# Neural Networks Are More Productive Teachers Than Human Raters: Active Mixup for Data-Efficient Knowledge Distillation from a Blackbox Model

Dongdong Wang [1*] Yandong Li [1*] Liqiang Wang[1] Boqing Gong[2]
[1]University of Central Florida [2]Google
{daniel.wang, liyandong}@Knights.ucf.edu lwang@cs.ucf.edu bgong@google.com

## Abstract

*We study how to train a student deep neural network for visual recognition by distilling knowledge from a blackbox teacher model in a data-efficient manner. Progress on this problem can significantly reduce the dependence on large-scale datasets for learning high-performing visual recognition models. There are two major challenges. One is that the number of queries into the teacher model should be minimized to save computational and/or financial costs. The other is that the number of images used for the knowledge distillation should be small; otherwise, it violates our expectation of reducing the dependence on large-scale datasets. To tackle these challenges, we propose an approach that blends mixup and active learning. The former effectively augments the few unlabeled images by a big pool of synthetic images sampled from the convex hull of the original images, and the latter actively chooses from the pool hard examples for the student neural network and query their labels from the teacher model. We validate our approach with extensive experiments.*

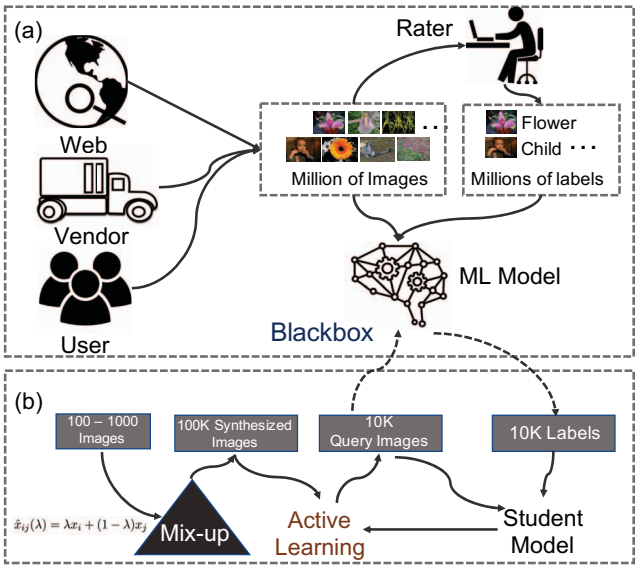


Figure 1. Data-efficient blackbox knowledge distillation. Given a blackbox teacher model and a small set of unlabeled images, we propose to employ mixup [49] and active learning [28] to train a high-performing student neural network in a data-efficient manner (b) so that we do not need to re-do the heavy and expensive data curation used to train the teacher model (a).

## 1. Introduction

Data curation is one of the most important steps for learning high-performing visual recognition models. However, it is often tedious and sometimes daunting to collect large-scale relevant data that have sufficient coverage of the inference-time scenarios. Additionally, labeling the collected data is time-consuming and costly.

Given a new task, how can we learn a high-quality machine learning model in a more data-efficient manner? We believe the answer varies depending on specific application scenarios. In this paper, we focus on the case that there exists a *blackbox* teacher model whose capability covers our task of interest. Indeed, there are many high-performing generic visual recognition models available as Web-based APIs, in our smart devices, or even as an obsolete model built by ourselves some while ago. The challenge is, however, we often have limited knowledge about their specifics, e.g., not knowing the exact network architecture or weights. Moreover, it could be computationally and/or financially expensive to query the models and read out their outputs for a large-scale dataset.

To this end, we study how to distill a *blackbox* teacher model for visual recognition into a student neural network in a data-efficient manner. Our objective is three-fold. First of all, we would like the distilled student network to perform well as the teacher model as possible at the inference time. Besides, we try to minimize the number of queries to the blackbox teacher model to save costs. Finally, we also shall use as a small number of examples as possible to save

* Equal contribution.

2575-7075/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPR42600.2020.00157

data collection efforts. It is hard to collect abundant data for rare classes or privacy-critical applications.

We propose to blend active learning [44, 28] and image mixup [49] to tackle the data-efficient knowledge distillation from a blackbox teacher model. The main idea is to synthesize a big pool of images from the few training examples by mixup and then use active learning to select from the pool the most helpful subset to query the teacher model. After reading out the teacher model's outputs, we simply treat them as the "groundtruth labels" of the query images and train the student neural network with them.

Image mixup [49, 13, 1] was originally proposed for data augmentation to improve the generalization performance of a neural recognition network. It synthesizes a virtual image by a convex combination of two training images. While the resultant image may become cluttered and semantically meaningless, it resides near the manifold of the natural images — unlike white-noise images. Given 1000 images, we can construct $O(10^5)$ pairs, each of which can further generate tens to thousands of virtual images depending on the coefficients in the convex combination. We conjecture that the big pool of mixup images provides good coverage of the manifold of natural images. Hence, we expect that a student network that imitates the blackbox teacher on the mixup images can give rise to similar predictions over the test images as the teacher model does.

Instead of querying the blackbox teacher model by all the mixup images, we resort to active learning to improve the querying efficiency. We first acquire the labels of the small number of original images from the blackbox teacher model and use them for training the student network. We then apply the *student* network to all the mixup images to identify the subset with which the current student network is the most uncertain. Notably, if two mixup images are synthesized from the same pair of original images, we keep only the one with higher uncertainty. We query labels for this subset, merge it into the previously labeled data, and then re-train the student network. We iterate this procedure of subset selection, querying the blackbox teacher model, and training the student neural network multiple times until reaching a stopping criterion.

To the best of our knowledge, we are the first to distill knowledge from a blackbox teacher model while underscoring the need for data-efficiency and query-efficiency. We empirically validate our approach by contrasting it to both vanilla and few/zero-shot knowledge distillation methods. Experiments show that, despite the blackbox teacher in our work, our approach performs on par or better than the competing methods that learn from whitebox teachers.

Note that the mixup images are often semantically meaningless, making them almost impossible for human raters to label. However, the blackbox teacher model returns predictions for them regardless, and the student network still gains from such fake image-label pairs. In this sense, we say that the blackbox teacher model is more productive than human raters in teaching the student network.

## 2. Related Work

**Knowledge Distillation.** Knowledge distillation is proposed in [16] to solve model compression problems, thus relieving the burden of ensemble learning. This work suggests that class probabilities, as "dark knowledge", are very useful to retain the performance of original network, and thus, light-weight substitute model could be trained to distill this knowledge. This approach is very useful and has been justified to solve a variety of complex application problems, such as pose estimation [37, 46, 33], lane detection [17], real-time streaming [31], object detection [6], video representation [41, 10, 11], and so forth. Furthermore, this approach is able to boost the performance of deep neural network with improvement on efficiency [35] and accuracy [25]. Accordingly, lots of research is conducted to enhance its performance from the perspective of training strategy [45, 20], distillation scheme [15, 4], or network properties [34] , etc.

However, there is an important issue. Traditional knowledge distillation requires lots of original training data which are very difficult to be obtained. To alleviate this data demand, few-shot knowledge distillation is proposed to retain teacher model performance with pseudo samplers which are generated in adversarial manner [21]. Another approach called data free knowledge distillation leverages extra activation records from teacher model to reconstruct original datasets, thus recovering teacher model [30]. Recently, a zero-knowledge distillation method is developed by synthesizing data with gradient information of teacher network [32]. Nevertheless, these approaches require the gradient information of teacher network, which enables them intractable in the real world.

**Blackbox Optimization.** Blackbox optimization is developed based on zero knowledge in the gradient information of queried models and widely used to solve practical problems. Recently, this work is widely used in deep learning, especially model attack. A rich line of blackbox attacking approaches [3, 18, 36, 2, 29] are explored by accessing the input-output pairs of classifiers, most of which are focusing on attacks resulting from accessing the data. [8] instead investigates that the adversaries are capable of recovering sensitive data by model inversion. However, there is no work for blackbox knowledge distillation.

**Active Learning.** Active learning is a learning process by interaction between oracle and learner agents. This strategy is widely used to solve learning problems which exhibit

costly data labelling since it could exploit existing data information to efficiently improve obtained model, thus reducing the number of queries. Lots of effective approaches are proposed to optimize this process, such as uncertainty-based [28, 48, 9] and margin-based methods [7, 38]. Form the review by [12], uncertainty-based methods, despite simple, are able to obtain good performance.

**Mixup.** Zhang *et al.* first proposed mixup to improve the generalization of deep neural network [49]. Between-Class learning [42] (BC learning) was proposed for deep sound recognition, and then, they extended this approach to image classification [43]. Following them, Pairing Samples [19] was proposed as a data augmentation approach by taking an average of two images for each pixel. More recently, an approach called AutoAugment [5], explores improving data augmentation policies by automatically searching.

## 3. Approach

We present our approach to the data-efficient knowledge distillation from a blackbox teacher model in detail in this section. Given a blackbox teacher model and a small number of unlabeled images, the approach iterates over the following three steps: 1) constructing a big candidate pool of synthesized images from the small number of unlabeled images, 2) actively choosing a subset from the pool with which the current student network is the most uncertain, 3) querying the blackbox teacher model to acquire labels for this subset and to re-train the student network.

### 3.1. Constructing a Candidate Pool

In real-world applications, data collection could consume a huge amount of time due to various reasons, such as privacy concerns, rare classes, data quality, etc. Instead of relying on a big dataset of real images, we begin with a small number of unlabeled images and use the recently proposed mixup [49] to augment this initial image pool.

Given two natural images $x_i$ and $x_j$, mixup generates multiple synthetic images by a convex combination of the two with different coefficients,

$$\hat{x}_{ij}(\lambda) = \lambda x_i + (1-\lambda)x_j, \tag{1}$$

where the coefficient $\lambda \in [0,1]$. Note that this notation also includes the original unlabeled data $x_i$ and $x_j$ when $\lambda = 1$ and $\lambda = 0$, respectively.

This technique comes handy and effective for our work. It can exponentially expand the size of the initial image pool. Suppose we have collected 1000 natural images, and we generate 10 mixup images for each image pair by varying the coefficient $\lambda$. We then arrive at a pool of about $10^6$ images in total. Besides, this pool of synthetic images also provides good coverage of the manifold of natural images. Indeed, this pool can be viewed as a dense sampling of the convex hull of the natural images we have collected. The test images likely fall into or close to this convex hull if the collected images are diverse and representative. Hence, we expect the student neural network to generalize well to the inference-time data by enforcing it to imitate the blackbox teacher model on the mixup images.

### 3.2. Actively Choosing a Subset to Query the Teacher Model

Let $\{\hat{x}_{ij}(\lambda), \lambda \in [0,1], i \neq j\}$ denote the augmented pool of images. It is straightforward to query the teacher model to obtain the (soft) labels for these synthetic images and then train the student network with them. However, this brute-force strategy incurs high computational and financial costs. Instead, we employ active learning to reduce the cost.

We define the student neural network's confidence over an input $x$ as

$$C_1(x) := \max_y P_S(y|x), \tag{2}$$

where $P_S(y|x)$ is the probability of the input image $x$ belonging to the class $y$ predicted by the current student network. Intuitively, the less confidence the student network has over the input $x$, the more the student network can gain from the teacher model's label for the input.

Therefore, we could rank all the synthetic images in the candidate pool according to the student network's confidences on them, and then choose the top ones as the query subset. However, this simple strategy results in near-duplicated images, for example $\hat{x}_{ij}(\lambda = 0.5)$ and $\hat{x}_{ij}(\lambda = 0.55)$. We avoid this situation by choosing at most one image from any pair of images.

In particular, instead of ranking the synthetic images, we rank image pairs in the candidate pool. We define the confidence of the student network over an image pair $x_i$ and $x_j$ as the following,

$$C_2(x_i, x_j) := \min_\lambda C_1(\hat{x}_{ij}(\lambda)), \quad \lambda \in [0,1], \tag{3}$$

which depends on a coefficient $\lambda^*$ for the image pair. Hence, we obtain a confidence score and its corresponding coefficient for any pair of the original images. The synthetic image $\hat{x}_{ij}(\lambda^*)$ is selected into the query set if the confidence score $C_2(x_i, x_j)$ is among the lowest $k$ ones. We study the size of the query set in the experiments.

### 3.3. Training the Student Network

With the actively selected query set of images, we query the blackbox teacher model and read out its soft predictions as the labels for the images. We then merge them with the previous training set, if there is, to train the student network

**Algorithm 1** Data-efficient blackbox knowledge distillation

**INPUT:** Pre-trained teacher model $\mathcal{M}^T$
**INPUT:** A small set of unlabeled images $X = \{x_i\}_{i=1}^n$
**INPUT:** Hyper-parameters (learning rate, subset size, etc.)
**OUTPUT:** Student network $\mathcal{M}^S$

1: Query $\mathcal{M}^T$ and acquire labels $Y_0$ for all images in $X$
2: Train an initial student network $\mathcal{M}_0^S$ with $(X, Y_0)$
3: Construct a synthetic image pool $\mathcal{P} = \{\hat{x}_{ij}(\lambda)\}$ by using the unlabeled images $X$ with eq. (1)
4: Initialize $\mathcal{P}_1^s = X, \mathcal{Y}_1 = \mathcal{Y}_0$.
5: **for** $t = 1, 2..., T$ **do**
6: Select a subset $\Delta\mathcal{P}_t^s$ from $\mathcal{P}$ with lowest confidence scores $\{C_2(x_i, x_j)\}$ returned by student $\mathcal{M}_{t-1}^S$
7: Query $\mathcal{M}^T$, acquire labels $\Delta\mathcal{Y}_t$ for all images $\Delta\mathcal{P}_t^s$
8: $\mathcal{P}_t^s \leftarrow \mathcal{P}_t^s \cup \Delta\mathcal{P}_t^s$, $\mathcal{Y}_t \leftarrow \mathcal{Y}_t \cup \Delta\mathcal{Y}_t$
9: Train a new student network $\mathcal{M}_t^S$ with $(\mathcal{P}_t^s, \mathcal{Y}_t)$
10: Update $\mathcal{P} \leftarrow \mathcal{P}$ - $\Delta\mathcal{P}_t^s$
11: **end for**

using a cross-entropy loss. The soft probabilistic labels returned by the teacher model give rise to slightly better results than the hard labels, so we shall use the soft labels in the experiments below.

### 3.4. Overall Algorithm

Algorithm 1 presents the overall procedure of our approach to the data-efficient blackbox knowledge distillation. Beginning with a teacher model $\mathcal{M}^T$ and a few unlabeled images $X = \{x_1, x_2, ..., x_n\}$, we firstly train an initial student network $\mathcal{M}_0^S$ with $(X, Y_0)$, where $Y_0$ contains the labels for the images in $X$ and is obtained by querying the teacher model. We then construct a big pool of synthetic images $\mathcal{P}$ with mixup [49] (eq. (1)) to facilitate the active learning stage. We iterate the following steps until the accuracy of the student network converges. 1) Actively select a subset $\Delta\mathcal{P}_t^s$ of the synthetic images $\mathcal{P}$ with the lowest confidence scores, $C_2(x_i, x_j)$, as predicted by the current student network so that the resulting subset $\Delta\mathcal{P}_t^s$ contains hard samples for the current student network $\mathcal{M}_{t-1}^S$. 2) Acquire labels $\Delta\mathcal{Y}_t$ of the selected subset of synthetic images $\Delta\mathcal{P}_t^s$ by querying the teacher model. 3) Train a new student network $\mathcal{M}_t^S$ with all the labeled images thus far, $(\mathcal{P}_t^s, \mathcal{Y}_t)$. Note that, in Line 6 of Algorithm 1, we only keep one synthetic image for any pair $(x_i, x_j)$ of the original images to reduce redundancy.

## 4. Experiments

We design various experiments to test our approach, including both comparison experiments with state-of-the-art knowledge distillation methods and ablation studies. Additionally, we also challenge our approach when the available data is out of the distribution of the main task of interest. In practice, across all experiments, we select $\lambda \in \{0.3, 0.7\}$ (with an interval of 0.04) to generate synthetic images to produce more diverse mixup images.

### 4.1. Comparison Experiments

Since our main objective is to explore how to train a high-performing student neural network from a blackbox teacher model in a data-efficient manner, it is worth comparing our approach with existing knowledge distillation methods although they were developed for other setups. The comparison can help review how data-efficient our approach is given the blackbox teacher model.

#### 4.1.1 Experiment Setting

**Datasets.** We run experiments on MNIST [26], Fashion-MNIST [47], CIFAR-10 [22], and Places365-Standard [50], which are popular benchmark datasets for image classification. The MNIST dataset contains 60K training images and 10K testing images about ten handwritten digits. The image resolution is 28×28. Fashion-MNIST is composed of 60K training and 10K testing fashion product images of the size 28×28. CIFAR-10 consists of 60K (50K training images and 10K test images) 32×32 RGB images in 10 classes, with 6K images per class. In addition to evaluating the proposed approach on the above described low-resolution images, we also test our approach on Places365-Standard, which is a challenging dataset for natural scene recognition. It has 1.8M training images and 18,250 validation images in 365 classes. We use the resolution of 256×256 for Places365-Standard in the following experiments.

**Evaluation Metric.** We mainly use the classification accuracy as the evaluation metric. Additionally, we also propose a straightforward metric to measure how much "knowledge" the student network distills from the teacher model. This metric is computed as the ratio between the student network's classification accuracy and the teacher's accuracy, and we call it the distillation *success rate*.

**Blackbox Teacher Models.** For each task except Places365-Standard, we prepare a teacher model by following the training setting provided in [32]. For Places365-Standard, there is no training setting reference for the knowledge distillation research yet, so we use a pre-trained model from the dataset repository [50] as our teacher model. On MNIST and Fashion-MNIST, we use the LeNet-5 architecture [27] as the teacher model and optimize it to achieve 99.29% and 90.80% top-1 accuracies, respectively. On CIFAR-10, we have an AlexNet [24] as the teacher model and train it to obtain 83.07% top-1 accuracy. As shown in Table 1, the above teacher

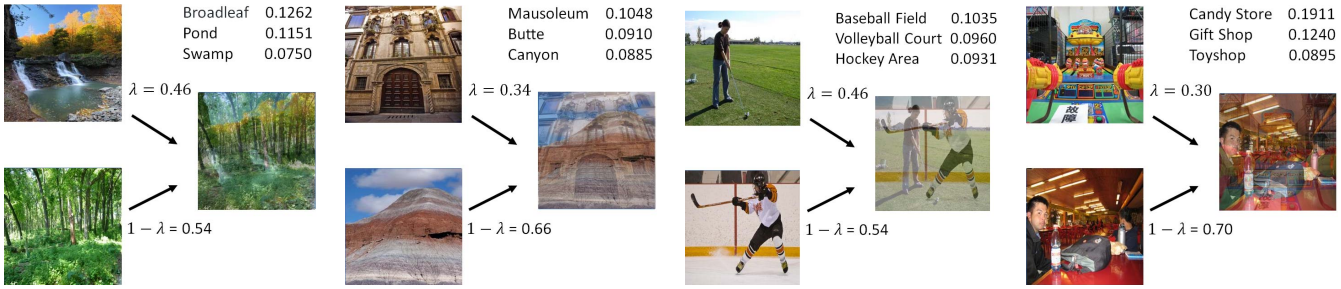


Figure 2. Mixup images whose confidence scores (cf. eq. (3)) are the lowest among all candidates in the third iteration. For each mixup image, we show the top three labels and probabilities returned by the blackbox teacher model.

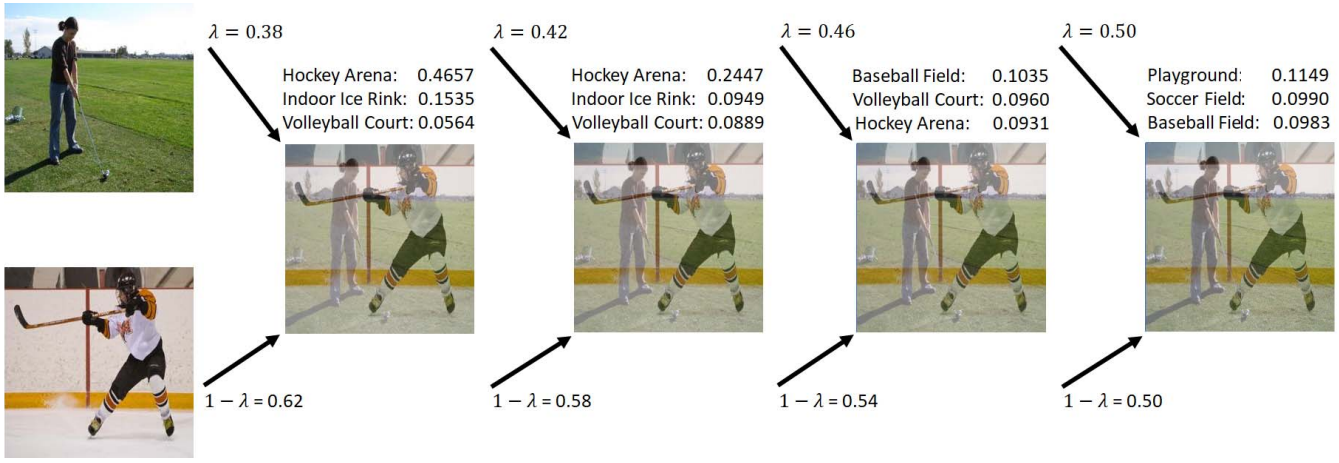


Figure 3. Different mixup images from the same pair of the original images by varying the mixup coefficient $\lambda$. We show the top three labels and probabilities predicted by the teacher model for each of them. It is interesting to see how the top-1 label changes from Hockey Arena, to Baseball Field, and to Golf Course.

models are comparable to the teacher models in [32]: 83.03% *vs.* 83.07% on CIFAR-10, 99.34% *vs.* 99.29% on MNIST, and 90.84% *vs.* 90.87% on Fashion-MNIST. For Places365-Standard, the teacher model is a ResNet-18 [14] and yields 53.68% top-1 accuracy.

**Competing Methods.** We identify three existing relevant methods for comparison.

- One is zero-shot knowledge distillation (ZSKD) [32], which distills a student neural network with zero training example from a *whitebox* teacher model. It synthesizes data by backpropagating gradients to the input through the whitebox teacher network.

- The second method is few-shot knowledge distillation (FSKD) [21], which augments the training images by generating adversarial examples. It is the most relevant work to ours, but it depends on the computationally expensive adversarial examples [40] and has no active learning scheme to reduce the query cost at all. The original work assumes a *whitebox* teacher neural network so that it is straightforward to produce the adversarial examples, whereas there exist blackbox attack methods [29, 3].

- The third is the vanilla knowledge distillation [16], which accesses the whole training set of the teacher model and is somehow an upper bound of our method.

#### 4.1.2 Quantitative Results

Table 1 shows the comparison results. For simplicity, we run the active learning stage for only one step (i.e., $T = 1$ in Algorithm 1). Section 4.2 presents the results of running it for multiple steps.

**Accuracy.** Our approach significantly outperforms FSKD over all the datasets. On CIFAR-10, MNIST, and Fashion-MNIST, ours yields 41%, 18%, and 14% success rate improvements over FSKD, respectively. On Places365-Standard, whose images are high-resolution about natural scenes, we also outperform FSKD by 14% success rate. Compared to ZSKD, which relies on a whitebox teacher network, our approach also shows higher accuracies and success rates except on MNIST. We were not able to re-

Table 1. Comparison results on Places365-Standard, CIFAR-10, MNIST, and Fashion-MNIST. The "Teacher" column reports the teacher model's accuracy on the test sets, "KD Accuracy" is the student network's test accuracy, "Success" stands for the distillation success rates, "Black/White" indicates whether or not the teacher model is blackbox, "Queries" lists the numbers of queries into the teacher models, and "Unlabeled Data" shows the numbers of original training images used in the experiments. (* results reported in the original paper)

| Task (Model) | Teacher | KD Accuracy | Success | Black/White | Queries | Unlabeled Data |
|---|---|---|---|---|---|---|
| Places365-Standard (ZSKD) [32] | – | – | – | – | – | 0 |
| Places365-Standard (FSKD [21]) | 53.69 | 38.18 | 71.11 | White | 480,000 | 80,000 |
| Places365-Standard (KD) | 53.69 | 49.01 | 90.35 | Black | 1,800,000 | 1,800,000 |
| Places365-Standard (Ours) | 53.69 | **45.71** | **85.14** | Black | 480,000 | 80,000 |
| CIFAR-10 (ZSKD) [32] | 83.03* | 69.56* | 83.78 | White | >2,000,000 | 0 |
| CIFAR-10 (FSKD [21]) | 83.07 | 40.58 | 48.85 | White | 40,000 | 2,000 |
| CIFAR-10 (KD) | 83.07 | 80.01 | 96.31 | Black | 50,000 | 50,000 |
| CIFAR-10 (Ours) | 83.07 | **74.60** | **89.87** | Black | 40,000 | 2,000 |
| MNIST (ZSKD) [32] | 99.34* | **98.77*** | 99.42 | White | >1,200,000 | 0 |
| MNIST (FSKD [21]) | 99.29 | 80.43 | 81.01 | White | 24,000 | 2,000 |
| MNIST (KD) | 99.29 | 99.05 | 99.76 | Black | 60,000 | 60,000 |
| MNIST (Ours) | 99.29 | 98.74 | **99.45** | Black | 24,000 | 2,000 |
| Fashion-MNIST(ZSKD) [32] | 90.84* | 79.62* | 87.65 | White | >2,400,000 | 0 |
| Fashion-MNIST (FSKD [21]) | 90.80 | 68.64 | 75.60 | White | 48,000 | 2,000 |
| Fashion-MNIST (KD) | 90.80 | 87.79 | 96.69 | Black | 60,000 | 60,000 |
| Fashion-MNIST(Ours) | 90.80 | **80.90** | **89.10** | Black | 48,000 | 2,000 |

produce ZSKD on Places365-Standard because its images are all high-resolution, making it computationally infeasible to generate a large number of gradient-based inputs. Similarly, the advantage of ours over ZSKD is larger on CIFAR-10 than other MNIST or Fashion-MNIST, probably because the CIFAR-10 images have a higher resolution. In contrast, the computation cost of our active mixup approach does not depend on the input resolution. Overall, the results indicate that active mixup has a higher potential to solve the larger-scale knowledge distillation in a data-efficient manner.

**Queries.** Our approach saves orders of queries into the teacher model compared to ZSKD. For example, we only query the blackbox teacher model up to 40K times for CIFAR-10. In contrast, ZSKD requires more than 2M queries and yet yields lower accuracy than ours. The big difference is not surprising because the gradient-based inputs in ZSKD are less natural than or representative of the test images than our mixup images. Besides, ZSKD incurs additional queries into the whitebox teacher model every time it produces an input.

#### 4.1.3 Qualitative Intermediate Results

We show some mixup images in Figures 2 and 3. These images are selected from the candidate pool constructed using the natural images in the Places365-Standard training set. Figure 2 shows some mixup images with low confidence scores. They can potentially benefit the student network more than the other candidate images if we use them to query the teacher model. Figure 3 demonstrates some mixup images synthesized from the same pair of natural images by varying the mixup coefficient $\lambda$. It is interesting to see that the mix of "Hockey Arena" and "Golf Course" leads to a "Baseball Field" at $\lambda = 0.46$ predicted by the blackbox teacher model. This indicates that our active mixup approach can effectively augment the originally small training set by not only bringing in new synthetic images but also comprehensive coverage of classes.

### 4.2. Ablation Study

We select CIFAR-10 and Places365-Standard to study our approach in detail since they represent the small-scale and large-scale settings, respectively. For CIFAR-10, we switch to VGG-16 [39] as the blackbox teacher model, which gives rise to 93.31% top-1 accuracy.

#### 4.2.1 Data-Efficiency and Query-Efficiency

We investigate how the results of our active mixup approach change as we vary the total number of unlabeled real images (data-efficiency) and the number of synthetic images selected by the active learning scheme (query-efficiency). Here we run only one step of the active learning stage ($T = 1$ in Algorithm 1) to save computation cost. Tables 2 and 3 show the results on CIFAR-10 and Places365-Standard, respectively. Each entry in the tables is a classification accuracy on the test set, and it is obtained by a

student network which we distill by using the corresponding number of unlabeled real images (Real images) and the number of selected synthetic images (Selected Syn.).

Table 2. Classification accuracy on CIFAR-10 with different numbers of real images and selected synthetic images.

| Selected Syn. \ Real images | 0.5K | 1K | 2K | 4K | 8K | 16K |
|---|---|---|---|---|---|---|
| 0 | 44.72 | 56.87 | 68.09 | 76.59 | 83.61 | 86.89 |
| 5K | 66.97 | 71.67 | 77.76 | 81.76 | 85.76 | 87.05 |
| 10K | 73.60 | 77.27 | 81.27 | 83.27 | 86.56 | 88.79 |
| 20K | 77.44 | 81.18 | 84.19 | 86.29 | 88.07 | 89.01 |
| 40K | 82.28 | 84.25 | 86.06 | 87.71 | 89.00 | 90.49 |
| 80K | 85.18 | 86.53 | 87.89 | 88.71 | 89.61 | 90.96 |
| 160K | 86.56 | 88.94 | 89.42 | 90.26 | 90.87 | 91.51 |

Table 3. Classification accuracy on Places365-Standard with different numbers of real images and selected synthetic images.

| Selected Syn. \ Real images | 20K | 40K | 80K |
|---|---|---|---|
| 100K | 40.72 | 41.95 | 43.52 |
| 200K | 41.15 | 42.86 | 44.77 |
| 400K | 41.94 | 43.42 | 45.71 |

We can see that the more synthetic images we select by their confidence scores (cf. eq. (3)), the higher-quality the distilled student network is. It indicates that the mixup images can effectively boost the performance of our method. Meanwhile, the higher the number of unlabeled real images we have, the higher the distillation success rate we can achieve. What's more interesting is that, when the number of synthetic images is high (e.g., 160K), the gain is diminishing as we increase the number of real images. Hence, depending on the application scenarios, we have the flexibility to trade-off the real images and synthetic images for achieving a certain distillation success rate.

We can take a closer look at Tables 2 and 3 to obtain an understanding about the "market values" of the selected synthetic images. In Table 2, 10K selected synthetic images and 8K unlabeled real images yield 86.56% accuracy; 20K synthetic images and 4K real images lead to 86.29% accuracy; and 40K synthetic images with 2K real examples give rise to 86.06% accuracy. The accuracies are about the same. There is a similar trend along the off-diagonal entries in Table 3, implying that if we reduce the number of real images by half, we can complement it by doubling the size of synthetic images to maintain about the same distillation success rate.

#### 4.2.2 Active Mixup vs. Random Search

We design another experiment to compare active mixup with the random search to understand the effectiveness of our active learning scheme. We keep 500 real images for CIFAR-10 and 20K for Places365-Standard. We then use them to construct 100K and 300K synthetic images, respectively. For a fair comparison, we let random search and active mixup share the same sets of natural images. Since our active learning scheme avoids selecting redundant images by using the improved confidence score in eq. (3), we also equip the random search such capability by using a single mixup coefficient of $\lambda = 0.5$ to construct the synthetic images. This guarantees that, like our approach, no two synthetic images selected by the random search are from the same pair of real images.

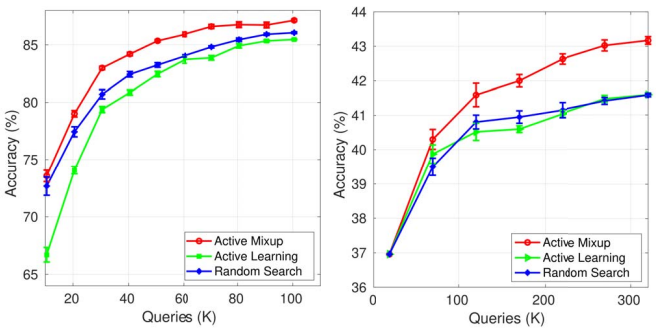


Figure 4. Test accuracy of student networks vs. number of queries into the blackbox teacher model on CIFAR-10 (left) and Places365-Standard (right). We use 500 and 20K natural images for the two datasets, respectively. The plot for CIFAR-10 starts from first active learning stage ($t = 1$ in Algorithm 1) and the one for Places365 starts from the initial student network training by natural images. The initial student network for CIFAR-10 trained by using natural images only yields 43.67% accuracy.

Figure 4 shows the comparison results of our active mixup and the random search. On CIFAR-10, we select 10K synthetic images every time and run the active learning stage for 10 steps ($T = 10$ in Algorithm 1). On Places365-Standard, we run it for six steps and choose 50K synthetic images per step. We can see that active mixup significantly outperforms random search over the whole course of knowledge distillation, verifying its effectiveness on improving the query-efficiency. More concretely, 80K actively selected synthetic images yield 86.76% accuracy, which is about the same as what 160K randomly selected synthetic images can achieve on CIFAR-10. Similarly, 40K synthetic images by active mixup lead to 84.2% accuracy, on par with the 85.18% accuracy by 80K randomly chosen synthetic images.

#### 4.2.3 Active Mixup vs. Vanilla Active Learning

Our active learning scheme (eq. (3)) improves upon the vanilla score-based active learning (eq. (2)) by selecting only one synthetic image at most from any pair of real images. This change is necessary because two nearly dupli-

cated synthetic images could both have very low scores according to eq. (2).

To quantitatively compare the two active learning methods, we run another experiment by replacing our active learning scheme with the vanilla version. The candidate pool is the same as ours, i.e., mixup images generated by varying $\lambda \in \{0.3, 0.7\}$ with an interval of 0.04. Figure 4 shows the results on both CIFAR-10 and Place365-Standard.

Generally, the vanilla active learning yields lower accuracy than our active mixup and the random search. This shows that the vanilla score-based active learning even fails to improve upon random search because it selects nearly duplicated synthetic images to query the teacher model. In contrast, our active mixup consistently performs the better than the vanilla active learning and random search. The prominent gap justifies that the constraint by $C_2$ in eq. (3) is crucial in our approach.

### 4.3. Active Mixup with Out-of-Domain Data for Blackbox Knowledge Distillation

In real-world applications, it may be hard to collect real training images for some tasks, e.g., due to privacy concerns. Under such scenarios, we have to use out-of-domain data to distill the student neural network. Hence, we further challenge our approach by revealing some images that are out of the domain of the training images of the blackbox teacher model.

We conduct this experiment on CIFAR-10 by providing our approach some training images in CIFAR-100 [23]. To reduce information leak, we exclude the images that belong to the CIFAR-10 classes and keep 2K images to construct the candidate pool. Equipped with these synthetic images, we run active mixup to distill student neural networks from a blackbox teacher model for CIFAR-10. The teacher model is VGG-16, which yields 93.31% accuracy on the CIFAR-10 test set.

Table 4. CIFAR-10 classification accuracy by the student neural networks which are distilled by using out-of-domain data.

| Selected Syn. | 10K | 20K | 40K | 80K |
|---|---|---|---|---|
| Accuracy (%) | 64.10 | 71.39 | 77.89 | 83.03 |

Table 4 shows the results of different numbers of selected synthetic images. We still run only one iteration of the active learning to save computation costs. The best distillation performance is 83% top-1 accuracy and success rate is 88.9%. Comparing the result to Table 2, especially the entry (87.89%) of 80K selected synthetic images and 2K real images, we can see that our approach leads to about the same performance by using the out-of-domain data as the in-domain data.

Table 5. CIFAR-10 classification accuracy by the student neural networks which are distilled by using out-of-domain data. We set the number of selected synthetic images to 40K and vary the numbers of real images.

| Real images | 500 | 1000 | 1500 | 2000 |
|---|---|---|---|---|
| Accuracy (%) | 70.21 | 74.60 | 75.54 | 77.89 |

To better understand how different factors influence the distillation performance, we also decouple the number of available real images from the number of selected synthetic images in Table 5. We fix the number of selected synthetic images to 40K and vary the numbers of real images. Not surprisingly, the more real images there are, the higher distillation accuracy the active mixup achieves. Furthermore, the number of synthetic images still plays a prominent role in distillation accuracy, according to Table 4. Without the original training data, mixup augmentation is probably more critical to enhancing the distillation performance than otherwise.

## 5. Discussion and Conclusion

In this paper, we formalize a novel problem, knowledge distillation from a blackbox teacher model in a data-efficient manner, which we think is more realistic than previous knowledge distillation setups. There are two key challenges to this problem. One is that the available examples are insufficient to represent the vast variation in the original training set of the teacher model. The other is that the blackbox teacher model often implies that it is financially and computationally expensive to query.

To deal with the two challenges, we propose an approach combining mixup and active learning. Although neither of them is new by itself, combining them is probably the most organic solution to our problem setup for the following reasons. First of all, we would like to augment the few available examples. Unlike conventional data augmentations (e.g., cropping, adding noise), which only probe the regions around the available examples, mixup provides a continuous interpolation between any pairwise examples. As a result, mixup allows the student model to probe diverse regions of the input space. We then employ active learning to reduce the query transactions to the teacher model. Extensive experiments verify the effectiveness of our approach to the data-efficient blackbox knowledge distillation.

## 6. Acknowledgements

This work was supported in part by NSF-1741431 and NSF-1836881.

## References


[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. 2

[2] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017. 2

[3] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26. ACM, 2017. 2, 5

[4] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4794–4802, 2019. 2

[5] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 3

[6] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7023–7032, 2019. 2

[7] Melanie Ducoffe and Frederic Precioso. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*, 2018. 3

[8] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333. ACM, 2015. 2

[9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1183–1192. JMLR. org, 2017. 3

[10] Chuang Gan, Boqing Gong, Kun Liu, Hao Su, and Leonidas J Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5589–5597, 2018. 2

[11] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7053–7062, 2019. 2

[12] Daniel Gissin and Shai Shalev-Shwartz. Discriminative active learning. *arXiv preprint arXiv:1907.06347*, 2019. 3

[13] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3714–3722, 2019. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[15] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. *arXiv preprint arXiv:1904.01866*, 2019. 2

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 5

[17] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning lightweight lane detection cnns by self attention distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1013–1021, 2019. 2

[18] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. *arXiv preprint arXiv:1804.08598*, 2018. 2

[19] Hiroshi Inoue. Data augmentation by pairing samples for images classification. *arXiv preprint arXiv:1801.02929*, 2018. 3

[20] Xiao Jin, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie Yan, and Xiaolin Hu. Knowledge distillation via route constrained optimization. *arXiv preprint arXiv:1904.09149*, 2019. 2

[21] Akisato Kimura, Zoubin Ghahramani, Koh Takeuchi, Tomoharu Iwata, and Naonori Ueda. Few-shot learning of neural networks from scratch by pseudo example optimization. *arXiv preprint arXiv:1802.03039*, 2018. 2, 5, 6

[22] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images, 2009. 4

[23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 8

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 4

[25] Jogendra Nath Kundu, Nishank Lakkakula, and R Venkatesh Babu. Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1436–1445, 2019. 2

[26] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 4

[27] Yann LeCun et al. Lenet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet*, 20:5, 2015. 4

[28] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR94*, pages 3–12. Springer, 1994. 1, 2, 3

[29] Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. *arXiv preprint arXiv:1905.00441*, 2019. 2, 5

[30] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*, 2017. 2

[31] Ravi Teja Mullapudi, Steven Chen, Keyi Zhang, Deva Ramanan, and Kayvon Fatahalian. Online model distillation for efficient video inference. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3573–3582, 2019. 2

[32] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. In *International Conference on Machine Learning*, pages 4743–4751, 2019. 2, 4, 5, 6

[33] Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, and Jiashi Feng. Dynamic kernel distillation for efficient pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6942–6950, 2019. 2

[34] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5007–5016, 2019. 2

[35] Mary Phuong and Christoph H Lampert. Distillation-based training for multi-exit architectures. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1355–1364, 2019. 2

[36] Tim Salimans, Jonathan Ho, Xi Chen, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017. 2

[37] Muhamad Risqi U Saputra, Pedro PB de Gusmao, Yasin Almalioglu, Andrew Markham, and Niki Trigoni. Distilling knowledge from a deep pose regressor network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 263–272, 2019. 2

[38] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in neural information processing systems*, pages 1289–1296, 2008. 3

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[40] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 5

[41] Mohammad Tavakolian, Hamed R Tavakoli, and Abdenour Hadid. Awsd: Adaptive weighted spatiotemporal distillation for video representation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8020–8029, 2019. 2

[42] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Learning from between-class examples for deep sound recognition. *arXiv preprint arXiv:1711.10282*, 2017. 3

[43] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Between-class learning for image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[44] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001. 2

[45] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1365–1374, 2019. 2

[46] Chaoyang Wang, Chen Kong, and Simon Lucey. Distill knowledge from nrsfm for weakly supervised 3d pose learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 743–752, 2019. 2

[47] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 4

[48] Yazhou Yang and Marco Loog. A benchmark and comparison of active learning for logistic regression. *Pattern Recognition*, 83:401–415, 2018. 3

[49] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 1, 2, 3, 4

[50] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2017. 4