# Cross-domain Knowledge Distillation for Retrieval-based Question Answering Systems

Cen Chen[1], Chengyu Wang[2], Minghui Qiu[2,*], Dehong Gao[2], Linbo Jin[2], Wang Li[1]
[1] Ant Group,, China   [2] Alibaba Group,, China
{chencen.cc,raymond.wang}@antfin.com
{chengyu.wcy,minghui.qmh,dehong.gdh,yuyi.jlb}@alibaba-inc.com

## ABSTRACT

Question Answering (QA) systems have been extensively studied in both academia and the research community due to their wide real-world applications. When building such industrial-scale QA applications, we are facing two prominent challenges, i.e., i) lacking a sufficient amount of training data to learn an accurate model and ii) requiring high inference speed for online model serving. There are generally two ways to mitigate the above-mentioned problems. One is to adopt transfer learning to leverage information from other domains; the other is to distill the "dark knowledge" from a large teacher model to small student models. The former usually employs parameter sharing mechanisms for knowledge transfer, but does not utilize the "dark knowledge" of pre-trained large models. The latter usually does not consider the cross-domain information from other domains. We argue that these two types of methods can be complementary to each other. Hence in this work, we provide a new perspective on the potential of the teacher-student paradigm facilitating cross-domain transfer learning, where the teacher and student tasks belong to heterogeneous domains, with the goal to improve the student model's performance in the target domain. Our framework considers the "dark knowledge" learned from large teacher models and also leverages the adaptive hints to alleviate the domain differences between teacher and student models. Extensive experiments have been conducted on two text matching tasks for retrieval-based QA systems. Results show the proposed method has better performance than the competing methods including the existing state-of-the-art transfer learning methods. We have also deployed our method in an online production system and observed significant improvements compared to the existing approaches in terms of both accuracy and cross-domain robustness.

## CCS CONCEPTS

• **Applied computing** → *Electronic commerce*; • **Information systems** → *Retrieval models and ranking*.

## KEYWORDS

transfer learning, knowledge distillation, retrieval-based question answering, text matching.

[*] Minghui Qiu is the corresponding author.

*WWW '21, April 19–23, 2021, Ljubljana, Slovenia*
© 2021 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449814

**ACM Reference Format:**
Cen Chen, Chengyu Wang, Minghui Qiu, Dehong Gao, Linbo Jin, and Wang Li. 2021. Cross-domain Knowledge Distillation for Retrieval-based Question Answering Systems. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3442381.3449814

## 1 INTRODUCTION

A Question Answering (QA) system is a typical information retrieval and NLP system that outputs a response given a user question query. It has been extensively studied in both academia and the research community due to its wide real-world applications such as Amazon Alexa, Apple Siri and Alibaba Alime. General approaches for building such QA systems include retrieval-based methods [38, 39, 49], generation-based methods [27, 30] and hybrid methods [25, 31, 40]. The fundamental problem for retrieval-based QA systems is to retrieve the most similar question from the QA knowledge base given a query, so as to provide the respective answer. Such a *text* (i.e., query-question) *matching problem* can be represented as Paraphrase Identification (PI) or some form of Natural Language Inference (NLI) [12, 49]. For example, if we could identify a question as paraphrase or if a question could be entailed by the query, we can directly retrieve the answer to that question from the underlying QA knowledge base as the response.

When dealing with such text matching problems in the real-world industrial-scale QA applications, we are facing two prominent challenges, i.e., i) the lack of abundant data to learn a model with high accuracy and ii) the requirement of high inference speed for online model serving. Recent advances on text matching rely heavily on the flourishment of deep learning models [15, 22]. On the one hand, those deep models are proven to be effective when rich in-domain labeled data is available. However, in real-world applications, it is challenging to obtain a sufficient amount of labeled data for every domain of interest, as data annotation is commonly time-consuming and costly. On the other hand, high Query-Per-Second (QPS) requirements for seamless online serving demand the deployed models to be light-weight. Thus the trained models have to be either designed to be simple in structure but effective in performance, or compressed if well-performed large models are originally trained. Therefore, there is a great incentive for researchers to establish effective algorithms that can utilize data or knowledge from related domains to train an accurate model for the target domain which is small in size.

A promising way to mitigate the above-mentioned problems is to adopt transfer learning. Transfer learning has been widely studied over recent years to improve the model performance of the data-insufficient target domain by leveraging knowledge acquired

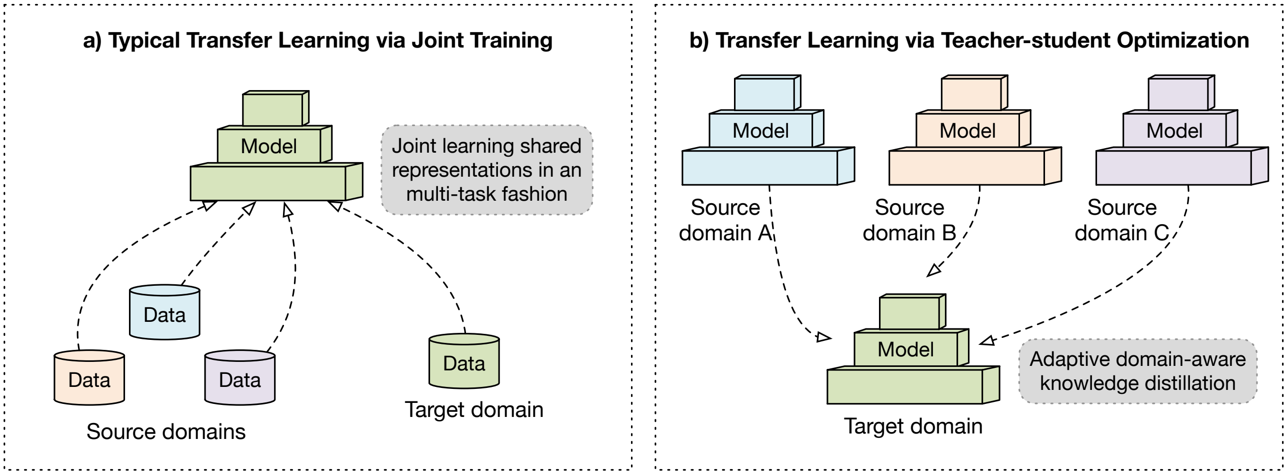
a) Typical Transfer Learning via Joint Training
Model
Joint learning shared representations in an multi-task fashion
Data
Data
Data
Data
Source domains
Target domain
b) Transfer Learning via Teacher-student Optimization
Model
Model
Model
Source domain A
Source domain B
Source domain C
Model
Adaptive domain-aware knowledge distillation
Target domain

**Figure 1: A conceptual illustration of the knowledge distillation with multiple cross-domain teachers for training a student network. Our goal is to selectively distill knowledge from multiple out-domain teachers taking into consideration of teacher domain expertise for effective adaptive knowledge transfer.**

from different but related source domains [23, 33]. The recent advance of transfer learning is mostly based on deep learning and typically considers *parameter sharing* mechanisms such as "fully-shared" and "shared-private" model architectures [16, 21, 47]. Those models transfer knowledge by jointly learning domain-shared feature representations and have demonstrated remarkable success in many real-world applications in NLP [16, 21, 33, 47]. As shown in Figure 1(a), such a transfer learning paradigm is able to benefit from multiple data sources and the model structure is often designed to be light-weight that enables fast online inference. For example, the study in [49] proposes a light-weight CNN-based method with transfer learning to boost model performance in an online QA system.

Meanwhile, the recent emergence of large pre-trained models, such as BERT [7] and XLNet [42], has revolutionized the learning paradigm of many NLP tasks and pushed the performance of those tasks to new heights. With those pre-trained models, a fine-tuning approach is then adopted to improve task-specific model performance. Although good model performance can be achieved with fine-tuning, the resulting model size is unavoidably large. This naturally leads to the question: *how can we utilize pre-trained models together with other data sources from different domains to facilitate knowledge transfer that is effective in performance and efficient in serving?*

To this end, we propose a new transfer learning framework named *Domain-Aware Knowledge Distillation* (DAKD) for cross-domain text matching. A conceptual illustration of the proposed method is outlined in Figure 1(b). Departing from the traditional transfer learning methods outlined in Figure 1(a),we turn to an alternative solution, i.e., extracting knowledge from pre-trained models in source domains to guide the training of small models in the target domain model, as presented in Figure 1(b).

Such teacher-student optimization process is also known as *Knowledge Distillation* (KD). KD was originally used for compressing pre-trained large teacher neural networks into smaller ones [11], and later it was found to be useful when training a student model that has the same architecture as the teacher in which the student excels the teacher [8]. The differences between our proposed DAKD framework and previous methods are summarized in Table 1. As seen, previous works pay more attention to model compression or knowledge transfer among the same task or different tasks in the same domain, using either the "dark knowledge", i.e., the predicted outputs of teacher models [5, 8, 11, 48] or hints, i.e., intermediate representations [28, 46, 48] of teacher models, mostly for computer vision applications.

**Table 1: Comparison of classical KD approaches on four dimensions, where "HD" indicates that teacher and student tasks belong to heterogeneous domains, "MT" denotes using multiple teachers for KD, "Dark" and "Hint" specify whether the "dark knowledge", i.e., predictions of the teacher models, or the hints, i.e., intermediate representations of the teacher models, are leveraged for training the student. The first three are classical approaches, while the middle ones are the recent works focusing on utilizing multiple teachers.**

| | HD | MT | Dark | Hint |
|---|---|---|---|---|
| Original KD (2015) [11] | | ✓ | ✓ | |
| FitNet (2015) [28] | | | | ✓ |
| A Gift from KD (2017) [46] | ✓ | | | ✓ |
| Born-Again Network (2018) [8] | | | ✓ | |
| Multiple Teachers (2017) [48] | | ✓ | ✓ | ✓ |
| Diverse Peers (2019) [5] | | ✓ | ✓ | |
| Domain-aware KD (Ours) | ✓ | ✓ | ✓ | ✓ |

This paper provides a new perspective on the potential of the teacher-student paradigm facilitating *transfer learning across domains*, where teacher and student tasks belong to heterogeneous domains, with the goal to improve the performance of the student model in the target domain. The teacher models can be either 1) small models with the same architecture as the student that are trained using the source domain data or 2) large models that fine-tune the pre-trained language models on the source domain data [1]. The resulting student model is a strong text matching model that is small in model size to meet the online serving needs [24, 26, 49].

As domain shift often exists, the student model employs the "*shared-private*" architecture to capture domain-specific and domain-invariant features. Domain-invariant features are guided by teachers' hints, while the learning targets are partially supervised by teachers' dark knowledge. Although teachers may be from different source domains, intuitively they can still provide constructive guidance to the student model from a different angle. Thus, to reduce domain gaps, teachers' domain *expertise scores* on the student task are measured for adaptive distillation. Domain-invariant features are learned with the objective to reduce the divergence between the distributions of student's intermediate layers and those of the teachers.

Extensive experiments have been conducted on two benchmark text matching datasets for retrieval-based QA systems. Results show DAKD has better performance than the competing methods including state-of-the-art transfer learning methods. We have also deployed our method in a chatbot system for an online A/B test and observed significant improvements. Additional experiments also show that the proposed framework has generalization capabilities that are helpful for other NLP tasks, such as review analysis.

The remainder of this paper is organized as follows. Section 2 describes the task followed by presenting the proposed approach in detail. All the experiments are shown in Section 3. Finally, Section 4 reviews the related work and Section 5 concludes the paper.

## 2 DAKD: THE PROPOSED APPROACH

In this section, we present the technical details on our proposed *Domain-Aware Knowledge Distillation* (DAKD) framework for cross-domain knowledge transfer on text matching in retrieval-based QA systems.

### 2.1 Problem Formulation and Model Overview

*2.1.1 Text Matching.* For retrieval-based QA systems, Paraphrase Identification (PI) and Natural Language Inference (NLI) are crucial tasks for question matching. Both tasks can be unified as a text (query-question) matching problem, which is typically modeled as a text pair classification task. Formally, given two collections of text pairs $\mathcal{X}^1 = \{X_1^1, X_2^1, ..., X_{l_1}^1\}$ and $\mathcal{X}^2 = \{X_1^2, X_2^2, ..., X_{l_2}^2\}$, where $l_1$ and $l_2$ denote the lengths (the size of the datasets) of $\mathcal{X}^1$ and $\mathcal{X}^2$ respectively. Our task is to predict a binary classification label $y \in \{0, 1\}$ that indicates whether a pair of texts $X^1$ and $X^2$ are semantically related.

Note that, for the PI task, the label $y = 1$ indicates $X^1$ can be identified as a paraphrase of $X^2$, while for the NLI task, the label $y = 1$ indicates $X^2$ can be inferred from the $X^1$, i.e., entailment.

*2.1.2 Transfer Learning Setting.* In the scope of DAKD, we consider the transfer learning setting where we are given labeled data from multiple source domains $\mathcal{D}_m^s$ and one target domain $\mathcal{D}^t$ [26]. Here, we have $m \in [1, K]$ and $K$ denotes the total number of source domains. We seek to use both $\mathcal{D}_m^s$ ($m \in [1, K]$) and $\mathcal{D}^t$ to help the learning of the student model in the target domain.

*2.1.3 Teacher-student Optimization.* The teacher-student optimization process was first introduced in *Knowledge Distillation* (KD) where the student network tries to mimic the behavior of pre-trained teacher networks [11]. Such teacher-student paradigm utilizes the supervision signals from the teachers to guide the student's learning task by adding extra terms to the loss function.

Let $(\mathcal{M}_1^{teacher}, \mathcal{M}_2^{teacher}, ..., \mathcal{M}_m^{teacher})$ denote the set of teacher networks. In the context of cross-domain transfer learning, those teacher networks can be either small networks with the same architecture as the student or large pre-trained language models with domain-specific fine-tuning. The student network $\mathcal{M}^{student}$ optimizes its own in-domain task objective and leverages the teachers' knowledge by regularizing the outputs and the intermediate hints.

Figure 2 shows an overview of the proposed DAKD framework. The teacher networks are first pre-trained using their respective source domain data $\mathcal{D}_m^s$. Different from standard KD, our teacher network is not required to be deeper or larger than the student. Subsequently, the parameters of teacher networks are frozen and the student network is trained using the target domain data $\mathcal{D}^t$ with the guidance from the teacher networks.

### 2.2 Learning from Heterogeneous Teacher Models

In this section, we introduce how the knowledge is transferred from heterogeneous domain teachers via KD and domain adaptation. For easy explanation, we start with the single-teacher setting, followed by the multi-teacher setting.

*2.2.1 Domain-aware Distillation.* Domain-aware distillation aims to distill the knowledge from the teacher model to the student model, considering the gap between the two domains. Specifically, the distillation process seeks to minimize the prediction differences. There are generally two types of distillation strategies, i.e., a soften version and a hard one. The former aligns a softer probability distribution over the classes, while the latter aligns the final one-hot predictions.

Let $\left(X_i^1, X_i^2, y_i\right)$ be an input instance. We denote $f^s(X_i^1, X_i^2)$ as the logits from the output layer and $g^s(X_i^1, X_i^2)$ as the prediction label from the teacher model. For the student side, $f^t(X_i^1, X_i^2)$ and $g^t(X_i^1, X_i^2)$ represents the logits and the prediction label, respectively. For a batch of data $\mathcal{B}$, We then define soft and hard KD losses as:

$$L_{soft}^{KD} = -\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} f^s(X_i^1, X_i^2) \cdot \log \left( f^t(X_i^1, X_i^2)/T \right), \tag{1}$$

$$L_{hard}^{KD} = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \left( g^s(X_i^1, X_i^2) - g^t(X_i^1, X_i^2) \right)^2, \tag{2}$$

where $T$ is the temperature value that is set as 1 by default. A higher $T$ value produces a softer probability distribution over classes.

Besides learning from the teacher, the student also seeks to fit the training labels. For binary classification, a cross-entropy loss is

[1] In the experiments, we have tested on both kinds of teacher settings to examine the effectiveness of our proposed framework regardless of the teacher model structure.

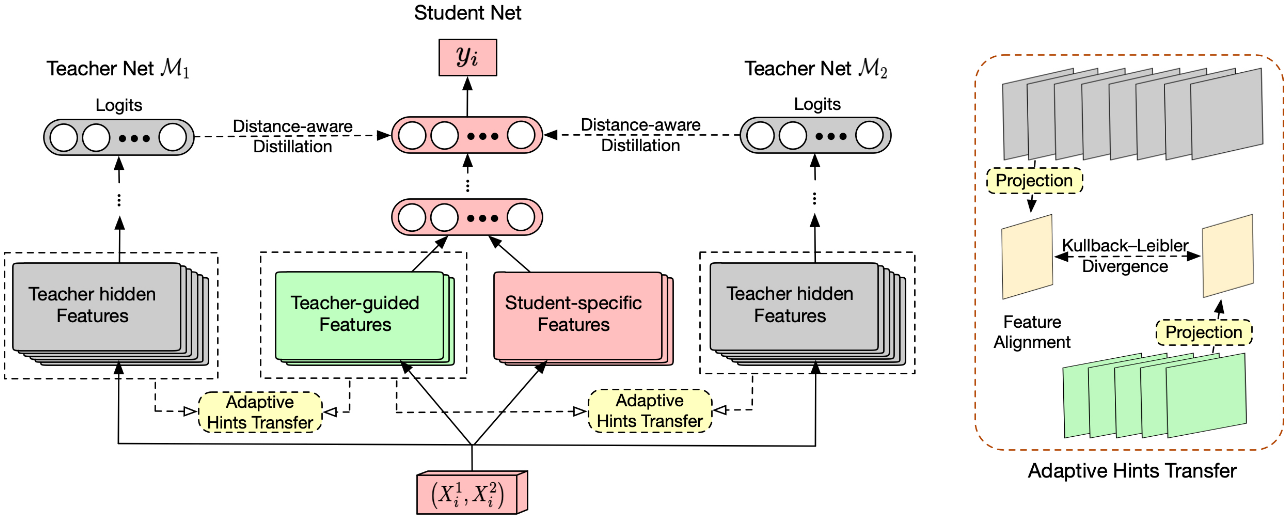

**Figure 2: The proposed DAKD framework with heterogeneous teachers. Those teachers with distance-aware guidance can be either in-domain or out-domain. The student models seeks to leverage both the dark knowledge, i.e., softened outputs, and the hints, i.e., hidden representations/features from the teachers during the distillation process. More specifically, domain expertise is taken into consideration when distilling the dark knowledge, while hints are adaptively transferred via features adaptation.**

adopted, defined as follows:

$$L^t = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\Big(y_i \log(g^t(X_i^1, X_i^2)\Big) + (1-y_i)\log\Big(1-g^t(X_i^1, X_i^2))\Big). \tag{3}$$

### 2.2.2 *Domain Expertise.*

Traditional KD models work well when the student and the teacher are in the same domain, which is often the assumption. Model performance can drop significantly with increasing domain discrepancy [4]. Nevertheless, in the transfer learning setting, source and target domains are different. We consider the teacher's expertise on the target domain when transferring the distilled knowledge from the teacher model to the student.

Specifically, we use a scalar $\lambda_i^m$ to represent the distance between the prediction of the teacher model $\mathcal{M}_m^{teacher}$ and the ground truth label $y_i$ for a particular input instance $(X_i^1, X_i^2)$. Such distances can be viewed as evidences of confidence for the teacher. Intuitively we can use the teacher's prediction error to measure this distance. The rationale is that if a teacher is able to make good predictions for a target input, the teacher should be trusted more for this judgment, thus can provide suitable guidance for the student. Hence, we define the domain expertise $\lambda_i^m$ as follows:

$$\lambda_i^m = \frac{1}{\exp^{(g^s(X_i^1, X_i^2)-y_i)^2}+1}. \tag{4}$$

Clearly, the weight $\lambda_i^m$ is the largest when the teacher model makes perfect predictions on the target input, i.e., $g^s(X_i^1, X_i^2) = y_i$.

From the experiments, we find the soft KD loss is able to encode more distribution information. Hence, a soft KD loss is adopted in our final model. To incorporate $\lambda_i^m$, the soft KD loss is rewritten as:

$$L_{soft}^{KD} = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\lambda_i^m f^s(X_i^1, X_i^2)\cdot\log\Big(f^t(X_i^1, X_i^2)/T\Big). \tag{5}$$

### 2.2.3 *Feature Adaptation.*

To improve the efficacy of knowledge transfer, we also incorporate the teacher's *hints*, where the hints are the intermediate features at the hidden layers in the teacher network. As domain difference poses a major obstacle in transferring features across domains, we use the "*shared-private*" architecture [16] for the student network, which consists of two subnetworks that allows modeling both domain-specific and domain-invariant features. As shown in Figure 2, domain-specific features are learned independently by the *student-specific* sub-network, while domain-invariant features in the *teacher-guided* sub-network are adaptively guided by the teacher's hints.

Specifically, for the text matching task, we denote $H^0$ and $H^1$ as the text representation layers for the teacher-guided sub-network and the student-specific sub-network in the student model, respectively. Let $H_m^{teacher}$ be the text representation layers in a pre-trained teacher model $\mathcal{M}_m^{teacher}$. We confine the student representation $H^0$ to be as close as the teacher representation $H_m^{teacher}$ by an adaptation loss:

$$L^d = Dist\Big(H_m^{teacher}(X_i^1, X_i^2), H^0(X_i^1, X_i^2)\Big), \tag{6}$$

$$= \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|} KL\Big(H_m^{teacher}(X_i^1, X_i^2), W_m^\top H^0(X_i^1, X_i^2)\Big) \tag{7}$$

where $|\mathcal{B}|$ denotes the batch size, $KL$ is the Kullback–Leibler divergence function, and $W_m$ serves as a projection matrix to project the student's representations to the teacher's.

**Algorithm 1** Training Procedure of DAKD

```
Require:
    $K$ source domain datasets $\mathcal{D}^s_m,\ m \in [1, K]$
    The target domain dataset $\mathcal{D}^t$
    Stage 1: Pre-training source domain models:
for each source domain $m \in [1, K]$ do
    Pre-train the teacher model $\mathcal{M}^{teacher}_m$ using the source domain dataset $\mathcal{D}^s_m$
end for
    Stage 2: Domain-aware teacher-student optimization:
for each batch $\mathcal{B}$ in training data $\mathcal{D}^t$ do
    for each teacher $m \in [1, K]$ do
        Feed batch data $\mathcal{B}$ to model $\mathcal{M}^{teacher}_m$
        Obtain meta-information $\mathcal{I}_m$ (logits, predictions, hints, etc.)
    end for
    Update model $\mathcal{M}^{student}$ using meta-information and training instances $(\mathcal{B}, \{\mathcal{I}_m\}^K_{m=1})$
end for
```

Thus, the student network's classification loss of the target domain $L^t$ is rewritten as follows:

$$g^t(X^1_i, X^2_i) = \sigma\Big(H^0(X^1_i, X^2_i) \oplus H^1(X^1_i, X^2_i)\Big), \tag{8}$$

$$L^t = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\Big(y_i \log(g^t(X^1_i, X^2_i)\Big) + (1 - y_i)\log\left(\left(1 - g^t\left(X^1_i, X^2_i\right)\right)\right) \tag{9}$$

where $\sigma$ is the output layer that transforms the text representations into the predictions.

Finally, the complete training loss of the student network is defined as:

$$\mathcal{L} = L^t + aL^d + bL^{KD}, \tag{10}$$

where $L^t, L^d, L^{KD}$ are the prediction loss, the adaption loss and the distillation loss, respectively. Scalars $a, b$ are applied to balance the losses. For simplicity, we omit the regularization terms of model parameters in the loss function.

*2.2.4 Combining Multiple Teachers.* Our framework can easily incorporate $K$ teachers. Similarly, the overall training loss of the student network is defined as follows:

$$\mathcal{L} = L^t + a\frac{1}{K}\sum_{m=1}^{K} L^d_m + b\frac{1}{K}\sum_{m=1}^{K} L^{KD}_m, \tag{11}$$

with $\lambda^m_i$ being replaced by the normalized expertise score with respect to all $K$ teachers:

$$\lambda^m_i = \frac{1/\exp^{(g^s_m(X^1_i, X^2_i) - y_i)^2}}{\sum_{m=1}^{K} 1/\exp^{(g^s_m(X^1_i, X^2_i) - y_i)^2}}. \tag{12}$$

## 2.3 Training Procedure

We present the overall training procedure of DAKD in Algorithm 1. The training procedure consists of two stages. For the transfer learning setting, the teacher network(s) are first pre-trained by source domain datasets, which can be either deep or shallow depending on task-specific model choices. The student network leverages knowledge from teachers that may come from multiple source domains by taking into consideration their supervision signals. Task-specific classification, domain-aware distillation, and feature adaptation losses are jointly minimized via the teacher-student optimization paradigm.

# 3 EXPERIMENTS

In this paper, we follow the previous works on retrieval based QA systems [26, 49] and conduct extensive experiments on two frequently studied text matching tasks, i.e., Natural Language Inference and Paraphrase Identification to quantitatively and qualitatively examine the effectiveness of our proposed method. Our proposed method is also deployed in the production system and the extrinsic evaluation is conducted to verify its benefits on the online QA system. Additional experiments on text analysis task are performed to verify the generalizability of our method to other NLP tasks.

## 3.1 Datasets and Experimental Setups

*3.1.1 Tasks and Datasets.* Two representative text matching datasets are used for evaluation. Data statistics are summarized in Table 2. Detailed task descriptions and the datasets are briefly introduced below:

- **Paraphrase Identification (PI)** task aims to examine whether two texts have the same meaning. We treat the Quora Question Pairs (*Quora QP*) [2] dataset released by Quora[3], as the source domain and the *AnalytiCup* [4] dataset released by CIKM AnalytiCup 2018 as the target domain. Quora QP is a large-scale dataset that covers a variety of topics, while the AnalytiCup dataset consists of question pairs only from the E-commerce domain for cross-lingual text matching. For data preprocessing in both tasks, we follow the previous works of [26, 49].
- **Natural Language Inference (NLI)** is a task to determine the relations between sentence pairs, i.e., entailment, contradiction, or neutral, between sentence pairs. We use *MultiNLI* [37] as the source domain dataset and *SciTail* [1] as the target domain. SciTail is a recently released challenging textual entailment dataset, specifically collected from the science sources, which contains around 550k hypothesis/premise pairs. MultiNLI [1] is a large crowdsourced benchmark corpus with textual entailment information from a wider range of text genres/sources, which is more diverse. We use the 1.0 version of MultiNLI with the examples drawn from five domains. Thus, in the experiments, MultiNLI can be used as a whole to train one teacher model or separately for multiple teachers. Note that, the labels in SciTail only consist of "entailment" and "neutral". Thus, following the study in [26], we delete the "contradiction" samples from MultiNLI for the transfer learning purpose.

[2] https://www.kaggle.com/c/quora-question-pairs
[3] https://www.quora.com/
[4] https://tianchi.aliyun.com/competition/introduction.htm?raceId=231661

**Table 2: Data statistics for paraphrase identification and natural language inference tasks.**

| Task | Dataset | S-T Paradigm | Domain | Train | Dev | Test |
|---|---|---|---|---|---|---|
| Paraphrase Identification (PI) | Quora QP | Teacher | Quora QA Pairs | 404,287 | - | - |
| | AnalytiCup | Student | E-commerce QA Pairs | 6,668 | 3,334 | 3,330 |
| Natural Language Inference (NLI) | MultiNLI | Teacher(s) | Fiction | 77,348 | 2,000 | 2,000 |
| | | | Travel | 77,350 | 2,000 | 2,000 |
| | | | Slate | 77,306 | 2,000 | 2,000 |
| | | | Telephone | 83,348 | 2,000 | 2,000 |
| | | | Government | 77,350 | 2,000 | 2,000 |
| | | | All domains | 392,702 | 10,000 | 10,000 |
| | SciTail | Student | Science | 23,596 | 1,304 | 2,126 |

*3.1.2 Baselines.* To examine the effectiveness of the DAKD framework, we compare it with several transfer and non-transfer learning baselines:

- DAM [24]: the decomposable attention model (DAM) without considering knowledge transfer from other domains. It is a light-weight model that enables fast online inference and has shown strong performance on text matching tasks in real-world applications [26].
- Fully-Shared (FS) [21]: a classic transfer learning method that utilizes a shared encoder to learn transferable representations for both source and target data.
- Shared-Private (SP) [16]: extends the fully-shared model by incorporating domain-private sub-networks.
- Knowledge Distillation (KD) [11]: vanilla teacher-guided KD via transferring only the "dark knowledge". Two variants are considered when evaluating the distillation strategies.

*3.1.3 Choice of Student Models.* In the real world, it is a common practice to consider not only the training but also the inference efficiency [49]. Following the previous works for industrial applications, we use DAM [24] as the base model for the student network. In theory, we can also use more advanced models such as BERT [7] as our base student model. Such models have huge parameter size (with billions of parameters), which makes them difficult to be deployed for real-world applications. Our framework is able to leverage the strong representation learning power of those large models by treating them as teachers and distilling their knowledge into an efficient small student model such as DAM for online systems.

*3.1.4 Evaluation.* As both NLI and PI can be viewed as sentence pair classification, we employ the classification Accuracy (*ACC*) and the Area under the ROC curve (*AUC*) as our evaluation metrics. Significant tests are performed on accuracy only, because AUC is an overall metric that does not support significant tests.

*3.1.5 Implementation Details.* For DAM, we set the size of the hidden layers as 200. The max sequence length is set as 40 for PI and 50 for NLI. DAM is used as the base model for the student network in all the experiments. We also conduct experiments on large models, where BERT is adopted as the teacher. As for the BERT teacher, in PI and NLI tasks, we concatenate the text pairs $X^1$ and $X^2$ and pad them with special tokens as an input to the BERT. Specifically, the input is in the format: {`[CLS]` $X^1$ `[SEP]` $X^2$ `[SEP]`} [5]. The max sequence length for the sentence-pair is set as 80 for both PI and NLI. After BERT encoding, we use the contextual representation of the `[CLS]` token as input and add two fully-connected layers with the size of 64×2 to generate the output predictions. ReLU is used as the activation function, and Adam [14] is used with the initial learning rate of 0.001 for optimization. For both tasks, the data segmentation method [49] is employed. All the models are implemented with TensorFlow[6] and trained with NVIDIA Tesla P100 GPU. All the models are tuned with the validation data in target domain and results are reported on the test datasets, each averaged over 5 random runs.

## 3.2 Learning from Single Heterogeneous Teacher

We first compare DAKD with transfer learning baselines under the single teacher setting. In this setting, for KD-based methods, only one teacher is trained with the training data in the source domain. Results are shown in Table 3.

Regardless of how the knowledge is transferred, we find that knowledge learned from the source domain improves the performance of models for the target domain, evidenced by the fact that FS, SP, and KD-based methods outperform the base model without knowledge transfer. Interestingly, SP does not always outperform FS, which means different transfer learning methods may have their advantages on different datasets. KD has shown comparable results with transfer learning baselines. DAKD outperforms all the baselines by a large margin especially on the AUC metric, which shows the effectiveness of the teacher-student paradigm in transferring knowledge. Unlike vanilla KD, the proposed DAKD framework is better at adapting information from a heterogeneous source domain to help model learning in the target domain.

## 3.3 Learning from Multiple Teachers

*3.3.1 Single teacher v.s. multiple teachers.* To help us get deeper insights on learning from heterogeneous teachers, we conduct experiments to compare the effectiveness of a versatile teacher trained by using all the five domain data in MultiNLI versus five teachers trained separately by each domain data, denoted with "single

[5] Without ambiguity, in `[CLS]` $X^1$ `[SEP]` $X^2$ `[SEP]`, we assume that $X^1$ and $X^2$ also represent tokens of the corresponding sentence pair.

[6] https://www.tensorflow.org/

**Table 3: Results for PI and NLI under the single teacher setting. $^\dagger$ means significant difference over other baselines with $p < 0.1$.**

| Datasets | PI | | NLI | |
|---|---|---|---|---|
| | ACC | AUC | ACC | AUC |
| Base Model | 0.846 | 0.869 | 0.730 | 0.766 |
| FS | 0.849 | 0.871 | 0.745 | 0.804 |
| SP | 0.854 | 0.875 | 0.727 | 0.798 |
| KD | 0.860 | 0.871 | 0.751 | 0.805 |
| DAKD | **0.867**$^\dagger$ | **0.881** | **0.755**$^\dagger$ | **0.811** |

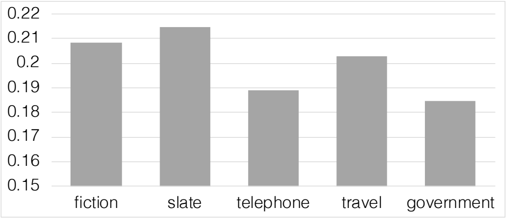

**Figure 3: Domain expertise $\lambda^m$ learned for multiple MultiNLI teachers.**

teacher” and “multiple teachers” respectively. Aside from DAKD, we also extend the original KD to multiple heterogeneous teachers by directly assembling the soft targets of the teachers. Classification AUC and accuracy results are presented in Table 5. Clearly, both teacher settings boost the performance of the target domain. The performance gain of multi-domain teachers for both KD and DAKD is larger than the gain of a single teacher. DAKD is the more effective than KD when incorporating knowledge from multiple cross-domain teachers. We speculate exposing the student to a diverse set of teachers can be beneficial in a similar way that the model ensemble benefits from multiple heterogeneous models, but further studies may be required to better understand this phenomenon.

*3.3.2 Visualizing domain expertise.* Furthermore, we average the domain expertise scores $\lambda_i^m$ as defined in Eqn. 4 for all the data instances and visualize the importance of different domains in Figure 3. We find the results are insightful. It reveals the importance of heterogeneous teachers from different source domains to the target task. Specifically, the domain “SciTtail” is close to “fiction” and “slate”, but not close to “telephone” and “government” in MultiNLI. This is intuitive as the SciTail data is created from multiple-choice science exams and web sentences. To summarize, the domain expertise scores are interpretable and provide us an insight into the importance of different domains, which boost the model performance in DAKD.

## 3.4 Learning from Pre-trained Language Models

Pre-trained language models such as BERT [7] have shown to be very successful in many NLP applications. A typical approach is fine-tuning pre-trained models for specific downstream tasks. To examine the generalization capability of DAKD, we adopt task-specific fine-tuned BERT [7] as our teacher model and DAM as the student model (denoted as DAKD from BERT), comparing with using DAM for both teacher and student models (denoted as DAKD from DAM).

The results are shown in Table 6. We find that fine-tuning BERT is a very strong baseline as it shows a clear advantage over the DAM-based methods. Despite its good performance, its inference time increases from 10ms to 200ms which is a bottleneck for online deployment. By replacing the teacher model from DAM to BERT, DAKD can further improve the results, e.g. from 0.867 to 0.899 in terms of ACC on the PI dataset. This shows a well-trained teacher model can provide the student with more transferable information. DAKD (from BERT) achieves the best performance with a very good inference speed (13 ms), which is suitable for deployment in online production, especially for industrial applications.

## 3.5 Detailed Model Analysis

*3.5.1 Distillation Strategies.* We compare DAKD with two types of KD variants with different distillation strategies on both PI and NLI datasets in Table 4. The soft KD variant distills knowledge by using soft targets, while the hard variant utilizes the predicted hard labels from the teachers. For fair comparison, the MultiNLI data is used as a whole for training one NLI teacher.

From the experimental results, we find that KD generally brings consistent performance gain for all models, except for Hard on the PI dataset (0.841 for Hard vs. 0.846 for Base). This can be justified by the fact that vanilla distillation may not always be beneficial due to domain difference. Moreover, using soft targets yields better results than using hard targets. This is intuitive as the soft targets can encode more distributional information. Furthermore, DAKD is the most robust for both datasets as it consistently outperforms the rest by a large margin, which echoes the benefits of domain adaptation in the KD process.

*3.5.2 Knowledge Transfer Strategies.* Recall that our method has two integral parts to facilitate knowledge transfer, i.e., dark knowledge and adaptive hints. We conduct an ablation study to examine the importance of these two components on both PI and NLI tasks, with results shown in Table 7. We have the following observations:

- We observe a clear performance drop when we remove the dark knowledge from DAKD, from 0.867 to 0.851 on the PI task and from 0.755 to 0.748 on the NLI task. This shows it is important to include such information for knowledge transfer.
- Adaptive hints are also crucial for boosting domain adaption capability. Without this component, the degenerated version of DAKD has less satisfactory results, i.e., the performance dropping from 0.867 to 0.859 on the PI task, and from 0.755 to 0.751 on the NLI task.

In a nutshell, both dark knowledge and adaptive hints are important to our method. With both of these components, DAKD demonstrates a clear advantage over the competing methods.

[7] We use the “bert-base-uncased” model with 12-layers pre-trained model from https://github.com/google-research/bert

**Table 4: Comparison of different distillation strategies for PI and NLI tasks. † means statistically significant difference over other baselines with $p < 0.1$ measured by the paired t-test. Note that AUC is an overall metric that does not support the significant test.**

| KD Strategies | Paraphrase Identification (PI) | | | | Natural Language Inference (NLI) | | | |
|---|---|---|---|---|---|---|---|---|
| | Base Model | KD-hard | KD-soft | DKDA | Base Model | KD-hard | KD-soft | DKDA |
| ACC | 0.846 | 0.841 | 0.860 | **0.867**† | 0.730 | 0.747 | 0.751 | **0.755**† |
| AUC | 0.869 | 0.862 | 0.871 | **0.881** | 0.766 | 0.801 | 0.805 | **0.811** |

**Table 5: Results for NLI using either one teacher trained by all the data in MultiNLI or five teachers trained separately by data in each domain.**

| MultiNLI to SciTail | ACC | AUC |
|---|---|---|
| Base Model | 0.730 | 0.766 |
| KD (single teacher) | 0.751 | 0.805 |
| KD (multiple teachers) | 0.753 | 0.807 |
| DAKD (single teacher) | 0.755 | 0.811 |
| DAKD (multiple teachers) | **0.761**† | **0.818** |

**Table 6: Results for learning from large teacher models. Inf. time refers to the inference time for predicting a sentence pair.**

| Setting | PI | | NLI | | Inf. Time |
|---|---|---|---|---|---|
| | ACC | AUC | ACC | AUC | (ms) |
| DAM | 0.846 | 0.869 | 0.745 | 0.804 | 10 |
| DAKD (from DAM) | 0.867 | 0.881 | 0.755 | 0.811 | 12 |
| BERT fine-tuning | 0.892 | 0.894 | 0.905 | 0.966 | 200 |
| DAKD (from BERT) | 0.899 | 0.902 | 0.907 | 0.971 | 13 |

**Table 7: Ablation study on two parts of knowledge transfer for PI and NLI tasks.**

| Setting | PI | | NLI | |
|---|---|---|---|---|
| | ACC | AUC | ACC | AUC |
| Base-DAM | 0.846 | 0.869 | 0.730 | 0.766 |
| DAKD-DAM (whole) | 0.867 | 0.881 | 0.755 | 0.811 |
| w/o dark knowledge | 0.851 | 0.868 | 0.748 | 0.805 |
| w/o adaptive hints | 0.859 | 0.871 | 0.751 | 0.804 |

## 3.6 Industrial Deployment

We have deployed DAKD in our chatbot engine under three language scenarios, i.e., English, Russian, and Spanish. For each user query, the chatbot engine retrieves the top-30 relevant questions. Our method is employed to find the most relevant question and fetch its corresponding answer as the output. One naive way to score all the questions is to distribute them asynchronously to a couple of machines and compute all the scores. To speed up the online serving speed, we pad the user query with all the 30 candidate questions to form batch data and feed the batch to our method to obtain batch scores. This can be done via a single machine which shows to be more efficient. As a result, we are able to support a peak QPS of 40 on a cluster of 5 service machines on an Intel Xeon E5-2430 server.

**Table 8: A/B test results for industrial online deployment.**

| Method | English | Russian | Spanish |
|---|---|---|---|
| Baseline (DAM with TL) | 0.872 | 0.890 | 0.870 |
| DAKD (from BERT) | 0.916 | 0.971 | 0.944 |
| Relative improvement | +5.1% | +9.1% | +8.5% |

To examine the effectiveness of DAKD, we conduct an online A/B test to compare our method with the online production method, i.e., DAM with transfer learning from a source domain. The results are shown in Table 8. Our method has better performance, with relative improvements of 5.1% ~ 9.1% in all the three application scenarios. Meanwhile, the serving latency and the throughput of our method are nearly identical to the previous online method, as the student model of our method is also based on DAM.

## 3.7 Application Study Beyond QA Systems

To examine the generalization capability of DAKD, we evaluate our method on a review analysis task. The goal is to examine the quality score of a given review. Due to the high volume of reviews in E-commerce sites, this task has drawn increasing attention from both academia and industry [19, 41]. Experiments are performed on reviews from five categories of products in the Amazon review dataset [20]. Categories include "Electronics", "Watches", "Cellphones", "Outdoor" and "Home", each with #354,301, #9,737, #18,542, #72,796, #219,310 review samples, respectively. To make a fair comparison with baselines, we adopt TextCNN as the base model for the student and the teachers and follow exactly the data processing, base model hyper-parameters, and experiment setup as in [3]. All experiment results are evaluated in terms of *Pearson correlation coefficient*. As shown in Table 9, we first perform a transfer learning study on the single teacher setting, where the "Electronics" domain is chosen as the source domain, and the remaining domains are served as target domains. Overall we have similar findings as in PI and NLI datasets. Comparing with other transfer techniques, DAKD achieves the best performance. It is intuitive to see target domains with less data are likely to benefit more from the source domain. For example, on the smallest domain "Watch", our method improves the most.

Moreover, experiments have been conducted with multiple teachers, where the "Phone" domain is used as the target domain and the remaining four are treated as source domains. Again we observe

**Table 9: Review analysis results where "Electronics" is used as the source domain while the rest are treated as target domains.**

| Correlation | Watch | Phone | Outdoor | Home |
|---|---|---|---|---|
| Base Model | 0.423 | 0.552 | 0.511 | 0.521 |
| FS | 0.510 | 0.611 | 0.575 | 0.598 |
| SP | 0.504 | 0.607 | 0.576 | 0.595 |
| DAKD | **0.531** | **0.621** | **0.590** | **0.603** |

the setting of using multiple teachers outperforms using a single teacher, improving the model performance from 0.621 to 0.632. After reviewing the domain expertise scores $\lambda^m$ for different teachers, we find that "Electronics" and "Home" domains are more important than "Watch" and "Outdoor" domains, as they are with higher expertise scores (larger than 0.25). Closer examination shows "Electronics" and "Home" domains have a lot of reviews in electronic devices that are close to the domain "Phone". "Outdoor" domain is more on outdoor clothing and equipment, and the "Watch" domain is more on analog watches, both of which are different from the "Phone" domain. Hence, those domain expertise scores learned are reasonable and insightful.

## 4 RELATED WORK

Our work is closely related to several lines of research topics including knowledge distillation and transfer learning.

### 4.1 Knowledge Distillation

Knowledge Distillation (KD) was first proposed by [11], which aims to improve a smaller student model by distilling information from a pre-trained larger teacher model via a teacher-student optimization paradigm. The intuition behind this is to leverage the guidance of the teacher. It has been widely applied in model compression and knowledge transfer among the same task or different tasks in the same domain [11, 32, 34, 46]. However, few studies consider to learn from heterogeneous teachers from different domains. It was also proven to be useful when training a student network that has the same architecture as the teacher in which the student excels the teacher [6, 8]. Several attempts have been made to adjust ways for teacher supervision to improve its effectiveness. Most of the KD works focus on utilizing either the "dark knowledge", i.e., predicted outputs [5, 8, 11, 48] or hints, i.e., intermediate representations [28, 46, 48] of the teacher model. The majority of KD models are applied in computer vision. Recent improvements for distillation techniques consider adversarial training [10, 35] and feature distribution matching [13] and few works consider cross-task or cross-modal distillation for specific CV tasks by embedding alignment [9, 45, 50]. A recent study explores a two-stage multi-teacher distillation for QA system [44], however, for a specific downstream task, the student model does not leverage cross-domain knowledge from teachers. In this work, we consider both distillation and adaptation, where we propose to measure the domain expertise of cross-domain teachers on the student task and "adapt" the knowledge respectively.

### 4.2 Transfer Learning

Transfer Learning (TL) has been extensively studied to improve the model performance in data-deficient target domains by leveraging knowledge from related source domains [23, 36]. Recently, a large amount of neural network-based TL methods have been proposed [16, 21, 43]. A simple but effective framework is fine-tuning, i.e., to train a model on the source domain data and then use the learned weights as the initialization to perform continued training on the target domain data. Another widely used technique for TL is to consider a shared neural network to learn shared features for both source and target domains [2, 21], often referred to as a fully-shared model. However, this simple model may not be able to capture domain-specific features that are useful for boosting domain performance. To address this issue, the shared-private model is proposed [16], which consists of a shared network and domain-specific networks to learn domain-invariant and domain-specific features. Both fully-shared and share-specific types of models can be regarded as parameter sharing based methods, as they all jointly train a shared network to capture transferable knowledge in a multi-task fashion.

There is another line of TL methods that aims to align the hidden feature representations by explicitly reducing the marginal/conditional distribution divergence between source and target domains. Such methods are often called domain adaptation. Depending on label availability, it further divided into unsupervised and supervised domain adaptation. Regardless of the different setting categories, they all try to minimize the feature representation divergence by some distribution difference metrics. Commonly used metrics include variants of the Maximum Mean Discrepancy (MMD), the Kullback-Leibler Divergence and the Wasserstein distance [17, 18, 29, 33].

The main differences between DAKD and existing TL methods are two-fold. First, unlike typical TL methods that joint train transferable features in a multi-task fashion, our method utilizes a student-teacher paradigm. Second, those TL methods seldom utilize dark knowledge and mostly focus on transferable features only, while our method can adaptively leverage both dark knowledge and hints provided by the teacher models.

## 5 CONCLUSION

In this work, we provide a new framework named *Domain-Aware Knowledge Distillation* (DAKD), which enhances the teacher-student paradigm to facilitate cross-domain transfer learning, where teacher and student tasks belong to heterogeneous domains, with the goal to improve the student model performance of the target domain. Our framework considers both the "dark knowledge" from teacher models and adaptive hints to alleviate domain differences. Extensive experiments on two benchmark datasets show the proposed method has better performance than baselines.

We have also deployed our method in an online production system and observed significant improvements. To examine the generalization capability of our method, we further evaluate our model performance on a review analysis task. In the future, we seek to evaluate more tasks to further examine the generalization power of our method.

## ACKNOWLEDGMENTS

This work is being supported by Open Research Projects of Zhejiang lab (No. 2019KD0AD01/004). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

## REFERENCES

[1] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*. 632–642.

[2] Cen Chen, Minghui Qiu, Yinfei Yang, Jun Zhou, Jun Huang, Xiaolong Li, and Forrest Sheng Bao. 2019. Multi-domain gated cnn for review helpfulness prediction. In *The World Wide Web Conference*. 2630–2636.

[3] Cen Chen, Yinfei Yang, Jun Zhou, Xiaolong Li, and Forrest Sheng Bao. 2018. Cross-Domain Review Helpfulness Prediction Based on Convolutional Neural Networks with Auxiliary Domain Discriminators. In *NAACL*. 602–607.

[4] Daoyuan Chen, Yaliang Li, Minghui Qiu, Zhen Wang, Bofang Li, Bolin Ding, Hongbo Deng, Jun Huang, Wei Lin, and Jingren Zhou. 2020. AdaBERT: Task-Adaptive BERT Compression with Differentiable Neural Architecture Search. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 2463–2469.

[5] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. 2019. Online Knowledge Distillation with Diverse Peers. *CoRR* abs/1912.00350 (2019). arXiv:1912.00350

[6] Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D Manning, and Quoc V Le. 2019. Bam! born-again multi-task networks for natural language understanding. *arXiv preprint arXiv:1907.04829* (2019).

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.

[8] Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born-Again Neural Networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. 1602–1611.

[9] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. 2016. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2827–2836.

[10] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. 2018. Knowledge Distillation with Adversarial Samples Supporting Decision Boundary. *CoRR* abs/1805.05532 (2018).

[11] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. *CoRR* abs/1503.02531 (2015). arXiv:1503.02531

[12] Qiang Huang, Jianhui Bu, Weijian Xie, Shengwen Yang, Weijia Wu, and Liping Liu. 2019. Multi-task Sentence Encoding Model for Semantic Retrieval in Question Answering Systems. In *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*. 1–8.

[13] Zehao Huang and Naiyan Wang. 2017. Like What You Like: Knowledge Distill via Neuron Selectivity Transfer. *CoRR* abs/1707.01219 (2017).

[14] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[15] Wuwei Lan and Wei Xu. 2018. Neural Network Models for Paraphrase Identification, Semantic Textual Similarity, Natural Language Inference, and Question Answering. In *COLING 2018*. 3890–3902.

[16] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial Multi-task Learning for Text Classification. In *ACL*. 1–10.

[17] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. Learning Transferable Features with Deep Adaptation Networks. In *ICML*. 97–105.

[18] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. 2017. Deep Transfer Learning with Joint Adaptation Networks. In *ICML*. 2208–2217.

[19] Lionel Martin and Pearl Pu. 2014. Prediction of Helpful Reviews Using Emotions Extraction. In *AAAI*. 1551–1557.

[20] Julian J. McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *RecSys*. 165–172.

[21] Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How Transferable are Neural Networks in NLP Applications?. In *EMNLP*. 479–489.

[22] Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. 2018. A Survey of the Usages of Deep Learning in Natural Language Processing. *CoRR* abs/1807.10854 (2018). arXiv:1807.10854

[23] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 22, 10 (2010), 1345–1359.

[24] Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. In *EMNLP*. 2249–2255.

[25] Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. 2017. AliMe Chat: A Sequence to Sequence and Rerank based Chatbot Engine. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*. 498–503.

[26] Chen Qu, Feng Ji, Minghui Qiu, Liu Yang, Zhiyu Min, Haiqing Chen, Jun Huang, and W. Bruce Croft. 2019. Learning to Selectively Transfer: Reinforced Transfer Learning for Deep Text Matching. In *WSDM*. 699–707.

[27] Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-Driven Response Generation in Social Media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*. 583–593.

[28] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. FitNets: Hints for Thin Deep Nets. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[29] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. 2019. Beyond Sharing Weights for Deep Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 4 (2019), 801–814.

[30] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*. 1577–1586.

[31] Yiping Song, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, Dongyan Zhao, and Rui Yan. 2018. An Ensemble of Retrieval-Based and Generation-Based Human-Computer Conversation Systems. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. 4382–4388.

[32] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient Knowledge Distillation for BERT Model Compression. In *EMNLP*.

[33] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. 2018. A Survey on Deep Transfer Learning. In *ICANN*. 270–279.

[34] Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling Task-Specific Knowledge from BERT into Simple Neural Networks. *arXiv preprint arXiv:1903.12136* (2019).

[35] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2018. KDGAN: Knowledge Distillation with Generative Adversarial Networks. In *NeurIPS*. 783–794.

[36] Karl R. Weiss, Taghi M. Khoshgoftaar, and Dingding Wang. 2016. A survey of transfer learning. *J. Big Data* 3 (2016), 9.

[37] Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *NAACL*. 1112–1122.

[38] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*. 55–64.

[39] Rui Yan, Dongyan Zhao, and Weinan E. 2017. Joint Learning of Response Ranking and Next Utterance Suggestion in Human-Computer Conversation System. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*. 685–694.

[40] Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W. Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A Hybrid Retrieval-Generation Neural Conversation Model. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*. 1341–1350.

[41] Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Sheng Bao. 2015. Semantic Analysis and Helpfulness Prediction of Text for Online Product Reviews. In *ACL*. 38–44.

[42] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*. 5754–5764.

[43] Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

[44] Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, and Daxin Jiang. 2020. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 690–698.

[45] Han-Jia Ye, Su Lu, and De-Chuan Zhan. 2020. Distilling Cross-Task Knowledge via Relationship Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12396–12405.

[46] Junho Yim, Donggyu Joo, Ji-Hoon Bae, and Junmo Kim. 2017. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 7130–7138.

[47] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *NIPS*. 3320–3328.

[48] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Learning from Multiple Teacher Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*. 1285–1294.

[49] Jianfei Yu, Minghui Qiu, Jing Jiang, Jun Huang, Shuangyong Song, Wei Chu, and Haiqing Chen. 2018. Modelling Domain Relationships for Transfer Learning on Retrieval-based Question Answering Systems in E-commerce. In *WSDM*. 682–690.

[50] Mingkuan Yuan and Yuxin Peng. 2019. Ckd: Cross-task knowledge distillation for text-to-image synthesis. *IEEE Transactions on Multimedia* (2019).