Comparison Knowledge Translation for Generalizable Image Classification

Zunlei Feng\textsuperscript{1,4,5}, Tian Qiu\textsuperscript{1}, Sai Wu\textsuperscript{1}, Xiaotuan Jin\textsuperscript{3}, Zengliang He\textsuperscript{3}, Mingli Song\textsuperscript{1,4,5}, Huiqiong Wang\textsuperscript{2,*}

\textsuperscript{1}Zhejiang University
\textsuperscript{2}Ningbo Research Institute, Zhejiang University
\textsuperscript{3}Hangzhou Honghua Digital Technology Co., Ltd.
\textsuperscript{4}Shanghai Institute for Advanced Study of Zhejiang University
\textsuperscript{5}Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies
\{zunleifeng, huiqiong_wang\}@zju.edu.cn

Abstract

Deep learning has recently achieved remarkable performance in image classification tasks, which depends heavily on massive annotation. However, the classification mechanism of existing deep learning models seems to contrast to humans’ recognition mechanism. With only a glance at an image of the object even unknown type, humans can quickly and precisely find other same category objects from massive images, which benefits from daily recognition of various objects. In this paper, we attempt to build a generalizable framework that emulates the humans’ recognition mechanism in the image classification task, hoping to improve the classification performance on unseen categories with the support of annotations of other categories. Specifically, we investigate a new task termed Comparison Knowledge Translation (CKT). Given a set of fully labeled categories, CKT aims to translate the comparison knowledge learned from the labeled categories to a set of novel categories. To this end, we propose a Comparison Classification Translation Network (CCT-Net), which comprises a comparison classifier and a matching discriminator. The comparison classifier is designed to classify whether two images belong to the same category or not, while the matching discriminator works together in an adversarial manner to ensure whether classified results match the truth. Exhaustive experiments show that CCT-Net achieves surprising generalization ability on unseen categories and SOTA performance on target categories.

1 Introduction

In the past decade, deep learning has achieved remarkable performance in the image classification task. However, it usually costs a vast number of annotations to train a practical model in a real scenario. On the contrary, with only a glance at an image of the object even unknown type, humans can quickly and precisely find other same category objects from massive images. The underlying recognition mechanism of existing deep classifiers is different from humans’ recognition mechanism in the classification task.

In fact, humans’ quick and precise recognition ability on unknown-type objects benefits from the daily practices on objects of various known categories [Jolles et al., 2010]. This discovery raises an interesting and vital question: Can the existing deep classifiers quickly and precisely classify novel categories with the support of a set of fully labeled categories?

Some works such as zero/few-shot learning and transfer learning attempt training deep networks to handle novel categories with the help of a set of fully labeled categories. The former aims to train models using only a few annotated samples, while the latter focuses on transferring the models learned on one domain to another novel one. Despite the recent progress in few-shot and transfer learning, existing approaches are still prone to either inferior results [Lu et al., 2020], or the rigorous requirement that the two tasks are strongly related [Zhuang et al., 2020] and a large number of annotated samples [Dosovitskiy et al., 2021; Simonyan et al., 2014]. It seems that the recognition mechanism of the above two kinds of methods still has a difference from the humans’ recognition mechanism.

When taking an image of the object even unknown type, as a reference, humans can effortlessly find other same category objects from massive images. Inspired by this fact, we study a new Comparison Knowledge Translation Task (CKT-Task), aiming to translate the comparison knowledge learned from massive public source categories where abundant annotations are available, into novel target categories where a few number of annotations or even no annotation are available for each class. In this paper, comparison knowledge is defined as the recognition ability for distinguishing whether two images belong to the same category.

To this end, we propose a Comparison Classification Translation Network (CCT-Net) for the above CKT-Task. CCT-Net contains a comparison classifier and a matching discriminator, both of which comprise two branches that take a pair of images as input. What’s more, the matching discriminator has an additional similarity score as input. The comparison classifier is designed to classify whether two images belong
to the same category or not, while the matching discriminator
works together in an adversarial manner to ensure whether
classified results match the truth. The comparison classifi-
er only focuses on target categories; meanwhile, the adver-
sarial optimization between the comparison classifier and the
matching discriminator will translate the comparison knowl-
edge of source categories into the comparison classifier.

Experiments demonstrate that, with only tens of labeled
samples, the proposed CCT-Net achieves close performance
on par with fully supervised methods. When trained with ful-
ly annotated samples, CCT-Net achieves state-of-the-art per-
formance. The most surprising is that the proposed CCT-Net
shows promising generalization ability on novel categories.

Our contribution is therefore introducing a new CKT-Task,
in aim to translate the comparison knowledge learned from
fully annotated source categories into novel ones with few la-
bel, which emulates the humans’ recognition mechanism in
the image classification task. Furthermore, we propose a ded-
crated solution CCT-Net that comprises a comparison classi-
fier and a matching discriminator. The proposed CCT-Net
is evaluated on a broad domain of image datasets, which shows
that CCT-Net achieves SOTA classification performance and
surprise generalization on novel categories.

2 Related Work

To improve the performance (precision, convergence time,
and robustness) of deep classifiers with as few annota-
tions as possible, various classification tasks, including zero/few-
shot learning [Lu et al., 2020], transfer learning [Zhuang et
al., 2020], distillation learning [Gou et al., 2020], un/semi-
supervised learning [Chen et al., 2020], have attracted inter-
est from many researchers. In what follows, we review here
two lines of work that are closely related to ours, zero/few-
shot learning and transfer learning.

Zero/few-shot learning can be classified into three cat-
egories: mode-based, metric-based, and optimization-
based methods. Metric-based methods, including Siame-
seNet [Koch et al., 2015], Match Net [Vinyals et al., 2016],
Relation Net [Sung et al., 2018], and Prototype Net [Jake
et al., 2017] are most related to ours. SiameseNet [Koch et
al., 2015] is composed of weights shared twin CNNs, which ac-
cpt a pair of samples as inputs, and their outputs at the top
layer are combined in order to output a single pairwise simi-
larity score. Prototypical Net [Jake et al., 2017] classified
samples of new categories by comparing the Euclidean dis-
tance between the representation of the input sample with a
learnable class prototype. Match Net [Vinyals et al., 2016]
adopted cosine distance to measure similarity between two
representations that are encoded with two different encoder-
s. Unlike Prototypical Net and Matching Net, which use the
non-parametric Euclidean distance or cosine distance to
measure the similarity between pairwise features, Relation
Net [Sung et al., 2018] adopted a learnable CNN to measure
pairwise similarity, which takes the concatenation of feature
maps of two samples as input and outputs their relation score.
[Lu et al., 2020] summarized more variants about those net-
works.

Unlike the above methods, we adopt the adversarial man-
ner to distinguish whether the image pair matches with the
predicted similarity score. The major difference is that the
comparison knowledge of source categories is translated in-
to the target categories rather than adopting the annotations
to supervise the classifier’s training. What’s more, the cross-
attention mechanism throughout the whole classification pro-
cess is adopted to enhance the discriminant ability of the com-
parison classifier.

Transfer learning can be categorized into three types: in-
ductive, transductive, and unsupervised transfer learning. The
idea of inductive transfer learning, including multi-task learn-
ing algorithm [Simoes et al., 2018] and self-taught learning
[Niyaz et al., 2018], is to increase approximation of the target
probability distribution in the target domain given target tasks
are different from the source tasks. In the transductive trans-
fer learning technique [Rajesh and Manikanthan, 2017], the
source domain has a lot of labeled data, while the target do-
main has no labeled data. Both source tasks and target tasks
of transductive transfer learning are similar, whereas there is a
difference in the domain only. Unsupervised transfer learning
[Siddhart et al., 2018] is the same as inductive, but the main di-
dference is that there is no labeled data in both the source
and the target domains.

The differences between transfer learning and CCT-Net
contain two aspects: the knowledge type and the way of
knowledge transfer. Transfer learning transfers the discrim-
inanit ability from source categories into a classification net-
work for target categories, while CCT-Net translates compar-
ison classification ability for a pair image, which is a general
classification ability for both source and target categories. For
the knowledge transfer way, transfer learning adopts the fine-
tune and co-training strategies to transfer the classification a-
bility. The adversarial training strategy in CCT-Net is adopted
to translate the comparison classification ability, which brings
the advantage that redundant discriminant ability for source
categories will not distract the discriminant ability for target
categories. Furthermore, CCT-Net has superior generaliza-
tion ability on novel categories, as shown in Section 5.3 and
Table 2.

3 Comparison Knowledge Translation Task

Inspired by the general boundary knowledge [Feng et al.,
2021; Cheng et al., 2021] devised for the segmentation task,
we introduce the comparison knowledge, the recognition a-
bility to distinguish whether two images belong to the same
category. Then, the Comparison Knowledge Translation Task
(CKT-Task) is defined as follows. CKT-Task aims to learn a
generalizable image classification framework that can quickly
and precisely classify novel categories like humans. In CKT-
Task, there are assumed to be labeled source dataset $S^m$ that
contains $m$ object categories and target dataset $S^n$ of $n$ objec-
t categories. The $m$ categories and $n$ categories are disjoint.
CKT-Task is supposed to translate the comparison knowledge
of $S^m$ into the comparison classifier $F_0$, which is devised for
only concentrating on classifying of $n$ categories. The com-
parison classifier $F_0$ is expected to learn the generalizable
distinguishing ability.
4 Method

Deep learning methods usually require sufficient annotations to train a well-behaved classifier. There are vast public classification datasets with plenty of annotations, which have been exploited by transfer learning and distillation learning for improving classification performance on target domains or categories. However, the difference specificity of domains and categories severely limits classification performance’s upper bound. In CKT-Task, we study the comparison knowledge that is generalizable for different categories. To this end, we propose the Comparison Classification Translation Network (CCT-Net) to improve the classification performance on target categories, which draws on the generalizable knowledge from vast public datasets. Fig. 1 depicts the whole framework of CCT-Net comprising a comparison classifier and a matching discriminator. The adversarial training strategy translates the comparison knowledge of the matching discriminator learned from vast source categories into the comparison classifier focusing on targeted categories.

4.1 Comparison Classifier

In CCT-Net, the comparison classifier $\mathcal{F}_\theta$ is designed to be a two-branch architecture. Each branch is composed of multiple encoders. The comparison classifier only focuses on the dataset $\mathcal{S}^n$ of $n$ target categories. The input of the comparison classifier is a pair of images, which have two cases similar and non-similar, as shown in the left part in Fig. 1.

With a target image pair $(I^t_1, I^t_2) \in \mathcal{S}^m$ as input, two branches learn a pair of representations $(r^1_t, r^2_t)$. Then, a fully connected layer predicts the similarity score $p(r^1_t - r^2_t)$ with the representation difference $r^1_t - r^2_t$ as input.

The two branches of the comparison classifier $\mathcal{F}_\theta$ share the same architecture but have different parameters. Inspired by the humans’ recognition mechanism that distinguishes the image pair from global to local, cross attention is introduced to compare features of two images at each layer. The comparison classifier will compare basic, middle-level, and high-level semantic features as layers go deeper.

4.2 Matching Discriminator

As shown in Fig. 1, the matching discriminator has the same architecture as the comparison classifier. Unlike the comparison classifier, the input of the matching discriminator contains two parts: a pair of images $(I^t_1, I^t_2)$ and a similarity score $p \in \{0, 1\}$. The matching discriminator is expected to distinguish whether the input image pair $(I^t_1, I^t_2)$ matches with the similarity score $p$.

The input image pairs of the matching discriminator include both the source dataset $\mathcal{S}^m$ and the target dataset $\mathcal{S}^n$. Given a target image pair $(I^t_1, I^t_2) \in \mathcal{S}^m$ and the similarity score $p(r^1_t - r^2_t)$ predicted by the comparison classifier $\mathcal{F}_\theta$, the matching discriminator first learns a pair of representations $(r^1_s, r^2_s)$. Then, the representation difference $r^1_s - r^2_s$ concatenated with a similarity vector $[p(r^1_t - r^1_s), p(r^1_t - r^2_s), ..., p(r^1_t - r^2_s)]$ is input into the fully connected layer of the matching discriminator, which will discriminate whether the similarity of the image pair $(I^t_1, I^t_2)$ matches with the predicted similarity score $p(r^1_t - r^2_t)$.

For the image pair $(I^t_1, I^t_2)$ from the source dataset $\mathcal{S}^m$, it is assigned with a similarity score $p^*$ (assigning 0 or 1 in this paper) and annotated with the matching condition $c^*$ (True or False, set as 1 or 0 in the code). Four kinds of input cases for the source category are summarized in Fig. 1. Given the source image pair $(I^s_1, I^s_2) \in \mathcal{S}^s$ and the assigned similarity score $p^*$, the matching discriminator first learn a pair of representations $(r^1_s, r^2_s)$. Then, the representation difference $r^1_s - r^2_s$ concatenated with the similarity vector $[p^*, p^*, ..., p^*]$ is input into the fully connected layer of the matching discriminator. The annotated matching condition $c^*$ for the triplet $(I^t_1, I^t_2, p^*)$ will supervise the matching discriminator $D_\phi$ to learn the matching discrimination ability with the following...
In the experiments, we adopt five datasets, including the source and target categories in the comparison experiments. The proposed task needs the disjoint source and target categories. So, categories of each dataset are evenly split into the source and target categories in the comparison experiments.

### Network architecture and parameter setting

For each encoder branch of the comparison classifier and the matching discriminator, ViT-B/16 [Dosovitskiy et al., 2021] is adopted as the backbone. Each encoder of CCT-Net comprises 12 attention heads, where 2 attention heads are used for cross-attention between two sub-branches in each layer. The fully connected layer of the comparison classifier and the matching discriminator share the same architecture (linear layer: 4096, LeakyReLU, linear layer: 1024, linear layer: 256). The length of the similarity vector is 1024, and the representation difference \( d_t^{1/s} - d_t^{2/s} \) is 768. More details are given in the supplements.

### 5.1 Comparing with SOTA Methods

In this section, the proposed method is compared with unsupervised methods: SCAN [Gansbeke et al., 2020] and SimCLR [Chen et al., 2020], few-shot methods: PrototypeNet [Jake et al., 2017] and Simple CNAPS [Bateni et al., 2020], semi-supervised methods: MixMatch [David et al., 2019] and FixMatch [Sohn et al., 2020], transfer learning methods: Transfer(10) and Transfer(100), and fully supervised methods: SiameseNet [Koch et al., 2015], VGG-16 [Simonyan et al., 2014], ResNet-50 [He et al., 2016], MobileNetV2 [Sandler et al., 2018], DenseNet-121 [Huang et al., 2017], VIT-B/16 [Dosovitskiy et al., 2021], and mini-ImageNet [Jia et al., 2009], to verify the effectiveness of the proposed CKT-Task and CKT-Net. The proposed task needs the disjoint source and target categories. So, categories of each dataset are evenly split into the source and target categories in the comparison experiments.

### Dataset

The experiments, we adopt five datasets, including MNIST [LeCun et al., 1998], CIFAR-10 [Krizhevsky, 2009], STL-10 [Adam et al., 2011], Oxford-IIIT Pets [Parkhi et al., 2012], and mini-ImageNet [Jia et al., 2009], to verify the effectiveness of the proposed CKT-Task and CKT-Net. The proposed task needs the disjoint source and target categories. So, categories of each dataset are evenly split into the source and target categories in the comparison experiments.
5.2 Translation between Different Dataset

This section provides the knowledge translation experiments between different datasets to verify the robustness of CKT-Task and CCT-Net. For all the translation tasks (Source → Target) in Table 3, only 20 annotated pairs of the target categories are used for CCT-Net. In the experiment, different datasets and different numbers of source and target categories are taken as two ablation factors. From Table 3, we can see that all the translation tasks achieve satisfactory results between different datasets with different source and target category numbers, which verifies the high extensibility and practicability of CKT-Task and CCT-Net.

5.3 Generalization Ability on Novel Category

To further verify the generalization ability of CKT-Task, the trained models are directly tested on novel categories (the trained models have never seen before). Table 2 shows the generalization results on three datasets. ‘dataset1 → dataset2’ denotes CCT-Net is only trained on the dataset1 and then tested on unseen dataset2. CCTadv and CCTcross denote CCT-Net without discriminator and cross attention, respectively. So, we only compare the proposed method with SiameseNet, which has the same two-branch architecture. ‘SiameseNet(all)’ denotes that SiameseNet is trained with fully annotated samples of dataset1 and then tested on dataset2.

From Table 2, we can see that CCT-Net(all) achieves about double increment than SiameseNet(all) on all datasets (CIFAR-10: (+47.6, +55.4), STL-10: (+16.6, +28.0), mini-ImageNet: (+28.2, +26.2), mini-ImageNet → STL-10: (+44.4, +47.3)), which demonstrates the surprising generalization ability of CKT-Task and CCT-Net on novel categories. Even without annotations of target categories, CCT-Net(0) still achieves better generalization ability than SiameseNet(all). What’s more, we further verify the importance of each component of CCT-Net on generalization ability. With annotated samples of target categories, CCTcross (CCT-Net without cross attention) achieves lower classification accuracy than CCTadv (CCT-Net without discriminator), which demonstrates that cross attention mechanism contributes more to the generalization ability of CCT-Net.

5.4 Convergence Speed on Novel Category

Except for the excellent generalization ability, CKT-Task and CCT-Net still have the advantage of fast convergence speed, shown in Fig. 2. Due to the different number of model parameters, we adopt ‘Time / parameters’ (training time (seconds) per 1M parameter) to compare the convergence speed of different methods on novel categories. From Fig. 2, we can see that the proposed CCT-Net have a faster convergence speed on novel category than existing methods, which indicates that CCT-Net imitates the humans’ quick recognition ability on novel category.
To verify the effectiveness of CCT-Net’s components, we do ablation study on the false pair samples (case 3 and case 4 in Fig. 1), discriminator, cross attention, input position of the condition $c^{l/s}$, and the different numbers of labeled samples of target categories. For the ablation study, the mini-ImageNet dataset is adopted. For $CCT_{\text{false}}$, $CCT_{\text{adv}}$, $CCT_{\text{cross}}$ and $CCT_{\text{head}}$, 20 labeled pairs for each target category are used. From Table 5, we can see that CCT-Net(20) achieves higher scores than other ablative methods, demonstrating all components’ effectiveness. $CCT_{\text{adv}}$ achieves the lowest score than other ablative methods, which indicates that the adversarial translation strategy is the most critical factor for CKT-Task and CCT-Net.

5.6 Ablation Study

Table 5: The ablation study results of CCT-Net. $CCT_{\text{false}}$, $CCT_{\text{adv}}$, $CCT_{\text{cross}}$, and $CCT_{\text{head}}$ denote CCT-Net without false pair samples, discriminator, cross attention, respectively. $CCT_{\text{cond}}$ denotes placing the similarity score in the head of the discriminator.

The ablation study results on annotated pairs of the target category are given in Table 4, where we can find that 50-annotated-pairs is a critical cut-off point, supplying relatively sufficient guidance. Even without any guidance of labeled pairs of target categories, CCT-Net still achieves 81.4% accuracy score on target categories, verifying the effectiveness of CKT-Task and CCT-Net again.

6 Conclusion

This paper studies a new Comparison Knowledge Translation Task (CKT-Task), which imitates humans’ quick and precise recognition ability on novel categories. The goal of CKT-Task is to translate the comparison knowledge of source categories into the deep classifier for new categories in the least effort and dependable way. Toward realizing CKT-Task, we introduce the Comparison Classification Translation Network (CCT-Net), which comprises a comparison classifier and a matching discriminator. The comparison classifier is devised to classify whether two images belong to the same category or not, while the matching discriminator works together in an adversarial manner to ensure whether classified results match the truth. With the adversarial training between the comparison classifier and the matching discriminator, the comparison knowledge of massive public source categories is successfully translated into the deep classifier for target categories. There is no special requirement for the source and target categories, which means that all public classification datasets can be used as the source datasets in the proposed CCT-Net. Exhaustive experiments show that CCT-Net achieves impressive generalization ability and SOTA performance on par with existing methods. Surprisingly, CCT-Net also achieves impressive results, including fast convergence speed and high accuracy on novel categories, revealing its superior generalization ability. We will focus on exploring more generalization knowledge and framework on other tasks in the future.
Acknowledgments

This work is supported by National Natural Science Foundation of China (No.62002318), Key Research and Development Program of Zhejiang Province (2020C01023), Zhejiang Provincial Science and Technology Project for Public Welfare (LG21F020020), Fundamental Research Funds for the Central Universities (2021FZZX001-23), Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (Grant No. SN-ZJU-SIAS-001), and Zhejiang Lab (No.2019KD0AD01/014).

References


