On the Importance of Distractors for Few-Shot Classification

Rajshekhar Das\(^1\) Yu-Xiong Wang\(^2\) José M.F. Moura\(^1\)
\(^1\)Carnegie Mellon University \(^2\)University of Illinois at Urbana-Champaign
rajshekd@andrew.cmu.edu yxw@illinois.edu moura@andrew.cmu.edu

Abstract

Few-shot classification aims at classifying categories of a novel task by learning from just a few (typically, 1 to 5) labelled examples. An effective approach to few-shot classification involves a prior model trained on a large-sample base domain, which is then finetuned over the novel few-shot task to yield generalizable representations. However, task-specific finetuning is prone to overfitting due to the lack of enough training examples. To alleviate this issue, we propose a new finetuning approach based on contrastive learning that reuses unlabelled examples from the base domain in the form of distractors. Unlike the nature of unlabelled data used in prior works, distractors belong to classes that do not overlap with the novel categories. We demonstrate for the first time that inclusion of such distractors can significantly boost few-shot generalization. Our technical novelty includes a stochastic pairing of examples sharing the same category in the few-shot task and a weighting term that controls the relative influence of task-specific negatives and distractors. An important aspect of our finetuning objective is that it is agnostic to distractor labels and hence applicable to various base domain settings. More precisely, compared to state-of-the-art approaches, our method shows accuracy gains of up to 12% in cross-domain and up to 5% in unsupervised prior-learning settings. Our code is available at https://github.com/quantacode/Contrastive-Finetuning.git

1. Introduction

The ability to learn from very few examples is innate to human intelligence. In contrast, large amounts of labelled examples are required by modern machine learning algorithms to learn a new task. This limits their applicability to domains where data is either expensive to annotate and collect or simply inaccessible due to privacy concerns. To overcome this limitation, few-shot classification has been proposed as a generic framework for learning to classify with very limited supervision [12, 31, 34, 59]. Under this paradigm, most approaches leverage prior knowledge from a (labelled) base domain to solve a novel task by either finetuning-based transfer [9, 64] or meta-learning [12, 15, 49, 57, 59, 62, 69]. In particular, when the base and novel domains are related, the hope is that representations learnt in the base domain can be generalized to novel tasks, thus facilitating positive knowledge transfer.

While the above paradigm is effective for tasks that can leverage large datasets like ImageNet [52] as the related base domain, for others, such as rare species classification [68] or medical image classification [70], acquiring necessary prior knowledge can be exceedingly difficult due to the absence of a related base domain with labelled data. To relax such data requirements, recent techniques explore alternative ways such as unsupervised learning [25, 28] or cross-domain learning [1, 15, 45, 66] to obtain representations useful for novel tasks. In the absence of labelled base data, approaches like [25, 27, 28] seek to benefit from self-supervised representation learning over unlabelled data in a related domain. In a more challenging scenario where related base data is hard to obtain, cross-domain techniques [10, 64, 66] exploit representations learnt in other domains.
that do not have the same task characteristics as the novel tasks.

Although the issue of learning a good prior representation remains a core focus in few-shot classification, it addresses only a part of the problem. In this work, we investigate the other important aspect, i.e., effective finetuning specific to the novel task. Our main motivation comes from recent findings [1, 9, 16] that demonstrate the outperformance of simple finetuning over more sophisticated prior learning techniques such as meta-learning. Despite its effectiveness, we suspect that finetuning might still suffer from overfitting as a consequence of small training set in a few-shot task. To alleviate this situation, we propose to leverage additional unlabelled data exclusive to the task. Such datapoints are referred to as distractors. For instance, in the case of classifying Honduran Emeralds and Kakapos (rare species of birds), examples of butterflies, cars or ducks can serve as distractors (Fig. 1). By the virtue of its task-exclusivity, distractors can be obtained from various data-abundant domains with categories that could be semantically unrelated to novel task categories. However, in this work, we restrict ourselves to just the base data as a source for distractors. This allows us to efficiently reuse the data under standard settings and directly compare with prior works.

To this end, we pose the imminent question – Can distractors improve few-shot generalization? The answer is, somewhat surprisingly, yes. To elucidate how, we propose ConFT, a simple finetuning method based on a contrastive loss that contrasts pairs of the same class against those from different classes. We show that with a few simple but crucial modifications to the standard contrastive loss, distractors can be incorporated to boost generalization. We hypothesize that in the absence of extensive in-domain supervision for prior experience, distractor-aware finetuning can yield non-trivial gains. Towards the design of the loss function, we adopt an asymmetric construction of similarity pairs to ensure that distractors contribute only through different-class pairs. Our key insight here is two-fold – 1) generalization in contrastive learning can be influenced by not only same-class but also different-class pairs; 2) construction of different-class pairs is extremely flexible in that it can include samples from task-specific as well as task-exclusive categories. As a test of generality, we study the effect of our finetuning approach in conjunction with two different prior learning setups, namely, cross-domain and unsupervised prior learning. Our contributions are as follows.

- We propose contrastive finetuning, ConFT, a novel finetuning method for transfer based few-shot classification.
- We show how distractors can be incorporated in a contrastive objective to improve few-shot generalization.
- The proposed method outperforms state-of-the-art approaches by up to 12 points in the cross-domain few-shot learning and up to 5 points in unsupervised prior learning settings.

2. Related Work

2.1. Few-Shot Classification

Modern algorithms for few-shot classification are predominantly based on meta-learning where the goal is to quickly adapt to novel tasks. These approaches can be broadly classified into three categories: initialization based [12, 38, 39, 50, 53], hallucination based [2, 20, 71], and metric-learning based [3, 31, 57, 59, 62, 69] methods. Despite the growing interest in sophisticated meta-learning techniques, recent works [1, 6, 9, 64] have demonstrated that even simple finetuning based transfer learning [13, 17, 32, 44, 76] can outperform them. Such baselines usually involve cross-entropy training over the base categories followed by finetuning over a disjoint set of novel classes. Following these results, we further the investigation of finetuning for few-shot classification.

Cross-Domain Few-Shot Classification: A number of recent works [1, 11, 16, 40, 43, 54, 66, 67] have been proposed to address the cross-domain setup where base and novel classes are not only disjoint but also belong to different domains. Interestingly, [6] demonstrated that in this setup too, finetuning based transfer approaches outperformed popular meta-learning methods by significant margins. Following that, [66] proposed to learn feature-wise transformations via meta-learning to improve few-shot generalization of metric-based approaches. While in standard finetuning, the embedding model is usually frozen to avoid overfitting, recent works like [1, 16] have shown that frozen embeddings can hinder few-shot generalization. In this work, we build upon these developments to propose a more effective finetuning method over the entire embedding model.

In the context of learning from heterogeneous domain, [65] introduced a benchmark for multi-domain few-shot classification. This benchmark has been adopted by some recent works [7, 10, 37, 55]. While multiple base domains can alleviate cross-domain learning, we test our approach on a more challenging setup [66] that only involves a single base domain. Recent works used [66] as a benchmark to evaluate the importance of representation change [40] and spatial contrastive learning [43] in cross-domain few-shot classification. Another related work [67] leveraged unlabelled data from the novel domain in addition to few-shot labelled data to improve the task performance in a similar benchmark [16]. In contrast to [67], we operate under a limited access to novel domain data, i.e., only the few-shot labelled data.

Unlabelled Data in Few-Shot Classification: Our use
of unlabelled data in the form of distractors is inspired from
cognitive neuroscience studies [36] describing the effect of
visual distractors on learning and memory. Prior works that
use additional unlabelled data for few-shot classification in-
clude [4, 14, 35, 51, 61, 72]. Complementary to [4, 14, 61] that
exploit unlabelled data via self-supervised objectives in the
prior learning phase, we use unlabelled data specifically for
task-specific finetuning. Nonetheless, combining both per-
spectives could yield further benefits and is left for future
work.

More related approaches [35, 51] combined heteroge-
neous unlabelled data, i.e., task-specific data and distrac-
tors, in a semi-supervised framework. Our distractor-aware
finetuning differs from these works in two important ways:
our few-shot classification is strictly inductive in that we do
not use unlabelled data specific to the task, and our method
leverages distractors instead of treating them as interfer-
ence that needs to be masked out. The most relevant meth-
ods [1, 13], like us, reused the base (or source) domain as a
source for additional data. The key difference, however, is
that their success relies on effective alignment of the base
and novel classes, whereas we benefit from contrasting the
two. While the importance of distractor-aware learning has
been investigated in the context of object detection [46, 78],
their benefit to few-shot generalization has not been studied
before.

Recently, [25, 27, 28] have studied few-shot classification
in the context of unsupervised prior-learning where the base
data is unlabelled. In this work, we evaluate the benefit of
contrastive finetuning under this setting and and compare it
to existing methods.

2.2. Contrastive Learning

Contrastive learning yields a similarity distribution over
data by comparing pairs of different samples [58]. Recently,
contrastive learning [18, 19, 56, 60] based methods have emerged
as the state of the art for supervised [26, 29, 74] and self-supervised
[5, 21, 23, 24, 42, 63, 75] representation learning. While the supervised approaches primarily ex-
plot ground-truth labels to construct same-class pairs, self-
supervised techniques leverage domain knowledge in the
form of data augmentation to generate such pairs. As a spe-
cial case, [29] maximized the benefit by integrating both
forms of contrastive losses into a single objective. In this
work, we use a modified version of the supervised con-
trastive loss when more than one labelled example is avail-
able per category. However, in the extreme case of 1-
shot classification, it switches to self-supervised contrastive
learning. Recent works such as [10, 43] also explored contrastive learning in the context of few-shot classification.
While they use contrastive objectives at the prior-learning
stage to learn a general-purpose representation solely on the
base domain, our method uses a contrastive objective
at finetuning to improve the downstream-task-specific rep-
resentation directly on the target domain task with base do-
main data as distractors. As a design choice, we adopt the
contrastive loss over other losses like cross-entropy since it
allows us to leverage distractor data that does not belong to
the novel categories but improves generalization.

3. Our Approach

To achieve the goal of few-shot generalization, our con-
trastive finetuning method, ConFT, optimizes for two simul-
taneous objectives. First, it aims to bring task-specific sam-
ple that share the same class close to each other; and sec-
ond, it strives to push apart samples that belong to different
classes. This two-fold objective can lead to compact clus-
ters that are well separated amongst each other. In the fol-
lowing sections, we first introduce some notations that we
then use to formally describe our approach. An overview of
our method is presented in Fig. 2.

3.1. Preliminaries

Consider an input space \( \mathcal{X} \) and a categorical label set
\( \mathcal{Y} = \{c_1, \ldots, c_M \} \) where each of the \( M \) classes is re-
presented via one-hot encoding. A representation space
\( \mathcal{R} \subset \mathbb{R}^r \) of the input is defined by the composition of an
augmentation function \( A : \mathcal{X} \rightarrow \mathcal{X} \) and a representation
model \( M_\theta : \mathcal{X} \rightarrow \mathcal{R} \), parameterized by \( \theta \). The augmen-
tation function is a composition of standard image transfor-
mations such as random cropping, color jittering, horizontal
flipping etc. Given a small number \( K \), a few-shot classification
\( \tau \) can be defined as the collection of a support set,
\( \tau_{\text{supp}} = \{(x_i, y_i) | x_i \in \mathcal{X}, y_i \in \mathcal{Y}, i \in I_{\text{supp}} \} \) with \( K \) exam-
ple per class, and a query set, \( \tau_q = \{\tilde{x}_j | \tilde{x}_j \in \mathcal{X}, j \in I_q \} \)
sampled from the same (but unobserved) classes. Here,
\( I_{\text{supp}} \) and \( I_q \) are the collection of indices for the support and
query sets, respectively. The few-shot classification goal is to
leverage the support set to obtain a classifier for the query
samples. In this case, the classifier is constructed over the
representation model obtained via contrastive finetuning of
a prior model, \( M_{\theta_0} \) over \( \tau_{\text{supp}} \).

3.2. The ConFT Objective

A key component of the ConFT objective is that it in-
cludes unlabelled samples, distractors, to improve few-shot
generalization. Formally, a distractor set, \( S_d = \{x_i | x_i \in \mathcal{X}, i \in I_d \} \), drawn from a domain \( D : \mathcal{X} \times \mathcal{Y}_D \) together
with the task-specific support set \( \tau_{\text{supp}} \), constitutes the train-
ing data for few-shot learning. Here, the distractor class
set \( \mathcal{Y}_D \) is assumed to be task-exclusive, \( \mathcal{Y}_D \cap \mathcal{Y} = \emptyset \). Start-
ing with a support set example \( i \) (a.k.a anchor), we
first construct an anchor-negative index set, \( N(i) = \{p \in I_{\text{supp}} | y_i \neq y_p \} \), and an anchor-positive index set \( P(i) \) such
that \( y_p = y_i, \forall p \in P(i) \). Samples indexed by \( N(i) \) are
treated as negatives within the task, whereas those indexed
Figure 2. **Contrastive Finetuning in Cross-Domain Few-Shot Learning:** Our contrastive finetuning approach to few-shot classification comprises of two stages: 1) The prior learning stage trains a representation model on labelled (under cross-domain settings) base data using a cross-entropy loss; 2) The pretrained representation model is then finetuned over task-specific samples as well as distractors using a contrastive loss. For each task sample \(z_i\), the contrastive objective (right) maximizes a similarity score, \(\text{sim}\), over same-class pairs while minimizing it over other pair types. In the absence of enough labelled examples, distractors can improve classification by pushing apart task-specific clusters (here, different classes of birds).

The representation obtained from transfer learning is used to represent the query samples. The objective of contrastive finetuning is to increase the similarity between the anchor and positive samples while decreasing the similarity with negative samples. The contrastive objective is defined as:

\[
\mathcal{L}_{\text{conf}}(\theta) = -\frac{1}{|\mathcal{I}_{\text{supp}}|} \sum_{i \in \mathcal{I}_{\text{supp}}} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log l_{ip},
\]

where \(\gamma\) is a temperature hyper-parameter. The finetuning objective is simply the minimization of \(\mathcal{L}_{\text{conf}}\) to yield optimal parameters \(\theta\) specific to task \(\tau\). To classify the query samples, we construct a nearest-mean classifier [11, 47, 58] atop the updated representation \(\mathcal{M}_{\theta_{\tau}}\). The class-specific weight vectors are computed as an average over the representations of \(K\) support examples pertaining to that class. The \(j^{th}\) query sample is then assigned to the class whose weight vector has the largest cosine similarity (and hence, nearest in the Euclidean sense) with the query representation. We use the accuracy of this classifier to compare various baselines in the experiment section.

The contrastive finetuning approach to few-shot classification defines our contrastive loss that uses a contrastive loss. For each task sample \(z_i\), the contrastive objective (right) maximizes a similarity score, \(\text{sim}\), over same-class pairs while minimizing it over other pair types. In the absence of enough labelled examples, distractors can improve classification by pushing apart task-specific clusters (here, different classes of birds).

In the case when the number of shots is odd, we omit one sample from each class to allow even pairing. The omission is, however, not an issue in the overall scheme of finetuning where multiple steps of gradient descent optimization ensures that eventually each sample gets to participate with equal chance. In the special case 1-shot learning, anchornegative sets are constructed similar to [5] using augmentation \(A\).

### 3.3. Relative Importance of Anchor-Negatives

Given the loss formulation of (1), both task-specific (few-shot) and task-exclusive (distractor) anchor-negatives influence the loss proportionate to their respective batch sizes. While the batch size of task-specific negatives \(N(i)\) is upper bounded by the number of ways \(M\) and the number of shots \(K\), the batch size of distractors can be made as large as that of the domain itself, \(i.e., |D|\). In standard contrastive learning paradigms with only task-specific and no task-exclusive training examples, large batch sizes of negatives are known to be beneficial for downstream task performance. However, in our case where both types of negatives exist, naively increasing distractor batch size can be counterproductive (shown in the supplementary). We suspect that too many distractors might overshadow the effect of task-specific negatives that can be more crucial for generalization. Also, the effect might vary according to the proximity of distractors with respect to task samples in the representation space. Nonetheless, there is a need to balance the undue influence of distractors by adjusting the batch size.
sizes. To avoid an extensive search for an optimal batch size specific to the distractor domain, we propose a domain-agnostic weighting scheme for the anchor-negatives proportional to their batch sizes as follows

$$l_{ip} = \frac{\exp(\frac{z_i z_p}{\gamma})}{\exp(\frac{z_i z_p}{\gamma}) + \alpha \sum_{n \in N(i)} \exp(\frac{z_i z_n}{\gamma}) + (2 - \alpha) \sum_{d \in I_d} \exp(\frac{z_i z_d}{\gamma})},$$

$$\alpha = 2 \frac{|I_d|}{|N(i)| + |I_d|}$$

We found that this simple weighting scheme makes the few-shot performance robust to batch size variations and also improves the overall performance (see the supplementary).

4. Prior Learning and Distractors

Thus far we have assumed the access to a distractor set, $S_d$ and a prior model $M_d$. In this section, we describe how to obtain them and how distractors boost generalization. Recall that our goal is to achieve few-shot generalization by finetuning a prior model over the few-shot task. However, due to the scarcity of task-specific labelled examples, a reasonably strong prior encoded in the model parameters $\theta_0$ is crucial for preventing overfitting, especially when using high-capacity models like neural networks. We next describe two different ways of learning such a prior that can serve as a good initialization for subsequent finetuning.

4.1. Types of Prior Learning

Cross-Domain Learning: In the cross-domain setup, we are provided a labelled dataset, $D_t = \{(x_i, y_i) | x_i \in \mathcal{X}_c, y_i \in \mathcal{Y}_c\}^{\mid D_t \mid}$ drawn from a source domain $\mathcal{X}_c \times \mathcal{Y}_c$, such that the categorical label set $\mathcal{Y}_c$ is disjoint from novel categories $\mathcal{Y}$. The key characteristic of this setup is that the distribution of $M$-way $K$-shot tasks, if constructed out of $D_t$, will be significantly different from novel tasks in the target domain. Such distribution shift could arise due to difference in task granularity (e.g., coarse-grained vs. fine-grained) or shift in input distribution or both. In this work, we consider the case where the shift in task granularity is notably more than the input distribution. Towards the goal of learning a reasonably strong prior, we adopt a simple objective that minimizes cross-entropy loss over all categories in $D_t$. During finetuning, the distractors are sampled from $D_t$, thus, naturally satisfying the non-overlapping categories assumption with respect to novel tasks.

Unsupervised Prior Learning: For unsupervised prior learning, we are given an unlabelled dataset, $D_u = \{v_i\}^{\mid D_u \mid}$ drawn from a source domain $\mathcal{X}_u \times \mathcal{Y}_u$, such that the corresponding labels in $\mathcal{Y}_u$ are unobserved. While there are no explicit assumptions about the task distribution gap in this setting, the strength of the learnt prior is likely to be more reasonable when the distribution gap is small. To learn a suitable prior using $D_u$, we use the SimCLR loss [5] as a form of self-supervised objective. Our choice of this objective over others [21] was based on its superior performance found in our preliminary experiments. Priors learnt via self-supervised contrastive objectives on large base datasets (like, ImageNet) have been shown to transfer well to many-shot downstream tasks. In this work we show that such objectives are effective even with smaller base datasets and few-shot downstream tasks. In our experiments, priors learnt in this way already outperform state-of-the-art approaches [25,27,28] that are then further improved by our proposed contrastive finetuning. In this setup, we use $D_u$ as the source for distractors where the assumption of non-overlapping categories is satisfied with high probability, provided the base dataset is relatively large and encapsulates a wide variety of categorical concepts.

4.2. Distractor-Aware Generalization

The most important and perhaps surprising aspect of our method is that distractors, despite being drawn from unrelated (to novel task) categories, can improve generalization. To understand the underlying mechanism, we propose to measure the change in quality of task-specific representation before and after finetuning. Particularly, given a few-shot task with $M$ classes, we define the subset of query samples, $I_q^c \subset I_q$ that share the same class and two other quantities – cluster spread $u^q_{\text{spread}}$ and cluster-separation $u^q_{\text{sep}}$ that measure the degree of clustering in the representation space. Specifically,

$$u^q_{\text{spread}}(\theta_t) = \frac{1}{M} \sum_{m=1}^{M} \sum_{i \in I_q^m \setminus \{t\}} \left(1 - z_i \cdot z_j\right),$$

$$u^q_{\text{sep}}(\theta_t) = \frac{1}{M} \sum_{m=1}^{M} \sum_{i \in I_q^m \setminus \{t\}} \left(1 - z_i \cdot z_j\right),$$

where $\theta_t$ are the parameters of the representation model after $t$ finetuning epochs. For each of the above quantities, we define the change, $\delta^q(t) = u^q_{\theta_t}(\theta_t) - u^q_{\theta_0}(\theta_0)$, and relative change, $\delta^q_{\text{rel}}(t) = \frac{\delta^q(t)}{\kappa(\theta_0)}$ where, the subscript can be $\text{sep}$ or $\text{spread}$ and division by a fixed value, $\kappa(\theta_0)$ ensures scale invariance. Finally, to quantify generalization within a given target domain, we define the average relative change, $E_{\mathcal{T}}[\delta^q_{\text{rel}}(t)]$ over a large number of tasks sampled from that domain. The average relative change can also be defined for

\footnote{Note that the query class labels are considered only for analysis purposes. In practice, they are not observed.}
4.3. A Multitask Variant of ConFT

While our original objective \( (2) \) is agnostic to distractor supervision, finetuning in the cross-domain setting can further benefit from distractor labels. To that end, we introduce an auxiliary loss \( \ell_{\text{mice}} \) during finetuning that minimizes the cross-entropy between predicted probabilities and one-hot encodings of the ground-truth label averaged over the base data, \( D_i \). This leads to a new multitask formulation

\[
\ell_{\text{mt-conf}} = \ell_{\text{conf}} + \lambda \ell_{\text{mice}},
\]

where we fix the relative weighting factor \( \lambda = 1 \) in our experiments and use a cosine classifier [6] for \( \ell_{\text{mice}} \). We found that this simple extension led to significant performance gains in some domains while marginal in others, depending on domain characteristics.

5. Experiments

Following sections first introduce some baselines (§5.1) and present our main results for contrastive finetuning in the cross-domain setup (§5.2). Then, §5.3 elucidates the generalization mechanism of ConFT followed by ablations in §5.4. Finally, §5.5 demonstrates the performance of our approach in the unsupervised prior learning setup.

Datasets and Benchmarks: We evaluate our proposed finetuning method in a variety of novel domains spanning across two different paradigms for prior learning. For cross-domain evaluations, we adopt the benchmark introduced by [66] that comprises of Cars [33], CUB [73], Places [77], and Plantae [68] as the novel domains and miniImageNet [49] as the base domain. Each dataset is split into \textit{train}, \textit{val} and, \textit{test} categories (please refer to the supplementary for details), where tasks sampled from the \textit{test} split are used to evaluate the few-shot performance in respective domains. We use the \textit{val} splits for cross-validating the hyperparameters and the train split of miniImageNet as our base data. For experiments in unsupervised prior learning, we use the same train split of miniImageNet to learn a self-supervised representation that is then evaluated for few-shot performance on miniImageNet-test. We present additional results on Meta-Dataset [65] in the supplementary.

Backbone (Representation Model): Following best practices in cross-domain few-shot learning, we adopt a ResNet10 [22] model for most of our experiments. In the unsupervised learning case, we use a four-layer CNN consistent with existing works except for a reduced filter size from 64 to 20 in the final layer. This modification was found to improve contrastive finetuning performance.

Optimization and Hyperparameters: In this work, we evaluate few-shot performance over \( 5 \)-way \( 1 \)-shot and \( 5 \)-way \( 5 \)-shot tasks with 15 query samples, irrespective of the prior learning setup. For the contrastive finetuning, we use an ADAM [30] optimizer with a suitable learning rate and early-stopping criteria. Our proposed method has a few hyperparameters such as the temperature (\( \gamma \)), learning rate, early-stopping criteria, and data augmentation (\( \mathcal{A} \)). However, recent studies [41] have highlighted that excessive hyperparameter tuning on large validation sets can lead to overoptimistic results in limited-labelled data settings like semi-supervised learning. Thus, we keep an extremely small budget for hyperparameter tuning. Among the mentioned hyperparameter, the one with the most number of parameters is the augmentation function \( \mathcal{A} \). In this work, we do not tune \( \mathcal{A} \) to any specific target domain. Instead, we use a fixed augmentation scheme introduced by [6] for the cross-domain setting and AutoAugment [8] for the unsupervised prior learning case. Please refer to the supplementary for a detailed summary of hyperparameters used in our experiments.

5.1. Baseline Comparisons

We begin our evaluations by comparing various baselines for finetuning in Table 1. These include two simple baselines (introduced in [6]) and two strong baselines (introduced in [16]). While the simple baselines freeze the backbones, the others allow finetuning over the entire embedding model. Another key difference is that the simple baselines are evaluated using standard linear evaluation [5, 6], whereas the rest are evaluated using nearest-mean classifiers. We compare the performance of all these baselines to our vanilla and multi-task (MT) versions of ConFT. Following previous works, the learning rates for the simple baselines are kept at 0.01, whereas for others (including ours), we use smaller learning rates (0.005 or 0.0005). We observe that among the baselines, the cosine classifier based baseline, FT-all (CC), outperforms the linear classifier based FT-all (LC). However, both versions of our finetuning approach significantly outperform all baselines across various dataset and shot settings.

5.2. ConFT for Cross-Domain Prior Learning

In this section, we present our main results on cross-domain few-shot learning (see Table 2). We compare our approach with various prior works on the LFT benchmark [66]. We observe that overall our proposed approaches, ConFT and MT-ConFT, significantly outperform the best previous results in Cars (by \( 3 \) to \( 12 \) points), Plantae (by \( 7 \) to \( 9 \) points) and CUB (by \( 1.7 \) to \( 2 \) points) domains. We also observe higher gains in the 5-shot setting than the 1-shot case, since more labelled examples can improve few-shot generalization. Further, we find that using the auxiliary support examples by simply swapping superscript ‘q’ with ‘s’. Also, in practice, we use \( u^{s}_{q}(\theta_0) \) as the fixed value for \( \kappa(\theta_0) \) irrespective of the superscript or subscript.
loss (MT-ConFT) is more beneficial in the 5-shot case. In fact, it performs worse than ConFT in the 1-shot cases for Cars, Places, and Plantae. Such a degradation could be due to a misalignment between the self-supervised objective (to which ConFT boils down in the 1-shot case) and the auxiliary cross-entropy loss. In the Places domain, “GNN w/ featTx” yields the best performance, whereas our approach outperforms the rest for the 5-shot case. We suspect that the use of a more sophisticated model in “GNN w/ featTx,” namely, graph neural net [57] built on top of a ResNet-10 model, leads to a better cross-domain generalization when the domain gap is smaller.

5.3. Effect of Distractors on Generalization

In this section, we investigate the central question – How do distractors improve generalization? We present two sets of plots in Figure 3 that track the change in cluster-spread and cluster-separation as the finetuning progresses. In the first set, we plot the average relative change, $\Delta [\delta_{rels}(t)]$ and $\Delta [\delta_{rel}(t)]$ (see §4.2) as a function of finetuning epochs, $t$ for both support and query samples in 2 different settings, namely, CUB (5-shot) and Cars (5-shot). We observe that for support examples (yellow lines), cluster-spread decreases with increasing epochs while the cluster-separation increases. This is indeed what is expected for training datapoints (here, support examples) and serves as a sanity check. For query samples (green plot), on the other hand, both cluster-spread and separation increase with the progress in finetuning epochs. The key observation, however, is that cluster-separation increases to a much greater extent than the cluster-spread, thus improving overall discriminability between classes represented by these clusters.
Table 4. Ablation 2. Novel task performance with varying sizes of the distractor domain, i.e., miniImageNet-train. Note that this is different from distractor batch size $|S_d|$. 

Table 5. Unsupervised Prior Learning. The results are averaged over 1000 novel tasks and are presented with 95% confidence intervals. Here, MFT refers to meta-style finetuning [12].

5.5. ConFT for Unsupervised Prior Learning

In Table 5, we demonstrate the generality of contrastive finetuning by evaluating on the unsupervised prior learning benchmark miniImageNet. The key distinction from cross-domain settings is that we do not have labelled base data to learn from. So, we leverage self-supervised contrastive learning [5] on the unlabelled base data and show that it outperforms state of the art by 1 to 2 points. Finetuning the resultant representation with our ConFT objective further improves the accuracy by 2 to 4 points. This is particularly significant, as the results come very close to supervised baselines that serve as performance upper bound in this setting [28].

6. Conclusion

We introduce a novel contrastive finetuning approach to few-shot classification. Specifically, our method leverages distractors to improve generalization by encouraging cluster separation of the novel task samples. We show that our method leads to significant performance gains in both cross-domain and unsupervised prior learning settings.

While the increase in cluster-separation hints towards the possible reason behind improved generalization, it is not clear how much of the improvement, if any, is a consequence of incorporating distractors. To delineate the effects of distractors from the contrastive loss itself, we present the second set of plots that compare the average relative change in query cluster spread and separation under the presence (red line) and absence (blue line) of distractors for the same data settings. We observe that with increasing finetuning epochs the gap between cluster-separation of only query samples under the presence and absence of distractors. The spread and separation quantities are averaged over 600 tasks for both top and bottom.


Table 5. Unsupervised Prior Learning. The results are averaged over 1000 novel tasks and are presented with 95% confidence intervals. Here, MFT refers to meta-style finetuning [12].

5.5. ConFT for Unsupervised Prior Learning

In Table 5, we demonstrate the generality of contrastive finetuning by evaluating on the unsupervised prior learning benchmark miniImageNet. The key distinction from cross-domain settings is that we do not have labelled base data to learn from. So, we leverage self-supervised contrastive learning [5] on the unlabelled base data and show that it outperforms state of the art by 1 to 2 points. Finetuning the resultant representation with our ConFT objective further improves the accuracy by 2 to 4 points. This is particularly significant, as the results come very close to supervised baselines that serve as performance upper bound in this setting [28].

6. Conclusion

We introduce a novel contrastive finetuning approach to few-shot classification. Specifically, our method leverages distractors to improve generalization by encouraging cluster separation of the novel task samples. We show that our method leads to significant performance gains in both cross-domain and unsupervised prior learning setups.

While the increase in cluster-separation hints towards the possible reason behind improved generalization, it is not clear how much of the improvement, if any, is a consequence of incorporating distractors. To delineate the effects of distractors from the contrastive loss itself, we present the second set of plots that compare the average relative change in query cluster spread and separation under the presence (red line) and absence (blue line) of distractors for the same data settings. We observe that with increasing finetuning epochs the gap between cluster-separation of only query samples under the presence and absence of distractors. The spread and separation quantities are averaged over 600 tasks for both top and bottom.


Table 5. Unsupervised Prior Learning. The results are averaged over 1000 novel tasks and are presented with 95% confidence intervals. Here, MFT refers to meta-style finetuning [12].

5.5. ConFT for Unsupervised Prior Learning

In Table 5, we demonstrate the generality of contrastive finetuning by evaluating on the unsupervised prior learning benchmark miniImageNet. The key distinction from cross-domain settings is that we do not have labelled base data to learn from. So, we leverage self-supervised contrastive learning [5] on the unlabelled base data and show that it outperforms state of the art by 1 to 2 points. Finetuning the resultant representation with our ConFT objective further improves the accuracy by 2 to 4 points. This is particularly significant, as the results come very close to supervised baselines that serve as performance upper bound in this setting [28].

6. Conclusion

We introduce a novel contrastive finetuning approach to few-shot classification. Specifically, our method leverages distractors to improve generalization by encouraging cluster separation of the novel task samples. We show that our method leads to significant performance gains in both cross-domain and unsupervised prior learning setups.

While the increase in cluster-separation hints towards the possible reason behind improved generalization, it is not clear how much of the improvement, if any, is a consequence of incorporating distractors. To delineate the effects of distractors from the contrastive loss itself, we present the second set of plots that compare the average relative change in query cluster spread and separation under the presence (red line) and absence (blue line) of distractors for the same data settings. We observe that with increasing finetuning epochs the gap between cluster-separation of only query samples under the presence and absence of distractors. The spread and separation quantities are averaged over 600 tasks for both top and bottom.


Table 5. Unsupervised Prior Learning. The results are averaged over 1000 novel tasks and are presented with 95% confidence intervals. Here, MFT refers to meta-style finetuning [12].

5.5. ConFT for Unsupervised Prior Learning

In Table 5, we demonstrate the generality of contrastive finetuning by evaluating on the unsupervised prior learning benchmark miniImageNet. The key distinction from cross-domain settings is that we do not have labelled base data to learn from. So, we leverage self-supervised contrastive learning [5] on the unlabelled base data and show that it outperforms state of the art by 1 to 2 points. Finetuning the resultant representation with our ConFT objective further improves the accuracy by 2 to 4 points. This is particularly significant, as the results come very close to supervised baselines that serve as performance upper bound in this setting [28].

6. Conclusion

We introduce a novel contrastive finetuning approach to few-shot classification. Specifically, our method leverages distractors to improve generalization by encouraging cluster separation of the novel task samples. We show that our method leads to significant performance gains in both cross-domain and unsupervised prior learning setups.
References


[32] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? CVPR, 2019. 2

[33] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In ICCV Workshops, 2013. 6


of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, 2018. 3


[38] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In ICML, 2017. 2


[56] Ruslan Salakhutdinov and Geoff Hinton. Learning a non-linear embedding by preserving class neighbourhood structure. In AISTATS, 2007. 3


[64] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? arXiv, 2020. 1, 2


[68] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In CVPR, 2016. 1, 6


[71] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In CVPR, 2018. 2


