Knowledge Distillation in Deep Networks under a Constrained Query Budget

Ankita Singh and Shayok Chakraborty

Department of Computer Science, Florida State University, USA

Abstract. Knowledge distillation addresses the problem of training a lightweight model (student) from a deeper, more complex model (teacher) so as to mimic its performance. Existing techniques mostly utilize the predictions furnished by the teacher on a given training set to perform the distillation and train the student. However, querying the teacher model for labels can be an expensive process in terms of computational/ financial overhead. In this paper, we tackle the problem of distilling knowledge from a blackbox teacher model into a student deep neural network, in a cost-efficient manner. Active learning algorithms automatically identify the salient and exemplar samples from large amounts of unlabeled data and are instrumental in reducing human annotation effort in inducing a machine learning model. We propose a novel active learning algorithm using which the student model can identify the most informative samples from a large amount of unlabeled data, which need to be queried from the teacher. We exploit the geometry of the unlabeled data to identify a batch of representative samples which can reconstruct the data with minimal error. We pose the sample selection as an NP-hard optimization problem and solve it efficiently using an iterative algorithm, with global convergence. Such an algorithm can be effective in distilling relevant knowledge from the teacher to the student under a constrained query budget. Our extensive empirical studies on five challenging datasets from two application domains (computer vision and text mining) corroborate the efficacy of our active sampling framework over competing baselines.

Keywords: Knowledge distillation · Deep learning · Subset selection

1 Introduction

Knowledge distillation (KD) is a method for transferring complex mapping functions learned by a high-capacity model or an ensemble of multiple models (the teacher) to a relatively simpler, lightweight model (the student) [9, 11]. Generally, the teacher models deliver good generalization performance; however, they have a high memory footprint and are computationally expensive. The student models, on the other hand, require much less memory and computation and are thus more suitable for real-time applications. KD has been used in a variety of applications, such as pose estimation [25], object detection [6] and video representation [7] among others.

In any KD application, the objective is to train the student network to imitate the teacher network, using a given training set. The transfer of knowledge from the teacher to the student is typically facilitated using a variety of methods, such as matching the soft-label probabilities, the l_2 norm between the feature representations or the attention maps, and the maximum mean discrepancy (MMD) between the distributions of the neuron selectivity patterns learned by the teacher and the student networks [11, 15, 29] among others. In all these methods, all the training examples are required to be passed to the teacher, and its outputs are used to compute a distillation loss and train the student. However, in certain applications, querying the teacher model can be expensive computationally and/or financially. For instance, the teachers are often models that are trained and hosted by companies on the cloud, commonly referred to as Machine Learning as a Service (MLaaS) platforms. Third-party developers access these models through Application Programming Interfaces (APIs). Each access to the API incurs a cost, which means that a price needs to be paid each time the teacher model is queried. In such applications, obtaining the teacher's output on all the training samples in order to train the student, can be prohibitive. This necessitates the development of an algorithm to distill knowledge from the teacher to the student network, when the number of label queries to the teacher cannot exceed a pre-specified budget.

We formally pose the research question as follows: *We are given a blackbox teacher model (deep neural network) trained on a given application of interest. The data used to train the teacher is not available; the specific architecture and trained parameters of the teacher are also not known. We are interested to train a student deep neural network using a knowledge distillation algorithm to imitate the teacher. For this purpose, we are given a small amount of labeled data L, and a large amount of unlabeled data U, with $|L| \ll |U|$. However, we are not allowed to query the labels of all the $|U|$ samples from the teacher due to computational and/or financial cost constraints. We are further given a query budget k which denotes the number of unlabeled samples whose labels can be queried from the teacher. Which k samples should we select for query so that the student model's generalization accuracy gets closest to that of the teacher?*

Active learning (AL) is a machine learning paradigm to automatically identify the most informative samples from large amounts of unlabeled data [34]. This tremendously reduces human annotation effort in inducing a model, as only the few samples that are selected by the algorithm, need to be labeled manually. Further, since the model gets trained on the exemplar data samples, it typically depicts better generalization accuracy than a passive learner, where the training data is sampled at random. AL has been successfully used in a variety of applications, such as computer vision [43], text analytics [38], computational biology [26], email classification [32] etc.

In this paper, we propose a novel AL algorithm to address the aforementioned research question. We pose the sample selection as a constrained NP-hard optimization problem (based on the data reconstruction error) and derive an iterative algorithm, with global convergence, to solve it. Our framework is

easy to implement and independent of the underlying KD algorithm, as well as the architectures of the teacher and student networks. The proposed algorithm is generic and can be used in any application to select an informative subset of samples from large amounts of data; we validate it on the KD application in this paper, as research in this area is still in a nascent stage.

2 Related Work

Active Learning: With the advent and popularity of deep neural networks, deep active learning (DAL) has attracted significant research attention, where the objective is to identify the salient unlabeled samples for manual annotation and simultaneously learn discriminating feature representations using a deep neural network [28]. Recently proposed DAL strategies include learning a task-agnostic loss function to identify the informative unlabeled samples [43], finding a core set of samples such that the deep model trained on this subset is competitive over the whole dataset [33], a method based on diverse gradient embeddings (BADGE) which combines uncertainty and diversity for active sample selection [2], a discriminative algorithm that selects samples such that the labeled and the unlabeled sets are maximally similar [8] and methods based on adversarial learning [36, 48]. Other related research in AL includes active learning in the presence of noisy annotators [14], actively completing an incomplete data matrix [30], combining active learning with transfer learning [37], actively selecting the informative features and samples [20] and AL with novel annotation mechanisms [12] among others.

Knowledge Distillation: Knowledge distillation has received increasing attention from the research community in recent years; please refer [9] for a comprehensive survey. A simple and effective idea to transfer knowledge is to match the responses [11], learned feature representations [29] or relationships between different layers [19] between the student and the teacher models. Metrics such as the KL-divergence, Maximum Mean Discrepancy (MMD), l_2 and l_1 norm distance are commonly used to compute the similarity and formulate the distillation loss terms to train the student network. Several distillation techniques have been explored to improve the transfer of knowledge from the teacher to the student in more complex settings, including adversarial distillation [22], multi-teacher distillation [45] and graph-based distillation [5] among others. Recently, a body of research has focused on reducing the amount of training data required for effective transfer of knowledge. Few-shot KD has been proposed to retain the teacher model's performance with pseudo samplers which are generated in an adversarial manner [16]. Zero-shot KD has also been explored by generating data using the gradient information of the teacher network [23]. However, these methods require the gradient information of the teacher network, which is difficult to obtain in real-world applications.

Active Learning for Knowledge Distillation: Even though both AL and KD have been extensively studied, AL for KD is in its nascent stage and has only been explored in recent years. As in conventional AL, uncertainty sampling has

been exploited to actively select unlabeled samples to train the student model [3, 27, 46]. Very recently, researchers have begun to study the performance of deep active learning algorithms such as BADGE and Coreset for KD [13, 18]. Wang *et al.* [39] proposed mixup together with active learning to augment the unlabeled pool with synthetic data samples, and then query the labels of the hard examples from the teacher to train the student. However, as noted by the authors, mixup may produce data samples that are semantically meaningless, and the knowledge gained by the student from such fake (sample-label) pairs may not be substantial. In contrast, we propose a method to identify the informative unlabeled samples to train the student without generating synthetic / fake data samples. We now describe our framework.

3 Proposed Framework

3.1 Problem Setup

In our problem setup, the student model is given a labeled set L and an unlabeled set U, where $|L| \ll |U|$. Let n be the number of unlabeled samples, where each sample is represented using a vector of d dimensions. Let $X \in \Re^{d \times n}$ denote the unlabeled data matrix, where each column represents a sample and each row represents a feature. Our objective is to select k samples from U to distill knowledge from the teacher and train the student model. Our method is motivated by research in transductive experimental design [35, 44], which attempts to select a representative subset such that the whole dataset can be approximated by a linear combination of the selected samples. We formulate the active sampling problem based on data geometry and attempt to select k samples using which the unlabeled data can be reconstructed with minimal error.

3.2 Active Sample Selection

Let $z \in \{0,1\}^{n \times 1}$ be a binary selection vector where $z_i = 1$ if unlabeled sample x_i is selected in the batch and $z_i = 0$ otherwise; let $diag(z)$ be a diagonal matrix with z along the main diagonal. We pose the sample selection as minimizing the following residual:

$$\min_{z,\widehat{C}} \left\| X - X diag(z) \widehat{C} \right\|_F^2 \tag{1}$$

$$\text{s.t.:} \quad z \in \{0,1\}^{n \times 1}, \qquad \sum_{i=1}^{n} z_i = k$$

where $||.||_F$ denotes the matrix Frobenius norm. The term $Xdiag(z)$ attempts to retain k columns (data samples) in the matrix X and set the remaining $(n-k)$ columns to 0; these k samples therefore denote the most representative samples to reconstruct the unlabeled data matrix X. $\widehat{C} \in \Re^{n \times n}$ is a matrix

of reconstruction coefficients. We decompose $\widehat{C} = CX$ where $C \in \Re^{n \times d}$ and express the problem as:

$$\min_{z,C} \|X - Xdiag(z)CX\|_F^2 \tag{2}$$
$$\text{s.t.:} \quad z \in \{0,1\}^{n \times 1}, \qquad \sum_{i=1}^{n} z_i = k$$

This can be written equivalently as:

$$\min_{Q} \|X - XQX\|_F^2 \tag{3}$$
$$\text{s.t.:} \quad \|Q\|_{2,0} = k$$

where $Q = diag(z).C$ and $\|Q\|_{2,0}$ denotes the $l_{2,0}$ norm of a matrix, that is, the number of non-zero rows in the matrix Q. To see the equivalence between Equations (2) and (3), we note that, if a particular row of $diag(z)$ has all 0s, that row in Q will also have all 0s. Hence, the number of non-zero rows in Q is equal to the number of non-zero entries in $diag(z)$, that is, $\|Q\|_{2,0} = k$. Based on this, we propose to optimize the following objective function:

$$\min_{Q} \|X - XQX\|_F^2 + \alpha \|Q\|_{2,0} \tag{4}$$

where $\alpha \geq 0$ is a regularization parameter. Once we solve for Q, we can compute the l_2 norm of each row of Q and select the k unlabeled samples corresponding to the k highest l_2 norm values. However, this is an NP-hard problem due to the matrix $l_{2,0}$ norm. Nie *et al.* [24] established that the $l_{2,1}$ norm of a matrix is the minimum convex hull of the $l_{2,0}$ norm, and minimizing the $l_{2,1}$ norm is equivalent to minimizing the $l_{2,0}$ norm, as long as the matrix is row-sparse. With this assumption, we can relax (4) into the following convex optimization problem:

$$\min_{Q} \|X - XQX\|_F^2 + \alpha \|Q\|_{2,1} \tag{5}$$

where $\|Q\|_{2,1}$ is the matrix $l_{2,1}$ norm, which is the sum of the l_2 norm of each row of a matrix. Our objective function contains the non-smooth term $\alpha \|Q\|_{2,1}$, which makes it challenging to guarantee an optimal solution by directly differentiating the objective. We employ the alternating direction method of multipliers (ADMM) to solve this problem [4]. We introduce a new variable $\widehat{Q}$ and express the problem as:

$$\min_{Q,\widehat{Q}} \|X - XQX\|_F^2 + \alpha \left\|\widehat{Q}\right\|_{2,1}$$
$$\text{s.t.:} \quad Q = \widehat{Q} \tag{6}$$

The augmented Lagrangian function can be written as:

$$L(Q, \widehat{Q}, \lambda, \theta) = \|X - XQX\|_F^2 + \alpha \left\|\widehat{Q}\right\|_{2,1} + \langle \lambda, Q - \widehat{Q} \rangle + \frac{\theta}{2} \left\|Q - \widehat{Q}\right\|_F^2 \tag{7}$$

where $\lambda \in \Re^{n \times d}$ is the matrix of Lagrangian multipliers, $\theta \in \Re^{1 \times 1}$ is a constraint violation penalty parameter and $\langle ., . \rangle$ denotes the matrix inner product operator.

Updating Q: Considering the terms with Q in Equation (7), we have the objective function to be minimized as:

$$L_Q = \|X - XQX\|_F^2 + \frac{\theta}{2} \left\| Q - \widehat{Q} + \frac{\lambda}{\theta} \right\|_F^2 \tag{8}$$

Setting $\frac{\partial L_Q}{\partial Q} = 0$, we get

$$2X^\top XQXX^\top + \theta Q = 2X^\top XX^\top + \theta \left(\widehat{Q} - \frac{\lambda}{\theta} \right) \tag{9}$$

Let $A = 2X^\top X$ and $B = 2X^\top XX^\top + \theta \left(\widehat{Q} - \frac{\lambda}{\theta} \right)$. Plugging back in Equation (9) we get:

$$AQXX^\top + \theta Q = B \tag{10}$$

Note that both A and $XX^\top$ are symmetric and positive-semidefinite matrices. We can therefore perform an eigen decomposition of both these matrices as follows:

$$A = U \Sigma_1 U^\top, \quad XX^\top = V \Sigma_2 V^\top \tag{11}$$

where U and V are orthogonal matrices, Σ_1, Σ_2 are diagonal matrices. Plugging this back in Equation (10) we get:

$$U \Sigma_1 U^\top QV \Sigma_2 V^\top + \theta Q = B \tag{12}$$

Multiplying both sides by $U^\top$ from left and V from right, we get:

$$\Sigma_1 U^\top QV \Sigma_2 + \theta U^\top QV = U^\top BV \tag{13}$$

Let $D = U^\top QV$. Plugging this back in Equation (13) we get:

$$\Sigma_1 D \Sigma_2 + \theta D = U^\top BV \tag{14}$$

Equating both sides element by element, we get:

$$D_{ij} = \frac{(U^\top BV)_{ij}}{(\Sigma_1)_{ii} . (\Sigma_2)_{jj} + \theta}, \quad i = 1 \ldots n, j = 1 \ldots d \tag{15}$$

Now, $D = U^\top QV$. Thus, we can solve Q as:

$$Q = UDV^\top \tag{16}$$

Updating $\widehat{Q}$: From Equation (7), considering the terms with $\widehat{Q}$, we have the objective function to be minimized as:

$$L_{\widehat{Q}} = \alpha \left\| \widehat{Q} \right\|_{2,1} + \frac{\theta}{2} \left\| Q - \widehat{Q} + \frac{\lambda}{\theta} \right\|_F^2 \tag{17}$$

Since the $l_{2,1}$ norm is the sum of the l_2 norms of each row of a matrix, we can decouple the minimization problem and solve for the matrix $\widehat{Q}$ row by row. The following lemma can be used to solve the above optimization problem [42].

Lemma 1. *For any $\kappa, \mu > 0$ and $g \in \Re^{n \times 1}$, the minimizer of*

$$\min_{t \in \Re^{n \times 1}} \kappa ||t||_2 + \frac{\mu}{2} ||t - g||_2^2$$

is given by

$$t = \begin{cases} \left(1 - \frac{\kappa}{\mu ||g||_2}\right) g & \text{if } ||g||_2 > \frac{\kappa}{\mu} \\ 0 & \text{if } ||g||_2 \leq \frac{\kappa}{\mu} \end{cases}$$

The solution to (17) is thus obtained as:

$$\widehat{Q}^i = \begin{cases} \left(1 - \frac{\alpha}{\theta ||s||_2}\right) s & \text{if } ||s||_2 > \frac{\alpha}{\theta} \\ 0 & \text{if } ||s||_2 \leq \frac{\alpha}{\theta} \end{cases} \tag{18}$$

where $s = \left(Q + \frac{\lambda}{\theta}\right)^i$, for $i = 1 \ldots n$ and M^i denotes the i^{th} row of matrix M.

Updating λ: The matrix λ can be updated using the following equation [4]:

$$\lambda \leftarrow \lambda + \theta(Q - \widehat{Q}) \tag{19}$$

The pseudo-code of our framework is outlined in Algorithm 1. As evident from the pseudo-code, our algorithm is independent of the underlying KD algorithm and the teacher-student network architectures, and can thus be seamlessly integrated across different teacher-student architectures and different applications. It is also very easy to implement.

Algorithm 1 The proposed active sample selection algorithm

Require: Unlabeled data matrix $X \in \Re^{d \times n}$, AL batch size k, parameters α, θ

1: **Initialize:** $Q = \widehat{Q} = \lambda = \{0\}^{n \times d}$
2: **repeat**
3: Compute the matrices A and B, as shown in Equation (10)
4: Perform eigen-decomposition and compute the matrices U, V, Σ_1 and Σ_2, as shown in Equation (11)
5: Compute the matrix D element by element, as shown in Equation (15)
6: Update the matrix Q, as shown in Equation (16)
7: Update the matrix $\widehat{Q}$ row by row using Equation (18)
8: Update the matrix λ using Equation (19)
9: **until** Convergence
10: Compute the l_2 norm of each row of the matrix Q. Identify the k rows with the highest l_2 norms and select the corresponding k unlabeled samples in the batch

3.3 Convergence Analysis

As evident from Algorithm 1, the sub-problems corresponding to $Q, \widehat{Q}$ and λ have closed form solutions. The convergence of Algorithm 1 can be obtained from the ADMM convergence results established in [4, 10], which is formalized in the following theorem:

Theorem 1. *For given parameters α and θ, the iterates $(Q^i, \widehat{Q^i}, \lambda^i)$ converge to the solution $(Q^\star, \widehat{Q^\star}, \lambda^\star)$ where $(Q^\star, \widehat{Q^\star})$ is the global optimal solution of Problem (6).*

Please refer [4, 10] for detailed proof.

3.4 Using Labeled Data for Active Sampling

Depending on the size of the initial training set L, it maybe desirable to use the uncertainty of the student model trained on L to select unlabeled samples from U, together with the method proposed in Algorithm 1. To this end, we compute an uncertainty vector $e \in \Re^{n \times 1}$ containing the prediction entropy of the student on all the unlabeled samples. Also, let q be the vector containing the l_2 norm of each row of the matrix Q, as detailed in Algorithm 1. We compute a weighted summation of these two vectors as follows:

$$v = \beta.q + (1 - \beta).e \tag{20}$$

where $0 \leq \beta \leq 1$ is a weight parameter governing the relative importance of the two terms. The k largest entries in the vector v are used to select the unlabeled samples in the batch. Note that our active sampling algorithm still remains independent of the network architecture and the KD algorithm, since the entropy vector can be computed merely from the probability values furnished by the student on the samples in U.

4 Experiments and Results

Datasets: Since knowledge distillation has been most extensively used in computer vision, we used three challenging and widely used computer vision datasets to study the performance of our framework: (*i*) **Fashion-MNIST (FMNIST)** [41]; (*ii*) **CIFAR-10** [17]; (*iii*) **CIFAR-100** [17]. We also studied the performance of our framework on two text mining datasets (detailed below).

Experimental Setup: Each dataset was divided into 4 subsets. The first subset was used to train the blackbox teacher model. The other subsets were used as the initial labeled set L, unlabeled set U and test set to actively train the student. The number of samples in each subset for each dataset, together with the accuracy of the teacher model, are detailed in Table 1. Each algorithm selected a batch of k unlabeled samples in each AL iteration (where k is the query budget / batch size). The labels of the selected samples were obtained from the

teacher and the newly labeled samples were added to the labeled training set. The student network was trained on the updated labeled set and its accuracy was computed on the test set. The process was continued for 15 iterations (taken as the stopping criterion in this work). All the results were averaged over 3 runs to rule out the effects of randomness. The vanilla knowledge distillation algorithm proposed by Hinton *et al.* [11] was used as the underlying KD algorithm for knowledge transfer.

The batch size k was taken as 300; the weight parameter β in Equation (20) was taken as 0.5, the parameters α and θ in Equation (7) were taken as 10^{-6} and 10^{-5} respectively. The matrices Q, $\widehat{Q}$ and λ were all initialized to 0. Following the convention in knowledge distillation research [9, 39], the teacher was considered the oracle in our empirical studies; that is, the labels furnished by the teacher in response to the sample queries were considered the ground-truth and were used to train the student network. The labels of the samples in the initial training set L were also obtained from the teacher model.

Teacher Student Network Architectures: The architectures of the teacher and student networks for each dataset are also shown in Table 1. Such architectures have been used with these datasets in previous KD research [31, 40].

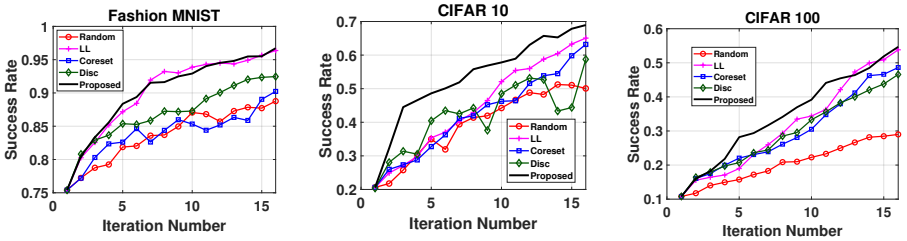
Table 1. Details of our experimental setup. The columns respectively denote the dataset, number of samples used to train the teacher model, the generalization accuracy of the teacher, the number of samples in the initial training set L, the unlabeled set U, the test set, the network architecture of the teacher model and the student model.

	Teacher Train	Teacher Acc.(%)	Initial Train	Unlabeled	Test	Teacher Arch	Student Arch
FMNIST	30,000	88.27	500	5,000	10,000	LeNet-5	LeNet-5-Half
CIFAR 10	30,000	75.34	500	5,000	10,000	AlexNet	AlexNet-Half
CIFAR 100	30,000	68.63	500	5,000	10,000	ResNet-34	ResNet-18
IMDB	25,000	84.28	500	5,000	6,500	BERT	DistilBERT
Tripadvisor	10,000	86.71	400	5,000	5,000	BERT	DistilBERT

Implementation Details: Please refer to the Supplemental File regarding the implementation details for training the teacher and student models.

Evaluation Metric: We used the *distillation success rate* [39] as the evaluation metric in this research. It computes the amount of knowledge the student network distills from the teacher and is computed as the ratio between the student's classification accuracy and the teacher's accuracy on the test set. A high value of this metric denotes better performance.

Comparison Baselines: The following AL algorithms were used as comparison baselines in our work: (*i*) **Random Sampling**, where a batch of unlabeled samples was selected at random; (*ii*) **Learning Loss for Active Learning (LL)** [43]; (*iii*) **Coreset** [33]; and (*iv*) **Discriminative AL (Disc)** [8]. *LL* and *Disc* are widely used techniques in recent active learning research [28]; *Coreset* has been used in the context of AL for knowledge distillation [18] and was hence selected as a comparison baseline.



(a) Fashion MNIST (b) CIFAR 10 (c) CIFAR 100

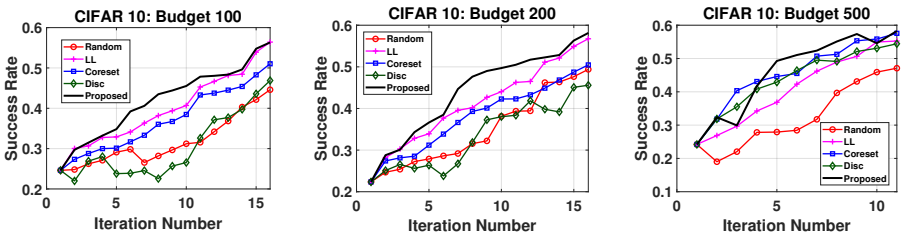
Fig. 1. Active Learning performance comparison. The x-axis denotes the iteration number and the y-axis denotes the distillation success rate on the test set. Best viewed in color.

4.1 Active Learning Performance

The AL performance results are depicted in Figure 1. In each figure, the x-axis denotes the iteration number, and the y-axis denotes the distillation success rate. *Random Sampling* does not produce good performance and achieves low distillation success rates with increasing size of the training set. The *Coreset* and *Disc* methods perform better than *Random Sampling* for the CIFAR-100 dataset. However, for CIFAR-10 and FMNIST datasets, the performance of *Coreset* is almost similar to *Random Sampling* and is sometimes inferior to *Random Sampling*; the *Disc* method mostly outperforms *Coreset* in the initial AL iterations. Both these observations are consistent with [8]. The *Learning Loss* method depicts the best performance among the baselines. The proposed method consistently depicts impressive performance and shows a steady growth in the distillation success rate with increasing label queries. It depicts the highest success rate for most of the AL iterations across all three datasets; it also attains the highest success rate at the end of 15 AL iterations, for all three datasets. Thus, by minimizing the reconstruction error, the proposed method is able to identify a batch of exemplar samples which well-represent the unlabeled data. These results unanimously depict the potential of the proposed AL technique to actively distill knowledge from the teacher to the student network, when the number of label queries to the teacher is constrained by a given budget.

4.2 Study of Query Budget

The goal of this experiment was to study the effect of query budget on the AL performance. The results on the CIFAR-10 dataset for query budgets $100, 200$ and 500 are presented in Figure 2. The results depict a similar trend as Figure 1. However, the performance of *Learning Loss* is not consistent across different budgets; it sometimes depicts marginally worse performance than *Coreset* and *Disc* (Figure 2(c)). Our framework outperforms the baselines consistently across all query budgets. This shows the practical usefulness of our algorithm, as the

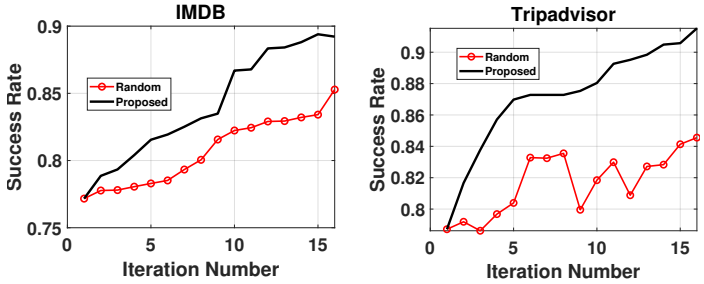


(a) Budget 100 (b) Budget 200 (c) Budget 500

Fig. 2. Study of query budget on the CIFAR 10 dataset. The results with budget 300 are presented in Figure 1(b) and are not included here. Best viewed in color.

query budget is often application specific and is dependent on the resources available for a given application.

4.3 Performance on Text Mining



(a) IMDB (b) Tripadvisor

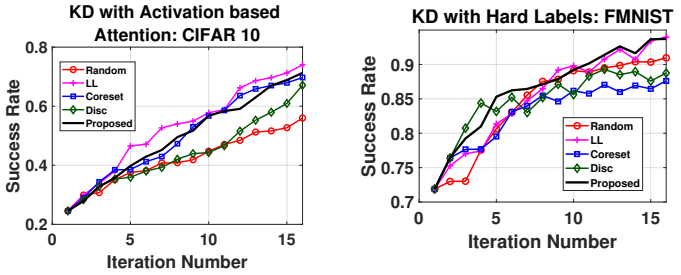
Fig. 3. Performance comparison on text mining datasets. Best viewed in color.

One of the useful features of our algorithm is its ability to generalize across multiple network architectures and hence, multiple applications. To demonstrate this, we studied its performance on text mining. We used the IMDB [21] and Tripadvisor [1] datasets for this experiment; the number of samples used are detailed in Table 1. The *BERT* model (based on a Transformer architecture) was used as the teacher and *DistilBERT* (a sub-network of BERT with half the number of layers) was used as the student for this experiment, similar to [31]. The KD algorithm was kept the same [11]. The baseline methods have largely been studied with CNN architectures for computer vision applications. The *Learning Loss* method, for instance, has mostly been applied with CNNs and its integration with transformer based architectures is not straightforward. We therefore compared our framework against *Random Sampling* in this study. The

results are presented in Figure 3. Our algorithm comprehensively outperforms *Random Sampling* and attains a much better success rate at the end of the AL iterations. For Tripadvisor, the improvement in the final success rate is about 7%. This corroborates the ability of our algorithm to seamlessly integrate across multiple teacher-student network architectures, and its ease of applicability in different domains.

4.4 Study of the Underlying KD Algorithm

Another useful feature of our framework is its independence of the underlying KD algorithm. In this experiment, we studied the performance of our framework in conjunction with the KD algorithm that uses activation based spatial attention as a mechanism of transferring knowledge from the teacher to the student network [47]. As in Table 1, we used AlexNet as the teacher and AlexNet-half as the student in this experiment. The results on the CIFAR-10 dataset are shown in Figure 4(a). We also analyzed our framework with a KD algorithm where the teacher furnishes only hard labels (instead of the soft-label probabilities) in response to each sample query. We used LeNet-5 as the teacher and LeNet-5-half as the student, as in Table 1. The results on the FMNIST dataset are depicted in Figure 4(b).



(a) KD with Activation based Attention: CIFAR 10

(b) KD with Hard Labels: FMNIST

Fig. 4. Study of the underlying knowledge distillation (KD) algorithm on the CIFAR 10 and FMNIST datasets. Best viewed in color.

Our framework once again depicts competitive performance, demonstrating its generalizability across different KD algorithms (even when the teacher produces only hard labels). The *Learning Loss* method also depicts good performance.

We also conducted experiments to study the effect of the weight parameter β in Equation (20) and the computation time of all the algorithms. These results are included in the Supplemental File due to space constraints.

5 Conclusion and Future Work

In this paper, we proposed a novel active learning algorithm for knowledge distillation applications. Such an algorithm can be immensely useful in training a lightweight student model to imitate a more complex teacher model, when the number of queries to the teacher cannot exceed a pre-specified budget. We posed the selection of exemplar training samples (to distill knowledge from the teacher to the student) as an NP-hard optimization problem and solved it using an iterative algorithm with global convergence. Our framework is independent of the underlying KD algorithm, as well as the architectures of the teacher and student networks, and can thus be seamlessly integrated across different KD applications. Our extensive empirical analyses verified the effectiveness of our framework for cost-effective blackbox knowledge distillation. Although validated on the KD application in this paper (as this is an under-explored research area), the proposed method is generic and can be used in any application to select an informative subset of samples from large amounts of data. As part of future research, we plan to study the performance of our framework on applications beyond computer vision and text mining, and also for regression and multi-label knowledge distillation.

6 Acknowledgment

This research was supported in part by the National Science Foundation under Grant Number: IIS-2143424 (NSF CAREER Award).

References

1. Alam, H., Ryu, W., Lee, S.: Joint multi-grain topic sentiment: Modeling semantic aspects for online reviews. Information Sciences **339**, 206 – 223 (2016)
2. Ash, J., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: Deep batch active learning by diverse, uncertain gradient lower bounds. In: International Conference on Learning Representations (ICLR) (2020)
3. Boreshban, Y., Mirbostani, S., Ghassem-Sani, G., Mirroshandel, S., Amiriparian, S.: Improving question answering performance using knowledge distillation and active learning. In: arXiv:2109.12662 (2021)
4. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning **1**, 1 – 122 (2011)
5. Chen, H., Wang, Y., Xu, C., Xu, C., Tao, D.: Learning student networks via feature embedding. IEEE Transactions on Neural Networks and Learning Systems (TNNLS) **32**(1), 25 – 35 (2021)
6. Deng, J., Pan, Y., Yao, T., Zhou, W., Li, H., Mei, T.: Relation distillation networks for video object detection. In: IEEE International Conference on Computer Vision (ICCV) (2019)
7. Gan, C., Gong, B., Liu, K., Su, H., Guibas, L.: Geometry guided convolutional neural networks for self-supervised video representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

8. Gissin, D., Shalev-Shwartz, S.: Discriminative active learning. In: arXiv:1907.06347 (2019)
9. Gou, J., Yu, B., Maybank, S., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision (IJCV) **129**, 1789 – 1819 (2021)
10. He, B., Liao, L., Han, D., Hai, Y.: A new inexact alternating directions method for monotone variational inequalities. Mathematical Programming **92**(1), 103 – 118 (2002)
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: arXiv preprint arXiv:1503.02531 (2015)
12. Hu, P., Lipton, Z., Anandkumar, A., Ramanan, D.: Active learning with partial feedback. In: International Conference on Learning Representations (ICLR) (2019)
13. Hu, Z., Hou, W., Liu, X.: Deep batch active learning and knowledge distillation for person re-identification. IEEE Sensors Journal **22**(14) (2022)
14. Huang, S., Chen, J., Mu, X., Zhou, Z.: Cost-effective active learning from diverse labelers. In: International Joint Conference on Artificial Intelligence (IJCAI) (2017)
15. Huang, Z., Wang, N.: Like what you like: Knowledge distill via neuron selectivity transfer. In: arXiv preprint arXiv:1707.01219 (2017)
16. Kimura, A., Ghahramani, Z., Takeuchi, K., Iwata, T., Ueda, N.: Few-shot learning of neural networks from scratch by pseudo example optimization. In: British Machine Vision Conference (BMVC) (2018)
17. Krizhevsky, A.: Learning multiple layers of features from tiny images. In: Technical Report, University of Toronto (2009)
18. Kwak, B., Kim, Y., Kim, Y., Hwang, S., Yeo, J.: Trustal: Trustworthy active learning using knowledge distillation. In: AAAI Conference on Artificial Intelligence (2022)
19. Lee, S., Song, B.: Graph-based knowledge distillation by multi-head attention network. In: British Machine Vision Conference (BMVC) (2019)
20. Li, C., Wang, X., Dong, W., Yan, J., Liu, Q., Zha, H.: Joint active learning with feature selection via cur matrix decomposition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **41**(6), 1382 – 1396 (2019)
21. Maas, A., Daly, R., Pham, P., Huang, D., Ng, A., Potts, C.: Learning word vectors for sentiment analysis. In: Association for Computational Linguistics (ACL) (2011)
22. Micaelli, P., Storkey, A.: Zero-shot knowledge transfer via adversarial belief matching. In: Neural Information Processing Systems (NeurIPS) (2019)
23. Nayak, G., Mopuri, K., Shaj, V., Babu, R., Chakraborty, A.: Zero-shot knowledge distillation in deep networks. In: International Conference on Machine Learning (ICML) (2019)
24. Nie, F., Wang, H., Huang, H., Ding, C.: Early active learning via robust representation and structured sparsity. In: International Joint Conference on Artificial Intelligence (IJCAI) (2013)
25. Nie, X., Li, Y., Luo, L., Zhang, N., Feng, J.: Dynamic kernel distillation for efficient pose estimation in videos. In: IEEE International Conference on Computer Vision (ICCV) (2019)
26. Osmanbeyoglu, H., Wehner, J., Carbonell, J., Ganapathiraju, M.: Active machine learning for transmembrane helix prediction. BMC Bioinformatics **11**(1) (2010)
27. Peng, F., Wang, C., Liu, J., Yang, Z.: Active learning for lane detection: A knowledge distillation approach. In: IEEE International Conference on Computer Vision (ICCV) (2021)
28. Ren, P., Xiao, Y., Chang, X., Huang, P., Li, Z., Gupta, B., Chen, X., Wang, X.: A survey of deep active learning. ACM Computing Surveys **54**(9), 1 – 40 (2021)

29. Romero, A., Ballas, N., Kahou, S., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: International Conference on Learning Representations (ICLR) (2015)
30. Ruchansky, N., Crovella, M., Terzi, E.: Matrix completion with queries. In: ACM Conference on Knowledge Discovery and Data Mining (KDD) (2015)
31. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. In: Neural Information Processing Systems (NeurIPS) Workshop (2019)
32. Sculley, D.: Online active learning methods for fast Label-Efficient spam filtering. In: Fourth Conference on Email and AntiSpam (2007)
33. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: International Conference on Learning Representations (ICLR) (2018)
34. Settles, B.: Active learning literature survey. In: Technical Report 1648, University of Wisconsin-Madison (2010)
35. Shi, L., Shen, Y.: Diversifying convex transductive experimental design for active learning. In: International Joint Conference on Artificial Intelligence (IJCAI) (2016)
36. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: IEEE International Conference on Computer Vision (ICCV) (2019)
37. Su, J., Tsai, Y., Sohn, K., Liu, B., Maji, S., Chandraker, M.: Active adversarial domain adaptation. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2020)
38. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. Journal of Machine Learning Research (JMLR) **2**, 45–66 (2001)
39. Wang, D., Li, Y., Wang, L., Gong, B.: Neural networks are more productive teachers than human raters: Active mixup for data-efficient knowledge distillation from a blackbox model. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
40. Wang, Z.: Zero-shot knowledge distillation from a decision-based black-box model. In: International Conference on Machine Learning (ICML) (2021)
41. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. In: arXiv:1708.07747 (2017)
42. Yang, J., Yin, W., Zhang, Y., Wang, Y.: A fast algorithm for edge preserving variational multichannel image restoration. SIAM Journal on Imaging Sciences **2**(2), 569 – 592 (2009)
43. Yoo, D., Kweon, I.: Learning loss for active learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
44. Yu, K., Bi, J., Tresp, V.: Active learning via transductive experimental design. In: International Conference on Machine Learning (ICML) (2006)
45. Yuan, F., Shou, L., Pei, J., Lin, W., Gong, M., Fu, Y., Jiang, D.: Reinforced multi-teacher selection for knowledge distillation. In: AAAI Conference on Artificial Intelligence (2021)
46. Yun, J., Kim, B., Kim, J.: Weight decay scheduling and knowledge distillation for active learning. In: European Conference on Computer Vision (ECCV) (2020)
47. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations (ICLR) (2017)
48. Zhu, J., Bento, J.: Generative adversarial active learning. In: Workshop at Neural Information processing Systems (NeurIPS-W) (2017)