A Broad Study on the Transferability of Visual Representations with Contrastive Learning

Ashraful Islam\textsuperscript{1,*}, Chun-Fu Chen\textsuperscript{2,3}, Rameswar Panda\textsuperscript{2,3}, Leonid Karlinsky\textsuperscript{3}, Richard Radke\textsuperscript{1}, Rogerio Feris\textsuperscript{2,3}

\textsuperscript{1}Rensselaer Polytechnic Institute  \; \textsuperscript{2}MIT-IBM Watson AI Lab  \; \textsuperscript{3}IBM Research

islama6@rpi.edu, chenrich@us.ibm.com, rpanda@ibm.com
leonidka@il.ibm.com, rjradke@ecse.rpi.edu, rsferis@us.ibm.com

Abstract

Tremendous progress has been made in visual representation learning, notably with the recent success of self-supervised contrastive learning methods. Supervised contrastive learning has also been shown to outperform its cross-entropy counterparts by leveraging labels for choosing where to contrast. However, there has been little work to explore the transfer capability of contrastive learning to a different domain. In this paper, we conduct a comprehensive study on the transferability of learned representations of different contrastive approaches for linear evaluation, full-network transfer, and few-shot recognition on 12 downstream datasets from different domains, and object detection tasks on MSCOCO and VOC0712. The results show that the contrastive approaches learn representations that are easily transferable to a different downstream task. We further observe that the joint objective of self-supervised contrastive loss with cross-entropy/supervised-contrastive loss leads to better transferability of these models over their supervised counterparts. Our analysis reveals that the representations learned from the contrastive approaches contain more low/mid-level semantics than cross-entropy models, which enables them to quickly adapt to a new task. Our codes and models will be publicly available to facilitate future research on transferability of visual representations.\textsuperscript{1}

1. Introduction

Self-supervised learning is an important research area whose goal is to learn superior data representations without any labelled supervision. Recently, self-supervised contrastive learning has shown promising results in image classification tasks [21, 7, 4]. In the contrastive learning paradigm, a model is trained to recognize different augmentations of the same image (commonly referred as positives) while discriminating them from other random images (referred as negatives) in the dataset. The promising performance of self-supervised contrastive learning led to the idea of leveraging label information in the contrastive learning paradigm. To this end, Khosla et al. [30] proposed a supervised contrastive learning framework that achieves better ImageNet accuracy than the standard cross-entropy model.

Representations learned from contrastive learning have been shown to perform better than supervised cross-entropy models in various downstream tasks, particularly the object detection task [21, 7, 30, 45, 54]. Despite recent progress, it is unclear why contrastive representations transfer better to other tasks, since most prior work focuses on in-domain evaluation, particularly ImageNet classification accuracy. In this paper, our goal is to understand the underlying mechanism of the superior transferability of contrastive learning. Towards this end, we conduct a comprehensive study regarding transfer learning of contrastive approaches on downstream image classification, few-shot evaluation, and object detection. We rigorously benchmark five methods with different training objective losses: cross-entropy, self-supervised contrastive, supervised con-

\footnotesize\textsuperscript{*}This work was done while the author was an intern at IBM.
\footnotesize\textsuperscript{1}https://github.com/asrafulashiq/transfer_broad

Figure 1: Average top-1 accuracy of different models on the downstream datasets. (a) Linear evaluation with a fixed feature extractor and (b) 5-way 5-shot few-shot classification. In both cases, we observe that contrastive pretrained models achieve superior performance compared to cross-entropy pretrained models. Adding a self-supervised contrastive loss (SelfSupCon) improves the performance for both supervised cross-entropy and supervised contrastive pretrained models. We argue that incorporating a self-supervised contrastive loss (SelfSupCon) increases the variability within the same-class features and makes the models learn both high-level semantics and low-level cues.
trastive, joint cross-entropy/self-supervised contrastive, and joint supervised/self-supervised contrastive.

We first compare the transfer performance of different ImageNet pretrained models on a collection of 12 downstream datasets from various domains. We find that contrastive methods perform much better than the supervised cross-entropy models, particularly in fixed feature transfer learning; however, the performance gap becomes smaller after full-network fine-tuning. We observe similar trends on other downstream tasks, including few-shot recognition and object detection and instance segmentation on the VOC0712 [15] and MS COCO [36] datasets. In particular, our results indicate that the joint objective of self-supervised contrastive loss and supervised cross-entropy/contrastive loss consistently outperforms the standard trained counterparts in different downstream tasks. Figure 1 shows the average top-1 accuracy of the different ImageNet pretrained methods we studied on the downstream datasets, for both fixed-feature linear evaluation and few-shot classification. Both the self-supervised contrastive model (denoted SelfSupCon) and supervised contrastive model (denoted SupCon) perform better than the cross-entropy model (denoted CE). Moreover, the combination of cross-entropy and self-supervised contrastive (denoted CE+SelfSupCon) performs better than cross-entropy or self-supervised contrastive alone. The same goes for the combination of self-supervised contrastive and supervised contrastive (denoted SupCon+SelfSupCon).

We next investigate why contrastive approaches show superior transferability by analyzing the similarity between hidden representations, intra-class separation, and robustness to image corruption. We find that contrastive approaches learn more low-level and mid-level information that can be easily adapted to a different domain than the supervised cross-entropy models, which mostly learns high-level semantics in the penultimate layers. Zhao et al. [54] hypothesized that one of the limiting factors of supervised cross-entropy models is the objective of minimizing intra-class variation. Our analysis also suggests that a model should have sufficient intra-class variation in the source domain to better transfer the learned representations to a different domain. Most standard supervised loss functions aim to increase inter-class distance and decrease intra-class variation, which might be harmful for transferability of features. We infer that contrastive approaches have larger within-class separation than the standard cross-entropy models, which could be one of the factors underlying their superior transferability. We also analyze the robustness and calibration of different models, and find that contrastive losses are more robust to different image corruptions and predict well-calibrated class probabilities that are more representative of true correctness likelihoods than cross-entropy models. Our key contributions in this work are as follows:

• We benchmark five methods including cross-entropy, self-supervised contrastive, supervised contrastive, and their combinations on downstream image classification, object detection, and few-shot recognition. All results show a similar trend that contrastive learning extracts better features for transfer learning.

• We show that combining supervised loss with self-supervised contrastive loss improves transfer learning performance. Specifically, learned representation from the joint objective of self-supervised contrastive and supervised contrastive loss significantly outperforms the model trained with cross-entropy by 5.63% under linear evaluation protocol and 3.46% in few-shot recognition (5-shot) on the 12 downstream datasets, 1.37% AP50 under object detection on VOC0712, and ~0.8% on MS COCO. The improvement of the joint objective over supervised contrastive model is small but consistent across all downstream tasks. The joint objective of cross-entropy and self-supervised contrastive loss also consistently performs better than the models trained with the individual objectives.

• We apply Centered Kernel Alignment (CKA) [33] and show that contrastive models contain more low-level and mid-level information in the penultimate layers than standard cross-entropy models. Furthermore, our analysis suggests that the contrastive models have higher intra-class variation than the standard cross-entropy models, even if the network is not explicitly trained to increase intra-class distance.

2. Related Work

Transfer Learning. Early results on transfer learning showed that convolutional neural networks (CNNs) trained on large-scale datasets could be used to extract features to train SVMs and logistic regression models that outperformed hand-crafted feature-based approaches [6, 14, 44]. Transfer learning can be a powerful tool to train a significantly smaller dataset than the base dataset without overfitting. However, the factors driving the performance are still not completely understood. Huh et al. [27] investigated the effect of the source dataset on transfer learning. Simon et al. [34] found that pretrained models with higher ImageNet accuracy also tend to perform well in the downstream task. Azizpour et al. [2] investigated the effect of network depth on the transfer performance. In this work, we show that contrastive training can improve transfer learning performance, and investigate the underlying principle behind it.

Self-Supervised and Supervised Contrastive Learning. Earlier work on self-supervised learning generated pseudo labels by patch position [13], image colorization [33], image inpainting [43], rotation [17], predictive coding [23, 23] and other pretext tasks. Recently, contrastive learning
has led to significant performance enhancement in self-supervised image representation learning. In particular, MoCo [21], SimCLR [7], SwAV [4], and others have shown dramatic improvement in representation quality learned from unlabeled ImageNet images. Khosla et al. [30] proposed a new contrastive loss to leverage the label information. Moreover, representations learned from contrastive learning have been shown to perform better than supervised cross-entropy models in various downstream tasks [52, 21, 7, 30, 29, 31, 28]. However, most of the studies perform limited comparison, particularly with regard to fixed-feature transfer, few-shot learning and robustness, and the underlying principle of why contrastive learning transfers better still remains unclear.

3. Analysis Setup

Given a source domain $D_s = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ with a marginal distribution $P_S$ and a target domain $D_T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ with a marginal distribution $P_T$, where $(x_i, y_i)$ is the image-label pair, and, in general, $P_S \neq P_T$, the objective of transfer learning is to learn a target prediction function $f_T(\cdot)$ using the knowledge of $D_S$. We study various target prediction tasks, namely, linear evaluation over fixed network for image classification, full-network fine-tuning for image classification, object detection, and few-shot image classification tasks.

3.1. Loss Functions

Supervised Cross-Entropy Loss. Supervised cross-entropy loss [3] is the standard loss function for multi-class classification. Given an input image $x$ and one-hot encoded target label $y$, denote the output representation of the encoder network as $v = f_0(x)$. The class logits are calculated as $l = Wv + b$, where $W \in \mathbb{R}^{K \times D}$ contains the weights and $b \in \mathbb{R}^K$ is the bias of the final linear layer. The supervised cross-entropy loss is defined as:

$$L_{CE}(l, y) = - \sum_{i=1}^{K} y_i \log \left( \frac{\exp(l_i)}{\sum_{j=1}^{K} \exp(l_j)} \right)$$

Self-Supervised and Supervised Contrastive Loss. In the contrastive learning paradigm, the network is trained by distinguishing between similar and dissimilar instances. We use Momentum Contrast (MoCo) [21], particularly MoCov2 [8], for studying the efficacy of contrastive representations for transfer learning. The encoder $f_0(\cdot)$ of MoCo is a convolutional neural network (CNN), followed by a multi-layer perceptron (MLP) head to embed the encoded features in a contrastive subspace. MoCo has two base networks; one is actively trained to extract query features, and the other is the moving average of the query encoder to extract positive and negative features (commonly known as keys). Denote the query as $q$ and the set of keys in the queue as $\{k_1, k_2, \ldots, k_M\}$. Assuming the key denoted by $k_+$ matches with the query $q$, the objective of MoCo is:

$$L_{SelfSupCon}(q) = - \log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{j=1}^{M} \exp(q \cdot k_j / \tau)},$$

where $M$ is the queue size. In self-supervised contrastive loss, the positive key $k_+$ is obtained from the augmented view of the same input image.

To leverage label information in contrastive learning, we follow the loss function from [30], where the positives are sampled from the contrastive features of the same classes as the query. The supervised contrastive loss is:

$$L_{SupCon}(q) = \sum_{1 \leq j \leq M} \log \frac{\exp(q \cdot k_j / \tau)}{\sum_{j'=1}^{M} \exp(q \cdot k_{j'} / \tau)}.$$
wise mentioned, we use top-1 accuracy as the evaluation
metric. When performing transfer learning to the down-
stream datasets, we created a separate validation set from
the training set, and swept the hyperparameters (learning
rate, batch size, and weight decay) for each dataset. Then
we use the optimal hyperparameters to train the full training
set (train+val), and evaluate on the test set. We also perform
multiple runs with different random seeds, and report mean
score among different runs. See Appendix B for detailed
results with confidence intervals over multiple runs.

4.3. Linear Evaluation over Fixed Network

We first use the linear evaluation over fixed network to
test the learned visual representations. For all methods,
we freeze the backbone, only train a linear layer on top of
the backbone, and optimize the cross-entropy loss with the
SGD optimizer. The learning rate, weight decay, and batch
size are selected by hyperparameter tuning on the validation
set. Figure 3 shows the performance of different models
on the 12 downstream datasets in terms of top-1 accuracy
(averaged over 5 runs). Note that our reproduced cross-
entropy trained model performs slightly differently than the
pretrained ResNet-50 model from PyTorch [42].

We note that the cross-entropy model performs the worst
as a fixed feature extractor, both in terms of the accu-
racy for an individual dataset and the final average across
all datasets. All of the contrastive approaches perform
better than cross-entropy. We also find that training the
cross-entropy model with strong augmentation as MoCo
does not help much either. The best performing model
SupCon+SelfSupCon performs, on average, 5.63% better than CE. For datasets that are different than Im-
ageNet, e.g., SVHN, Sketch, Omniglot, and DTD, CE
performs much worse than contrastive models. We in-
fier that features learned from the cross-entropy model are
not directly helpful for datasets that are much different
than ImageNet, while features learned from contrastive
approaches are applicable to the datasets from different
domains. Moreover, CE+SelfSupCon performs better than
CE or SelfSupCon and SupCon+SelfSupCon performs better than the individual SupCon or SelfSupCon model, suggesting that self-supervised contrastive learning improves the transferability of supervised cross-entropy and supervised contrastive learning.

4.4. Full-Network Fine-Tuning

We further fine-tune the full network to study the transferability of all the methods. Here we include the results of image classification and object detection.

Image Classification. We fine-tune the pretrained models along with the final linear header on downstream datasets. We perform different hyperparameter sweeping on the validation set and report the scores on the test set for the optimal hyperparameters. Table 2 shows the top-1 accuracy of all datasets for full-network fine-tuning for 5 runs. While CE performs much worse than other contrastive models in the linear evaluation experiments, we did not observe similar behavior for full-network fine-tuning. SupCon+SelfSupCon achieves 88.56% top-1 accuracy, just 0.43% better than CE, which achieves 88.13% top-1 accuracy. We also report the performance of fine-tuning with only 1000 training samples, where CE+SelfSupCon performs 1.09% better than CE, and SupCon+SelfSupCon performs 1.48% better than CE. We infer that contrastive pretrained methods are slightly more effective in a limited data regime than cross-entropy based models. However, when we have a sufficient amount of data, all models achieve similar performance.

Object Detection and Instance Segmentation. We conduct experiments on object detection and instance segmentation on VOC0712 [15] and MS COCO [36] to validate the learned representations from different models based on Detectron2 [51]. We follow the settings in [21] to fine-tune the whole network while only training a few epochs (1× schedule) in the Detectron2 setting. The results are shown in Table 3. SupCon provides slightly better results than CE and SelfSupCon. Furthermore, CE+SelfSupCon and SupCon+SelfSupCon achieved consistent improvement on AP over their counterparts by ~0.8% and ~0.5% on MS COCO, respectively. Again, the results echo our previous observations in linear evaluation experiments.

4.5. Few-shot Classification

For few-shot learning, we use the ResNet-18 backbone, and pretrain all models on the Mini-ImageNet dataset. As suggested by [47], a model providing good embedding is essential for few-shot learning; thus, we simply trained a logistic regression classifier on top of the fixed network for the few-shot classification. Table 4 shows the average top-1 accuracy of 600 episodes for 5-way 5-shot and 20-shot experiments. We observe a similar trend that contrastive approaches consistently perform better than cross-entropy across all downstream datasets. CE+SelfSupCon and SupCon+SelfSupCon achieve the best scores, which suggests that self-supervised contrastive learning improves upon both supervised cross-entropy and supervised contrastive learning. On the other hand, when performing in-

<table>
<thead>
<tr>
<th>Full Dataset</th>
<th>CropDisease</th>
<th>DeepWeeds</th>
<th>Flowers102</th>
<th>EuroSAT</th>
<th>Resisc45</th>
<th>ISIC</th>
<th>ChestX</th>
<th>Omniglot</th>
<th>SVHN</th>
<th>Kaokore</th>
<th>Sketch</th>
<th>DTD</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>99.92</td>
<td>97.37</td>
<td>93.47</td>
<td>98.78</td>
<td>96.03</td>
<td>88.06</td>
<td>55.67</td>
<td>90.34</td>
<td>97.03</td>
<td>88.13</td>
<td>79.49</td>
<td>73.26</td>
<td>88.13</td>
</tr>
<tr>
<td>SelfSupCon</td>
<td>99.91</td>
<td>97.39</td>
<td>93.36</td>
<td>98.85</td>
<td>96.28</td>
<td>88.13</td>
<td>56.41</td>
<td>91.10</td>
<td>97.25</td>
<td>88.92</td>
<td>77.16</td>
<td>75.49</td>
<td>88.35</td>
</tr>
<tr>
<td>SupCon</td>
<td>99.91</td>
<td>96.89</td>
<td>95.37</td>
<td>98.67</td>
<td>96.03</td>
<td>87.92</td>
<td>55.08</td>
<td>90.56</td>
<td>97.11</td>
<td>87.88</td>
<td>80.14</td>
<td>74.19</td>
<td>88.31</td>
</tr>
<tr>
<td>CE+SelfSupCon</td>
<td>99.88</td>
<td>97.28</td>
<td>95.30</td>
<td>98.91</td>
<td>96.18</td>
<td>88.29</td>
<td>54.77</td>
<td>90.20</td>
<td>97.14</td>
<td>88.25</td>
<td>80.85</td>
<td>74.12</td>
<td>88.43</td>
</tr>
<tr>
<td>SupCon+SelfSupCon</td>
<td>99.91</td>
<td>97.38</td>
<td>96.17</td>
<td>98.75</td>
<td>96.21</td>
<td>88.55</td>
<td>55.12</td>
<td>90.86</td>
<td>97.04</td>
<td>87.73</td>
<td>80.17</td>
<td>74.86</td>
<td>88.56</td>
</tr>
</tbody>
</table>

Table 2: Performance of different models on the downstream datasets in terms of top-1 accuracy (%) (averaged over 5 runs) for full-network fine-tuning. Contrastive pretrained methods are slightly more effective in a limited data regime than cross-entropy based models.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>VOC0712 FasterRCNN-R50-C4</th>
<th>MaskRCNN-R50-C4</th>
<th>MS COCO (Trained with 1× schedule) MaskRCNN-R50-FPN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detectors</td>
<td>Methods</td>
<td>AP (mAP)</td>
<td>AP (mAP)</td>
</tr>
<tr>
<td>CE</td>
<td>FasterRCNN-R50-C4</td>
<td>81.58</td>
<td>54.63</td>
</tr>
<tr>
<td>SelfSupCon</td>
<td>FasterRCNN-R50-C4</td>
<td>82.48</td>
<td>57.33</td>
</tr>
<tr>
<td>SupCon</td>
<td>FasterRCNN-R50-C4</td>
<td>82.64</td>
<td>56.12</td>
</tr>
<tr>
<td>CE+SelfSupCon</td>
<td>FasterRCNN-R50-C4</td>
<td>83.06</td>
<td>57.07</td>
</tr>
<tr>
<td>SupCon+SelfSupCon</td>
<td>FasterRCNN-R50-C4</td>
<td>82.95</td>
<td>57.26</td>
</tr>
</tbody>
</table>

The results of VOC0712 is the average of 5 runs. AP (Detectors): AP of object detection; AP (Detectors): AP of instance segmentation.

Table 3: Object detection and instance segmentation results on VOC0712 and MS COCO (averaged over 5 runs).
<table>
<thead>
<tr>
<th></th>
<th>Mini-IN*</th>
<th>CropDisease</th>
<th>DeepWeeds</th>
<th>Flowers102</th>
<th>EuroSAT</th>
<th>Resisc45</th>
<th>ISIC</th>
<th>ChestX</th>
<th>Omniglot</th>
<th>SVHN</th>
<th>Kadoore</th>
<th>Sketch</th>
<th>DTD</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CE</td>
<td>72.47</td>
<td>86.58</td>
<td>48.33</td>
<td>81.31</td>
<td>78.51</td>
<td>72.86</td>
<td>44.28</td>
<td>26.16</td>
<td>94.26</td>
<td>37.63</td>
<td>64.64</td>
<td>57.64</td>
<td>60.01</td>
<td></td>
</tr>
<tr>
<td>SelfSupCon</td>
<td>67.71</td>
<td>83.29</td>
<td>49.00</td>
<td>84.29</td>
<td>81.65</td>
<td>72.36</td>
<td>45.20</td>
<td>26.01</td>
<td>97.60</td>
<td>34.70</td>
<td>62.75</td>
<td>51.00</td>
<td>61.35</td>
<td></td>
</tr>
<tr>
<td>SupCon</td>
<td>75.20</td>
<td>83.44</td>
<td>47.74</td>
<td>82.93</td>
<td>80.94</td>
<td>74.48</td>
<td>42.97</td>
<td>26.23</td>
<td>96.78</td>
<td>33.22</td>
<td>54.10</td>
<td>45.10</td>
<td>62.79</td>
<td></td>
</tr>
<tr>
<td>CE+SelfSupCon</td>
<td>76.13</td>
<td>84.68</td>
<td>50.02</td>
<td>86.88</td>
<td>82.63</td>
<td>75.11</td>
<td>44.66</td>
<td>27.93</td>
<td>96.19</td>
<td>31.36</td>
<td>43.52</td>
<td>37.63</td>
<td>63.70</td>
<td></td>
</tr>
<tr>
<td>SupCon+SelfSupCon</td>
<td>72.81</td>
<td>84.26</td>
<td>50.35</td>
<td>86.72</td>
<td>82.57</td>
<td>74.94</td>
<td>45.82</td>
<td>28.19</td>
<td>96.72</td>
<td>30.67</td>
<td>45.26</td>
<td>37.63</td>
<td>63.47</td>
<td></td>
</tr>
<tr>
<td>20-shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CE</td>
<td>80.81</td>
<td>92.51</td>
<td>58.43</td>
<td>89.00</td>
<td>84.77</td>
<td>82.08</td>
<td>52.88</td>
<td>29.86</td>
<td>97.76</td>
<td>35.69</td>
<td>46.87</td>
<td>67.85</td>
<td>67.84</td>
<td></td>
</tr>
<tr>
<td>SelfSupCon</td>
<td>76.95</td>
<td>90.42</td>
<td>58.99</td>
<td>90.77</td>
<td>87.72</td>
<td>82.20</td>
<td>53.57</td>
<td>30.61</td>
<td>97.10</td>
<td>34.70</td>
<td>47.75</td>
<td>66.32</td>
<td>70.71</td>
<td></td>
</tr>
<tr>
<td>SupCon</td>
<td>83.32</td>
<td>90.73</td>
<td>55.90</td>
<td>90.45</td>
<td>87.71</td>
<td>83.40</td>
<td>51.11</td>
<td>31.06</td>
<td>98.87</td>
<td>44.98</td>
<td>54.94</td>
<td>58.23</td>
<td>71.08</td>
<td></td>
</tr>
<tr>
<td>CE+SelfSupCon</td>
<td>84.24</td>
<td>91.27</td>
<td>58.63</td>
<td>92.70</td>
<td>89.13</td>
<td>84.54</td>
<td>52.80</td>
<td>33.42</td>
<td>98.63</td>
<td>41.60</td>
<td>54.72</td>
<td>73.32</td>
<td>75.78</td>
<td></td>
</tr>
<tr>
<td>SupCon+SelfSupCon</td>
<td>82.07</td>
<td>91.54</td>
<td>59.24</td>
<td>92.51</td>
<td>89.04</td>
<td>84.18</td>
<td>54.08</td>
<td>33.92</td>
<td>98.71</td>
<td>40.84</td>
<td>55.25</td>
<td>82.84</td>
<td>71.39</td>
<td></td>
</tr>
<tr>
<td>*: test on the novel classes.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Few-shot classification accuracies (average score over 600 episodes) for 5-shot and 20-shot on the Mini-ImageNet and 12 downstream datasets. Contrastive approaches consistently perform better than cross-entropy across all downstream datasets.

5. Discussion and Analysis

Contrastive approaches learn more low/mid-level features. Figure 4 (top row) shows the similarity between different stages of the same ResNet-50 model in terms of the centered kernel alignment (CKA) [33]. The initial stages mostly learn low-level features, while the final stages learn more semantic information. The CE model has the least similarity between the representations of the final ResNet stage and the initial stage with a CKA score of 0.10. We infer that the final ResNet stages of CE contain mostly domain-specific high-level semantics. While SupCon is also trained in a supervised fashion to increase inter-class distance, it has a higher CKA score of 0.12 between the final and initial layers, which suggests that the final layers of SupCon contain more low-level and mid-level information than CE. As expected, the final stage of SelfSupCon contains the most low/mid-level information, and CE+SelfSupCon and SupCon+SelfSupCon show higher CKA scores between the final layers and initial layers than their supervised counterparts.

Similarly, Figure 4 (bottom row) shows the similarity between representations of the same ResNet stages between different models. We observe that the representations from ResNet stage-1 to stage-3 are highly similar between different methods, with CKA scores more than 0.9. However, the representations become slightly differentiated in ResNet stage-4 and highly dissimilar in the final ResNet stage. For example, the CKA between CE and SelfSupCon in the final layer is only 0.47. Surprisingly, CKA between CE and SupCon is 0.42 despite the fact that both models were trained with ImageNet1k with full supervision, which suggests that SupCon learns a much different feature representation than CE. Moreover, all contrastively trained models are more similar to each other than to cross-entropy model.

Supervised learning models learn feature representations using objectives that also increase the inter-class separation. However, we argue that increasing the intra-class variation, though possibly harmful for in-domain performance, is beneficial for learning rich feature representations in transfer learning. t-SNE visualizations in Figure 5 show that the clusters in contrastive methods are more spread out than in
vanilla cross-entropy, which also supports our claim.

**Contrastive models improve model calibration.** There are several metrics for measuring the models’ calibration. Here we adopt the Expected Calibration Error (ECE) and Negative Log Likelihood (NLL) [19]. See Appendix C.1 for the experimental setup.

Table 5 reports the performance for ImageNet pretrained models on the ImageNet validation dataset in terms of Top-1 accuracy (higher is better), Negative Log Likelihood (lower is better), and Expected Calibration Error (lower is better). We see that well-calibrated models do not necessarily have higher accuracy. In particular, SelfSupCon shows the best calibration performance, but worst top-1 accuracy score. Moreover, CE+SelfSupCon has better ECE and NLL scores than either CE or SelfSupCon, and SupCon+SelfSupCon shows better calibration scores than SupCon or SelfSupCon. We also evaluate the ImageNet pretrained models on the Stylized ImageNet validation set [16] to see whether contrastive approaches learn both texture-based and shape-based representations. In Table 5, we show that contrastively trained models perform better on the Stylized ImageNet validation set than the cross-entropy model. While SupCon+SelfSupCon has slightly lower accuracy than SupCon, it improves the calibration by 2.93%. Note that SelfSupCon performs the worst in terms of top-1 accuracy, which is expected since the backbone has not been trained with label information, and both of the datasets in Table 5 contain ImageNet classes. However, SelfSupCon performs better than CE in terms of ECE score. Overall, our experiments suggest that contrastive approaches produce more calibrated predictions than the cross-entropy model on both in-domain evaluation and transfer learning. Calibration performance on the 12 downstream datasets is provided in Appendix C.2.

<table>
<thead>
<tr>
<th>Method</th>
<th>ImageNet-R Top-1(%)</th>
<th>NLL</th>
<th>ImageNet-A Top-1(%)</th>
<th>ECE(%)</th>
<th>ImageNet-C mCE(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>35.83</td>
<td>19.45</td>
<td>3.35</td>
<td>55.03</td>
<td>60.80</td>
</tr>
<tr>
<td>SelfSupCon</td>
<td>42.01</td>
<td>13.40</td>
<td>7.36</td>
<td>46.78</td>
<td>54.23</td>
</tr>
<tr>
<td>SupCon</td>
<td>41.01</td>
<td>17.87</td>
<td>7.71</td>
<td>51.19</td>
<td>54.27</td>
</tr>
<tr>
<td>CE+SelfSupCon</td>
<td>43.67</td>
<td>15.60</td>
<td>6.56</td>
<td>34.66</td>
<td>54.40</td>
</tr>
</tbody>
</table>

Contrastive learning is robust to image corruption. Many deep learning models lack robustness to natural corruptions. In Table 6, we report the robustness performance of different models on the ImageNet-R [24], ImageNet-A [26], and ImageNet-C [25] datasets. Contrastive approaches, with the exception of SelfSupCon, show superior performance for both the ImageNet-A and ImageNet-R datasets, in both top-1 accuracy and expected calibration error (ECE) [19]. The improvement is particularly noticeable between CE and CE+SelfSupCon. CE+SelfSupCon improves the accuracy over CE by 5.18% for ImageNet-A and 4.38% for ImageNet-R, and lowers the calibration error by 1.58% for ImageNet-A and 3.72% for ImageNet-R.

The rightmost column of Table 6 reports performance of different models on ImageNet-C in terms of (unnormalized) mean corruption error (mCE) of the Noise, Blur, Weather, and Digital corruptions. Lower mCE denotes that the model is more robust to different corruption types. All the models are trained on clean ImageNet1K dataset. We observe that CE+SelfSupCon, SupCon, and SupCon+SelfSupCon perform the best across different models, and also provide better representations that are transferable to different domains, as shown in the linear evaluation and few-shot experiments. We also note that there is no single contrastive model that works best for all metrics in terms of robustness; however, contrastive loss, in general, improves the neural network robustness.

### 6. Ablation Studies

**Ablations on weights of SelfSupCon loss for model with joint objective.** As described in Section 3.1,
transfer performance of the datasets for models from different pretraining epochs. The fixed feature linear evaluation on the 12 downstream representations. Figure ImageNet-training epochs on the transferability of visual study the effect of pretrained checkpoints from different ImageNet.

Does transferability improve with longer training? Table 7: Performance comparison between CE and CE\text{(strong)} for linear evaluation, full-network fine-tuning, and 5-shot few-shot classification in terms of average top-1 accuracy over the 12 downstream datasets.

CE+SelfSupCon is trained with the objective of $L_{ce} + \alpha L_{SelfSupCon}$, where $\alpha$ is the weight on self-supervised contrastive loss. Figure 6 reports the effect of $\alpha$ on average transfer accuracy of 12 downstream datasets and ImageNet1K validation accuracy. We obtain the highest ImageNet accuracy for $\alpha = 1$; however, the highest transfer accuracy is reported at $\alpha = 2$, which indicates that higher ImageNet accuracy does not always imply higher transfer accuracy. Higher values of $\alpha$ impose more intra-class variation, and there is an optimal value where transferability of the model is maximized. Imposing more intra-class distance might hurt transfer performance, as shown in figure that when $\alpha > 2$ the transfer accuracy gradually decreases.

Effect of augmentations on CE model. For the CE model, we adopt standard data augmentation as used in ResNet-50 ImageNet training. We also train a cross-entropy model with additional augmentations, such as color-jitter, random gray-scale, and Gaussian blur, so that all the models are pre-trained with similar augmentations. We denote this model as CE\text{(strong)}. Table 7 shows mean accuracy of 12 downstream datasets for CE\text{(strong)} in linear evaluation, full-network fine-tune, and few-shot classification (averaged over 5 runs). CE\text{(strong)} is just slightly better than CE for transfer learning; however, contrastive approaches are still significantly better than both CE and CE\text{(strong)}, particularly for fixed-feature transfer.

Does transferability improve with longer training? We study the effect of pretrained checkpoints from different ImageNet-training epochs on the transferability of visual representations. Figure 7 shows average transfer accuracy for fixed feature linear evaluation on the 12 downstream datasets for models from different pretraining epochs. The transfer performance of the CE pretrained model improves very little after 80 pretraining epochs, suggesting that CE learns transferable representation mostly during the initial phase of the pretraining, and it learns more source-domain specific representation at later pretraining epochs. On the other hand, transferability of contrastive methods improve gradually with longer source dataset pretraining. Moreover, during the initial pretraining stages, we find that both CE+SelfSupCon and SupCon+SelfSupCon still perform better in transfer learning than other models, suggesting that the joint objectives could be beneficial even in resource constrained environments.

7. Conclusion In this paper, we conduct extensive analysis on the transferability of contrastive learning on the downstream image classification, few-shot recognition, and object detection tasks. Our study suggests that contrastive models consistently perform better in transfer learning than standard cross-entropy models, and combining self-supervised contrastive loss with cross-entropy or supervised contrastive loss improves transfer learning performance. We find several factors that make representations from contrastive learning more transferable than supervised cross-entropy model. The penultimate layer representations of contrastive learning are much different than the cross-entropy model; in particular, contrastive models contain more low-level and mid-level information in final layers, and the contrastively trained model shows larger intra-class separation, and contrastive models are more robust to image corruptions.

8. Acknowledgement This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-19-C-1001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

<table>
<thead>
<tr>
<th>Method</th>
<th>Linear-Evaluation</th>
<th>Finetune</th>
<th>Few-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>75.67</td>
<td>88.13</td>
<td>60.01</td>
</tr>
<tr>
<td>CE\text{(strong)}</td>
<td>75.91</td>
<td>88.27</td>
<td>61.31</td>
</tr>
</tbody>
</table>

Figure 6: Effect of different weights on the $L_{SelfSupCon}$ term in the CE+SelfSupCon model. Best viewed in color.

Figure 7: Left: Average linear evaluation accuracy (%) of all the downstream datasets, and Right: ImageNet validation accuracy, for intermediate checkpoints from different ImageNet-pretraining epochs. Transferable representations from the cross-entropy model do not improve much with more pretraining epochs on ImageNet. However, for contrastive approaches, longer pre-training improves transferability. Best viewed in color.
References


