., 2013), MNLI (Williams et al., 2018), MRPC (Dolan and Brockett, 2005), RTE (Bentivogli et al., 2009), QNLI (Rajpurkar et al., 2016) and QQP (Chen et al., 2018). For evaluation metrics, we follow previous works (Wu et al., 2023; Zhang et al., 2023; Zhou et al., 2022) to report accuracy (MNLI, SST-2, QNLI, QQP, and RTE), F1 score (MRPC), and Matthews correlation coefficient (CoLA). Following (Zhang et al., 2023; Zhou et al., 2022), the regression-oriented STS-B (Cer et al., 2017) is not validated due to its problem settings. Note that all discriminative tasks of GLUE are associated with extremely-low dimension of logits output ($d = 3$ for MNLI and $d = 2$ for the remainings tasks).

4.2. Implementation Details

Our SinKD is implemented with PyTorch and Transformers (Wolf et al., 2020). For comparability, we follow AD-KD (Wu et al., 2023) to set BERT$_{\mathsf{base}}$ as the teacher and a smaller BERT$_6$ (Turc et al., 2019) as the student for task-specific fine-tuning. For generalizability, we also validate SinKD on T0 (Sanh et al., 2021) and GPT-Neo (Black et al., 2021). Note that for all GLUE tasks except MNLI, two definitions of $\mathbf{D}$ (Eqs. 13,20) are equivalent given the constraint of $\sum_{m=1}^{d}\mathbf{t}_{i(m)} = 1$ and $d = 2$. Consequently, we use the default $\mathbf{D}$ by Eq. 13. Out of simplicity, we set $p = 1$ (ℓ_1-norm) for $\mathbf{D}$. The hyper-parameters are optimized via grid search to determine the learning rate $lr \in \{2e-5, 3e-5, 4e-5, 5e-5\}$, $\alpha \in \{0.8, 0.9, 1.0\}$, $b \in \{16, 32, 64\}$, and $\tau_{\mathsf{KL}} \in \{1, 2, 3, 4\}$. We empirically set $\tau_{\mathsf{SD}} = 2$, $\lambda = 0.1$, $T = 20$, and $\beta = 0.8$. Discussions on the effect of T, λ, τ_{KL}, b, τ_{SD}, α, and β can be found in Sec. 5.3.

4.3. Baselines

We compare SinKD with SOTA KD methods on logits and representations. For logits-based KD, we include the vanilla KD (Hinton et al., 2015), RCO (Jin et al., 2019), DML (Zhang et al., 2018), PD (Turc et al., 2019), and ReAugKD (Zhang et al., 2023). For representation-based KD, we compare PKD (Sun et al., 2019), Tiny-BERT (Jiao et al., 2020), RKD (Park et al., 2019), CKD (Park et al., 2021b), SFTN (Park et al., 2021a), TAKD (Mirzadeh et al., 2020), ProKT (Shi et al., 2020), MGSKD (Shi et al., 2020), MetaDistill (Zhou et al., 2022), and AD-KD (Wu et al., 2023). For a fair comparison, we follow (Wu et al., 2023)

to exclude MiniLM (Wang et al., 2020) and Mobile-BERT (Sun et al., 2020) as their two-stage settings involve both task-agnostic and task-specific distillation. In contrast, we emphasize a more generalized one-stage setting where no extra efforts are required for pre-training. Note that baseline results are quoted (Wu et al., 2023; Zhang et al., 2023).

5. Results and Discussions

5.1. Comparison with SOTA

Tab. 1 shows that SinKD outperforms all baselines on most datasets. Specifically, SinKD achieves an average increase of 0.47% and 1.17% over AD-KD (Wu et al., 2023) and ReAugKD (Zhang et al., 2023), respectively. Compared with AD-KD, SinKD reduces the performance gap between the student and the teacher over 57%, highlighting that SinKD effectively narrows such gap by injecting structural knowledge from teacher to student. Our improvements can be attributed to the unique properties of Sinkhorn distillation, where the integrated characteristics of distributions are respected during distillation and thereafter facilitate impartial, efficient knowledge transfer for robust convergence. We also notice that SinKD does not rank the top on QNLI, possibly due to suboptimal hyper-parameters for this specific task. Meticulous tuning of hyper-parameters might yield better results, but will impair comparability and therefore is beyond the scope of the present study.

5.2. Ablation Study

Sinkhorn loss benefits the student the most among all losses. In order to study the impact of each loss component, we carry out ablation studies on three variations of SinKD: 1) SinKD without Sinkhorn loss, 2) SinKD without KL divergence loss, and 3) SinKD without cross-entropy loss. As revealed in Tab. 2, significant decreases over all tasks can be observed when Sinkhorn loss is removed. In addition, the drop of performance without KL divergence loss suggests that the proposed SinKD is supplementary to the vanilla KL divergence in distribution measurements. With respect to the cross-entropy loss, its supervision from ground-truth labels directly improves the student model and consequently should be kept intact during distillation. Each component contributes to diminishing the gap between the student and the teacher. Our proposed Sinkhorn loss brings the most pronounced gains over other losses, confirming the validity of Sinkhorn distance as a stable metric for convergence to global optimum.

Batch-wise SinKD excels sample-wise SinKD. Tab. 3 demonstrates the superiority of the batch-wise over the sample-wise SinKD on all tasks, implying that the Sinkhorn distance is indeed adept in handling the deviation of the student from the teacher with a high-dimensional distribution. The sample-wise distillation treats each instance independently while neglecting the overall tendency of the student in tracking distributions of the teacher.

SinKD surpasses distillation methods based on variants of f-divergence. To investigate if the existing distillation methods with f-divergence measures can achieve competitive results, we replace our Sinkhorn loss with losses based on: 1) RKL divergence, 2) JD divergence, and 3) total variation distance (TVD). To fairly compare with SinKD, each loss mentioned above is combined with cross-entropy loss and KL divergence loss during distillation. Tab. 4 shows that Sinkhorn distillation outperforms three other distillation methods on all datasets, verifying the superiority of Sinkhorn distance over variants of f-divergence measures in matching distributions. Additionally, it is worth noting that among the other three methods, TVD exhibits slight advantages over RKL and JS divergence on average. Such finding is consistent with previous work (Wen et al., 2023).
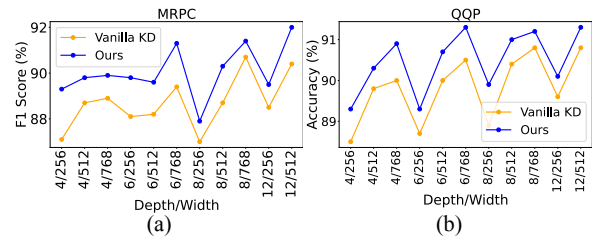

Figure 3: Performance at different student scales on (a) MRPC & (b) QQP. Best viewed magnified.

SinKD generalizes well on student LLMs across scales. To thoroughly assess the influence of size of student LLMs on the performance of SinKD, we conduct an extensive analysis with comparison between the vanilla KD and SinKD. Without loss of generality, we take two tasks (MRPC and QQP) for demonstration. A broad range of model scales (Turc et al., 2019) are employed to explore the adaptability and robustness of SinKD when applied on student models with various configurations. Note that both the vanilla KD and our SinKD are logits-based KD methods, which are independent of model structure by nature and thus enjoy high versatility. As illustrated in Fig. 3, SinKD consistently outperforms the vanilla KD on both two tasks across all scales. Such generalizability on model size confirms the potential of SinKD as an efficient and reliable KD method.

Method	#Params.	COLA (MCC)	SST-2 (ACC)	MNLI-(m/mm) (ACC)	MRPC (F1)	RTE (ACC)	QNLI (ACC)	QQP (ACC)	Avg
BERT$_{base}$ (T) (Devlin et al., 2018)	110M	60.3	93.7	84.9/84.8	91.4	71.1	91.7	91.5	83.28
BERT$_6$ (S) (Turc et al., 2019)	66M	51.2	91.0	81.7/82.6	89.2	66.1	89.3	90.4	79.53
Task-specific Representation-based Distillation									
PKD (Sun et al., 2019)	66M	45.5	91.3	81.3/-	85.7	66.5	88.4	88.4	77.63
TinyBERT (Jiao et al., 2020)	66M	53.8	92.3	83.1/83.4	88.8	66.5	89.9	90.5	80.3
RKD (Park et al., 2019)	66M	53.4	91.7	-	86.1	68.6	89.5	90.9	80.03
CKD (Park et al., 2021b)	66M	55.1	93.0	83.6/84.1	89.6	67.3	90.5	91.2	81.11
SFTN (Park et al., 2021a)	66M	53.6	91.5	-	85.3	68.5	89.5	90.4	79.80
TAKD (Mirzadeh et al., 2020)	66M	53.8	91.4	-	85.0	68.5	89.6	90.7	79.83
ProKT (Shi et al., 2020)	66M	54.3	91.3	-	86.3	68.4	89.7	90.9	80.15
MGSKD (Liu et al., 2022)	66M	49.1	91.7	83.3/83.9	89.8	67.9	90.3	91.2	80.00
MetaDistill (Zhou et al., 2022)	66M	58.6	92.3	-	86.8	69.4	90.4	91.0	81.42
AD-KD (Wu et al., 2023)	66M	58.3	91.9	83.4/**84.2**	91.2	70.9	**91.2**	91.2	82.45
Task-specific Logits-based Distillation									
Vanilla KD (Hinton et al., 2015)	66M	53.6	91.1	82.7/83.1	89.4	66.8	90.1	90.5	80.25
RCO (Jin et al., 2019)	66M	53.6	91.4	-	85.1	67.6	89.7	90.6	79.67
DML (Zhang et al., 2018)	66M	53.7	91.5	-	85.1	68.4	89.6	90.3	79.77
PD (Turc et al., 2019)	66M	-	91.1	82.5/83.4	89.4	66.7	89.4	90.7	-
ReAugKD (Zhang et al., 2023)	66M	59.4	92.5	-	86.3	70.4	90.7	91.2	81.75
SinKD (ours)	66M	**60.2**	**93.1**	**83.8/84.2**	**91.3**	**71.1**	90.5	**91.3**	**82.92**

Table 1: Comparison with SOTA methods on GLUE with BERT$_{base}$ as the teacher (T) and BERT$_6$ as the student (S). All scores are averaged except the accuracy of MNLI-(m/mm).

Method	COLA (MCC)	SST-2 (ACC)	MNLI (ACC)	MRPC (F1)	RTE (ACC)	QNLI (ACC)	QQP (ACC)
SinKD (ours)	60.2	93.1	83.8/84.2	91.3	71.1	90.5	91.3
w/o $\mathcal{L}_{SD}$	53.6	91.1	82.7/83.1	89.4	66.8	90.1	90.5
w/o $\mathcal{L}_{KL}$	56.2	91.7	82.3/83.0	90.1	69.3	90.2	90.7
w/o $\mathcal{L}_{CE}$	58.0	92.3	83.5/84.1	91.1	70.4	90.4	91.3
w/o $\mathcal{L}_{KL}$&$\mathcal{L}_{SD}$	51.2	91.0	81.7/82.6	89.2	66.1	89.3	90.4

Table 2: Effect of different loss terms on GLUE.

5.3. Discussion on Hyper-parameters

T **as the number of Sinkhorn iterations** We vary the number of iterations T and results (see Tab. 5) reflect the importance of selecting an appropriate T. An increase of T to 20 respectively improves F1 scores for MRPC (91.3) and accuracy for QQP (91.3), suggesting that sufficient iterations is crucial to approximation and convergence. Nevertheless, raising the iterations to 50 yields no further improvement. It indicates the existence of a saturation point, beyond which additional iterations are not beneficial but redundant. Hence, we set $T = 20$ throughout experiments.
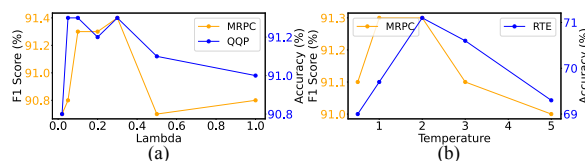

Figure 4: Effect of (a) λ on MRPC & QQP and (b) τ_{SD} on MRPC & RTE. Best viewed magnified.

λ **as the weight of entropy-regularization** The Sinkhorn distance is derived from the entropy-regularized OT problem, where the regularization term promotes a more dispersed and less concentrated OT plan. In other words, entropy-regularization would enhance the numerical stability and computational tractability of the solution to OT problem. Theoretically, λ dictates the balance between the accuracy of the OT approximation and the stability of the solution. A larger λ results in a smoother and more stable solution, albeit potentially less accurate. A smaller λ yields a more accurate solution at the risk of numerical instability. As demonstrated in Fig. 4(a), a λ within the range of 0.1 to 0.3 appears to achieve an optimal trade-off among various aspects. Out of consistency, we choose $\lambda = 0.1$ throughout experiments.

τ_{SD} **as the temperature in Sinkhorn loss** Fig. 4(b) systematically investigates the influence of τ_{SD} on distillation on the tasks of MRPC and QQP. Our findings indicate that the default empirical setting $\tau_{SD} = 2$ is appropriate for both two

Level	COLA (MCC)	SST-2 (ACC)	MNLI (ACC)	MRPC (F1)	RTE (ACC)	QNLI (ACC)	QQP (ACC)
Sample-wise	54.2	91.5	83.2/83.7	90.4	69.0	90.3	91.2
Batch-wise	60.2	93.1	83.8/84.2	91.3	70.0	90.5	91.3

Table 3: Comparison between the sample-wise and batch-wise SinKD on GLUE.

Method	Complexity	COLA (MCC)	SST-2 (ACC)	MNLI (ACC)	MRPC (F1)	RTE (ACC)	QNLI (ACC)	QQP (ACC)
RKL	$O(bd)$	53.9	91.6	82.9/83.4	90.5	67.1	90.1	91.1
JS	$O(bd)$	54.2	92.2	83.1/83.7	90.7	68.9	90.3	91.2
TVD	$O(bd)$	54.1	92.1	83.3/83.8	90.9	70.0	90.2	91.2
SinKD	$O\left(b^2\left(d+T\right)\right)$	60.2	93.1	83.8/84.2	91.3	71.1	90.5	91.3

Table 4: Comparison with distillation methods based on variants of f-divergence on GLUE.

Number of Interations T	MRPC (F1)	QQP (ACC)
2	90.6	91.0
5	90.9	91.0
10	90.9	91.1
20	91.3	91.3
50	91.3	91.3

Table 5: Effect of T on MRPC & QQP.

tasks. A smaller τ_{SD} may cause the student model to concentrate solely on learning the most salient features, neglecting the nuanced but valuable information present in less probable categories for classification. On the other hand, a larger τ_{SD} results in smoother and more uniform probability distributions, which confuses the student model to discern between essential and irrelevant information.
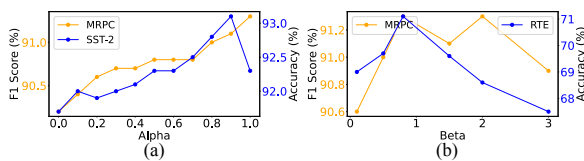

Figure 5: Effect of (a) α on MRPC & SST-2 and (b) β on MRPC & RTE. Best viewed magnified.

b **as the number of batch size** The batch size is closely associated with the efficiency of geometric structural learning since the distribution divergences are measured within each batch of samples for the proposed Sinkhorn distance minimization. An increased batch size contributes to enhancing the student's understanding of complex geometric inter-sample relations present within the dataset. Empirical evidence, as presented in Tab. 6, elucidates a positive correlation between the batch size and evaluation results (F1 scores for the MRPC benchmark and accuracy for SST-2). Such performance gains are theoretically grounded in the premise that larger batches provide a more expansive dimensional space, allowing for a more comprehensive representation of the geometric configuration during each optimization step. A larger batch size b effectively provides the model with exposure to the intrinsic geometric variance of the dataset, potentially accelerating the transfer and assimilation of the teacher model's knowledge. However, such benefit becomes negligible when the batch size increases beyond 32, where both metrics for MRPC and SST-2 remain almost unchanged. This observation suggests the existence of a saturation point, which delineates the boundary where the advantages of augmenting the geometric sampling space are outweighed by the computational overhead.

Batch size b	MRPC (F1)	SST-2 (ACC)
2	90.5	91.3
8	90.8	92.4
16	91.3	92.8
32	91.1	93.1
64	91.3	93.1

Table 6: Effect of b on MRPC & SST-2.

Temperature τ_{KL}	MRPC (F1)	SST-2 (ACC)
1	90.5	92.6
2	90.8	93.1
3	91.1	92.7
4	91.3	92.5

Table 7: Effect of τ_{KL} on MRPC & SST-2.

τ_{KL} **as the temperature in KL loss** Tab. 7 provides the results of how the temperature τ_{KL} affects the knowledge distillation. For the MRPC dataset, a monotonically increasing trend in the F1-score is observed as τ_{KL} ranges from 1 to 4. The best results of F1-score are achieved at $\tau_{\text{KL}} = 4$. Conversely, the accuracy for SST-2 is maximized at a lower temperature ($\tau_{\text{KL}} = 2$), beyond which a diminution occurs. It exemplifies the dualistic role of τ_{KL}: 1) refining the granularity of probability distributions at lower temperatures and 2) fostering generalization at higher settings. The optimal value of τ_{KL} is to be task-dependent and therefore necessitates task-specific tuning.

α **and** β **as the loss weights** In the total training objectives of SinKD, we introduce α and β to balance the contributions from the cross-entropy loss, KL divergence loss, and Sinkhorn distance loss. A comprehensive evaluation of various combinations of α and β can be found in Fig. 5. Each time, we only adjust one parameter and keep the other one fixed. Our findings indicate that a larger α generally produces better performance, corroborating that knowledge transfer from the teacher model does play an indispensable role. In line with the results of SinKD without the cross-entropy loss (see Tab. 2), $\alpha = 1$ causes a drastic decline on SST-2, suggesting that "soft" guidance from the teacher model is not equivalent to "hard" supervision from ground-truth labels. Additionally, we observe that $\beta = 0.8$ yields promising results for both two tasks. Consequently, we keep $\beta = 0.8$ fixed and find the optimal α in {0.8, 0.9, 1.0} for each task.

Method	RTE (ACC)	CB (ACC)	Method	RTE (ACC)	CB (ACC)
$T0_{11B}$ (T)	89.1	100	$GPT_{1.3B}$ (T)	75.4	86.9
$T0_{3B}$ (S)	87.1	94.6	GPT_{125M} (S)	64.4	80.4
KL	87.4	94.6	KL	64.7	83.3
KL+RKL	87.8	96.4	KL+RKL	64.3	83.3
KL+JS	88.1	96.4	KL+JS	64.6	82.1
KL+SinKD	**89.9**	**98.2**	KL+SinKD	**65.0**	**84.5**

Table 8: Results of T0 Table 9: Results of GPT-on SuperGLUE. Neo on SuperGLUE.

5.4. Generalizability on Generative LLMs

To demonstrate the potential of our SinKD on generative LLMs, we perform distillation on various transformer architectures including the encoder-decoder T0 (Sanh et al., 2021) and the decoder-only GPT-Neo (Black et al., 2021). Specifically, $T0_{11B}$ and GPT-$Neo_{1.3B}$ serve as the teacher while $T0_{3B}$ and GPT-Neo_{125M} as the student. We validate SinKD on the SuperGLUE (Wang et al., 2019) benchmark against SOTA KD methods based on

1) KL divergence, 2) RKL divergence, and 3) JS divergence. We choose two datasets of RTE (Bentivogli et al., 2009) and CB (De Marneffe et al., 2019) for demonstrative experiment. Tab. 8 and Tab. 9 show that the proposed SinKD surpasses all other KD methods. Such findings showcase that SinKD can generalize to generative LLMs whose output logits are of high dimension equivalent to the size of the tokenizer vocabulary. Moreover, the performance gap between T0 and GPT-Neo can be ascribed to two reasons: 1) Architecture. The encoder-decoder architectures are generally more suitable for discriminative tasks compared with the decoder-only architectures since the former better comprehend the input-output relationships with bi-directional modeling. 2) Model scale. According to the scaling laws (Brown et al., 2020), the performance of GPT-Neo is expected to grow exponentially with billions of parameters increased. Under the limited GPU budget, experiments on larger decoder-only models are currently unavailable.

6. Conclusion

In this paper, we resort to the Sinkhorn distance for divergence measure and present the SinKD to address the limitations of existing distillation methods. Besides, we propose a batch-wise reformulation to capture geometric intricacies of distributions across samples in the high-dimensional space. Extensive experiments on the GLUE and SuperGLUE benchmarks confirm the superiority of our SinKD over SOTA methods from the aspect of comparability, validity, and generalizability.

A potential limitation is that we employ task formatting to adapt discriminative tasks under generative settings via prompts for experiments on GPT-Neo. The manual design of these prompts requires engineering experience and could significantly influence performance. Future work includes exploring application to representation-based KD and extension to other tasks (*e.g.*, document summarization, machine translation).

Broader Impact It is prospective to apply SinKD for distillation beyond the field of NLP. Its advantage in handling the "batchified" high-dimensional distributions would facilitate KD of the increasingly larger vision and language models for small-yet-competent ones with high cost-efficiency.

7. Acknowledgements

References

Martin Arjovsky and Leon Bottou. 2017. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. *TAC*, 7:8.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*, pages 1877–1901.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Pengfei Chen, Rongzhen Zhao, Tianjing He, Kongyuan Wei, and Qidong Yang. 2022. Unsupervised domain adaptation of bearing fault diagnosis based on join sliced wasserstein distance. *ISA transactions*, 129:504–519.

Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2018. Quora question pairs.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. 2017. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*.

Gongfan Fang, Yifan Bao, Jie Song, Xinchao Wang, Donglin Xie, Chengchao Shen, and Mingli Song. 2021. Mosaicking to distill: Knowledge distillation from out-of-domain data. *Advances in Neural Information Processing Systems*, 34:11920–11932.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. 2015a. Learning with a wasserstein loss. *Advances in neural information processing systems*, 28.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. 2015b. Learning with a wasserstein loss. *Advances in neural information processing systems*, 28.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. 2018. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR.

Jiawei Gu, Xuan Qian, Qian Zhang, Hongliang Zhang, and Fang Wu. 2023a. Unsupervised domain adaptation for covid-19 classification based on balanced slice wasserstein distance. *Computers in Biology and Medicine*, page 107207.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023b. Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30.

Shuncheng He, Yuhang Jiang, Hongchang Zhang, Jianzhun Shao, and Xiangyang Ji. 2022. Wasserstein unsupervised reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6884–6892.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174.

Xiao Jin, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie Yan, and Xiaolin Hu. 2019. Knowledge distillation via route constrained optimization. In *ICCV*, pages 1345–1354.

Sanjula Kammammettu and Zukui Li. 2023. Scenario reduction and scenario tree generation for stochastic programming using sinkhorn distance. *Computers & Chemical Engineering*, 170:108122.

Taehyeon Kim, Jaehoon Oh, Nak Yil Kim, Sangwook Cho, and Se-Young Yun. 2021. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 2628–2635.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Shijie Li, Inigo Jauregi Unanue, and Massimo Piccardi. 2023. Improving machine translation and summarization with the sinkhorn divergence. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 149–161. Springer.

Chang Liu, Chongyang Tao, Jiazhan Feng, and Dongyan Zhao. 2022. Multi-granularity structural knowledge distillation for language model compression. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1001–1011.

Yanbin Liu, Linchao Zhu, Xiaohan Wang, Makoto Yamada, and Yi Yang. 2023. Bilaterally normalized scale-consistent sinkhorn distance for few-shot image classification. *IEEE Transactions on Neural Networks and Learning Systems*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198.

Thong Thanh Nguyen and Anh Tuan Luu. 2022. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11103–11111.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. 2016. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29.

Dae Young Park, Moon-Hyun Cha, Daesin Kim, Bohyung Han, et al. 2021a. Learning student-friendly teacher networks for knowledge distillation. *Advances in neural information processing systems*, 34:13292–13303.

Geondo Park, Gyeongman Kim, and Eunho Yang. 2021b. Distilling linguistic context for language model compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 364–378.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976.

Gabriel Peyré, Marco Cuturi, et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid

Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multi-task prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Wenxian Shi, Yuxuan Song, Hao Zhou, Bohan Li, and Lei Li. 2020. Learning from deep model via exploring local targets.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4323–4332.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*.

Lifu Tu, Richard Yuanzhe Pang, Sam Wiseman, and Kevin Gimpel. 2020. Engine: Energy-based inference networks for non-autoregressive machine translation. *arXiv preprint arXiv:2005.00850*.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

SS Vallender. 1974a. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786.

SS Vallender. 1974b. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786.

Cédric Villani and Cédric Villani. 2009. The wasserstein distances. *Optimal Transport: Old and New*, pages 93–111.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. 2023. f-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10817–10834.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Siyue Wu, Hongzhan Chen, Xiaojun Quan, Qifan Wang, and Rui Wang. 2023. Ad-kd: Attribution-driven knowledge distillation for language model compression. *arXiv preprint arXiv:2305.10010*.

Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. 2020. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724.

Lantao Yu, Yang Song, Jiaming Song, and Stefano Ermon. 2020. Training deep energy-based models with f-divergence minimization. In *Interna-

tional Conference on Machine Learning, pages 10957–10967. PMLR.

Jianyi Zhang, Aashiq Muhamed, Aditya Anantharaman, Guoyin Wang, Changyou Chen, Kai Zhong, Qingjun Cui, Yi Xu, Belinda Zeng, Trishul Chilimbi, et al. 2023. Reaugkd: Retrieval-augmented knowledge distillation for pre-trained language models.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328.

Wangchunshu Zhou, Canwen Xu, and Julian McAuley. 2022. Bert learns to teach: Knowledge distillation with meta learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7037–7049.