What's Missing From Self-Supervised Representation Learning?

Dave Epstein*
Columbia University

Yiliang Shi*
Columbia University

Eugene Wu
Columbia University

Carl Vondrick
Columbia University

Abstract

Despite tremendous effort to train visual recognition systems without human supervision, there is still no substitute for large, labeled training datasets. We perform a large-scale analysis to quantitatively understand the difference between the representations learned by self-supervised learning and supervised learning. Adopting a large collection of trained models for different computer vision tasks, we probe for functional similarities between visual recognition systems. Experiments and visualizations suggest that two key differences between self-supervised and supervised models are its representations for 3D geometry and deformable objects, which also substantially contribute to its failures. Our hope is that such analysis will expose future research directions in self-supervised visual learning.

1. Introduction

Although there has been striking progress in visual recognition in the last decade, computer vision models still require large amounts of labeled training data [31]. This challenge has led a surge of research to develop approaches for self-supervised representation learning [7, 9, 41, 24, 11, 23] in order to learn visual features without a substantial amount of human supervision. Although lacking ground-truth labels, unlabeled visual data naturally has context, such as spatial arrangement [9, 22], temporal order [36, 20], or cross-modal synchronization [24, 3]. Leveraging these incidental relationships to create pre-training tasks has emerged as a popular paradigm to learn visual features.

However, despite tremendous effort, there is still no substitute for large, labeled datasets. Figure 1 illustrates examples where self-supervised representations still perform worse than supervised representations. However, collecting and annotating large datasets for each computer vision task is not scalable in practice, and limits the versatility of visual recognition. Since the performance of self-supervised visual models lags behind that of their supervised counterparts often by considerable margins, we ask the question: what is missing from self-supervised visual learning?

We present a large-scale empirical analysis to uncover

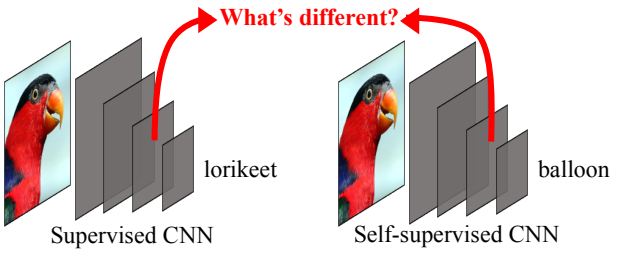

Figure 1: **Diagnosing Self-supervised Learning:** We show object class predictions using supervised features and self-supervised features. What is different about self-supervised representations?

what supervised models are learning that self-supervised models do not. Adopting a large collection of neural networks trained on computer vision tasks, we use recent methods [26, 28] to visualize and interpret neural networks in order to look "under the hood" of the self-supervised representation learning. However, in order to efficiently scale up and make our analysis automatic, we formulate our approach as a question of similarity between the representations learned by the models. By comparing the representations of supervised, self-supervised, and a bank of models, we can estimate what self-supervised features are learning, and more importantly, what they may miss.

Our experiments and visualizations provide a window into some of the functional shortcomings of current self-supervised representations. Our results show that, unsurprisingly, all self-supervised representations that we analyze lack strong representations for the object and scene semantics. However, since our approach to interpret the representation is comparison-based, we are also able to rank which tasks are the most similar and dissimilar to current self-supervised representations. For example, our analysis suggests that supervised models are better at learning 3D representations than self-supervised models, supporting that incorporating 3D into self-supervised learning remains an important research direction. Moreover, deformable objects remain challenging to represent for self-supervised models, and that self-supervised models share much more in common among themselves than they do with the supervised

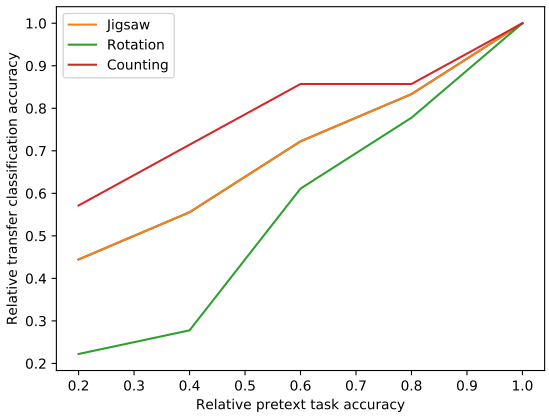

Figure 2: **Task Performance Correlation:** We plot performance of ImageNet accuracy versus performance of pretext task for some self-supervised methods. An increase in pretext task performance is highly correlated with an increase in transferred feature performance (Pearson correlation of 0.87). This suggests that self-supervised representation learning should be a reasonable approach. Why is there still a performance gap?

model trained for classification.

Figure 2 shows relative performance on pretext tasks is predictive of object recognition performance, suggesting that there are useful learning signals from self-supervised tasks. However, in practice, there are still large gaps in performance. Our goal is to understand from where this gap may come, and this paper is organized around investigating four questions that we will experimentally analyze. In section 2, we first review related work. Next, we tackle the question of analyzing characteristics of a supervised classification model, before delving into the next, a quantitative comparison of missing elements in self-supervised models. Following that, we attempt to evaluate the quality of self-supervised tasks with respect to its general capabilities and performance results. Finally, we analyze the attention maps of self-supervised and supervised models to analyze their behavior. Our hope is that this analysis will spur the next wave of research in self-supervised visual learning. We plan to release all code, data, and results.

2. Related Work

Analyzing Visual Recognition: This paper contributes to a large body of inquiry to analyze visual recognition systems. Due to the complexity of modern computer vision systems and their opaque failures, several works have focused on diagnosing the errors that visual recognition systems make [13, 25, 33]. Other work has studied the impact of training data, such as the size of the dataset [44, 31] or choice of the dataset [16, 32]. Several works have also studied the learned models directly [29, 19, 18, 2, 39]. In a related vein, several meta-analysis of problems in computer vision have been conducted [27, 6]. This paper is also analyzing visual recognition systems, however our focus is on self-supervised learning. Due to the practical impacts of learning without human supervision and the gaps in performance to supervised models, our hope is that this analysis will provide intuition for where self-supervised learning should go next.

Visualizing and Interpreting Models: The topic of interpreting neural networks have long been a subject of discussion. We build off an extensive line of research to develop tools to visualize and interpret neural networks, which provides the foundation for this study. In this paper, we would like to answer both the high level question of what tasks a model is performing, as well as finer granularity questions on specific characteristics of inputs that a model attends to. Common visualization techniques includes highlighting the most important portions of a image (saliency), parts of the original image that cause the most variation in network output (backpropagation), and parts of a image that contributes most to a classification (Grad-CAM) [28]. There are many variants of these techniques that adjust the regions highlighted. One disadvantage of visual analysis is that it depends on humans to observe patterns in visualization. Network dissection [4] quantitatively explores these patterns by computing IoU scores between neuron activations and labeled ground truth. Similar veins of work relate activations in recurrent networks to linguistic forms in the the context of natural language processing [14]. Lastly, recent techniques such as SVCCA [26] and PWCCA [21] enable the direct comparison of neural networks through their intermediate representations. In this paper, we use a combination of SVCCA and PCA to obtain high level interpretations of the entire model, as well as Grad-CAM and backpropagation techniques to analyze classification specifically.

Self-supervised Learning: The focus of this paper is to analyze self-supervised visual models and understand the reasons why their performance is lower than supervised models. Although self-supervised learning is twenty five years old, first proposed by de Sa in 1994 [7], there recently has been significant interest in this problem. A common approach has been to remove some already available information about image data and train a network to predict it. For example, images may be split up into grids and shuffled, or grid tiles' relative positions obscured [9, 22]; color [41, 17, 35] information may be removed and predicted; or orientation may be changed [11]. Other approaches use video to attempt to predict object motion [36, 34] and ego-motion [43, 1], learn similar representation for tracked

objects across frames [37], or leverage sound information [24, 3]. Networks have also been trained to count object entities [23], encode cross-channel information to reconstruct images [42], learn transitive invariance [38], and fuse multiple pretext tasks to improve performance [10]. Our work selects three self-supervised tasks that take diverse approaches to feature learning, are easy to understand, and provide state-of-the-art performance when transferred to ImageNet classification.

3. What tasks does an ImageNet classification model need to learn?

Before we can analyze what may be missing from self-supervised representations, we must first determine what supervised models learn.

How should we analyze neural networks to find out what they learn? We can functionally characterize what a model learns by comparing it with models trained on other visual tasks. If our classification network is similar to, *e.g.*, an edge detection network according to a reasonable metric, we can say that the classification network has learned edge detection. The Taskonomy task bank [40] contains a set of pre-trained visual task estimators. From these, we pick twenty to serve as the tasks we test our classification network against.

To facilitate comparison with self-supervised models later, we train an ImageNet classifier with the AlexNet architecture [15, 8]. All models from the Taskonomy test bank have identical-architecture encoders (4 ResNet [12] blocks) and shallow decoders with varied architecture. To facilitate comparisons across tasks, we thus focus on the task encoder networks. Although there could be issues with dataset bias because ImageNet and Taskonomy are likely different distributions of images [32], we empirically found the ImageNet models generalize to Taskonomy images without accuracy loss, suggesting these models are not impacted by domain shift.

We estimate the similarity between representation A and representation B of two neural networks by observing that the two neural networks are providing different views of the same input data. To measure similarity, rather than using naive Euclidean distance, which is sensitive to arbitrary transformations that could be applied to network weights without changing behavior, we build off [26, 21] and use an adaptation of Canonical Correlation Analysis (CCA). CCA is invariant to affine transformations common to convolution layers and has been shown to work in these settings.

We can score the similarity between these neural-net views of data by finding the low-dimensional subspace with maximal linear correlation between the two views. More formally, let $X \in \mathbb{R}^{D \times N}$ be a matrix of hidden activations from one network with D dimensional features over N examples. Likewise, let $Y \in \mathbb{R}^{D' \times N}$ be another matrix of hidden activations. We define the distance between these networks to be $\max_{w,u} \text{corr}\left(w^T X, u^T Y\right)$ where $w \in \mathbb{R}^{D \times K}$ is a projection matrix. This is the standard CCA objective and can be solved with off-the-shelf solvers.[1] It is empirically shown that convolution layers at different depths learn different features. As such, we match the layers of AlexNet and ResNet by depth [2], and define the CCA score of two networks to be the average CCA score across all pairs.

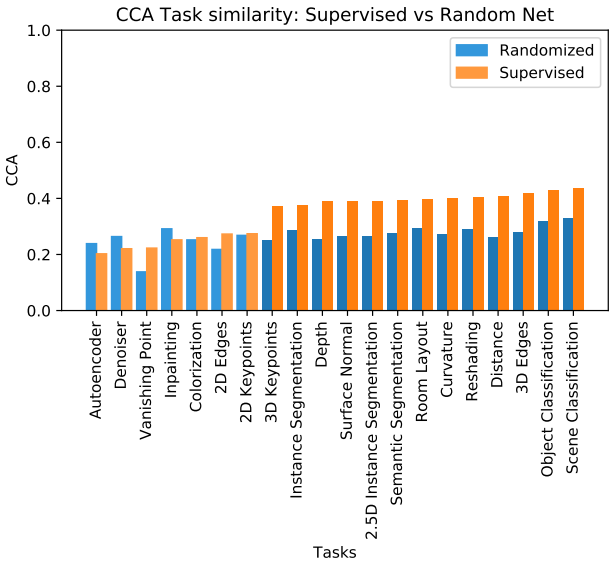

Figure 3: **What are supervised models learning?** We plot the similarity (vertical axis) between a bank of models (horizontal) with an ImageNet-supervised network (orange bars) and a network with random weights (blue bars). The tasks that are most useful for ImageNet classification will have a higher orange bar than blue bar. For visualization purposes, we lighten the bars where the difference is small or negative.

Which classical tasks best correlate with a supervised classification network? Figure 3 shows the similarities of a fully supervised classification network and a network with random weights with classical computer vision tasks. As expected, tasks with the highest similarity are object and scene classification, whereas tasks with the lowest similar-

[1] We use $N = 120$ since we found our results did not change much with larger N. With $N = 120$, the standard error of CCA scores computed from 10 different batches is 0.005 per comparison. We also use $K = 20$. Optimization typically took 4 hours on a multi-core server. As we are working with images, we use layer-wise comparisons across convolution layers. We resize the height and width of intermediate layer activations on each datapoint to 32x32 before flattening, giving $D = 1024 \cdot 120$. We treat each channel of the intermediate representation as a dimension of X and Y, giving us $num\ channels \times D$ matrices.

[2] That is, we pair `conv i` of self-supervised or supervised AlexNets with `block i` of the task networks, for `i=1,...,4`. We do not compare `conv5` with the final output layer of the encoder due to a large mismatch in dimensionality which led to unreliable results.

ity - even lower than that of a random network - are autoencoding and denoising, tasks which try to preserve the internal representation of a scene with no requirement of semantic understanding. Random networks show a similarity score of approximately 0.2 in our calculations, providing a baseline for all measurements.

Answer: We observe that the supervised model is significantly closer than a random net to 13 task networks that we analyze (*e.g.* 3D edges, surface normal estimation, and instance segmentation), suggesting that the supervised network learns features related to these tasks. Tasks in our task bank can be roughly grouped into semantic, lower-level, 2D, and 3D vision tasks. We observe that other than semantic tasks, the supervised network exhibits significant similarity to many 3D scene understanding tasks (reshading, distance estimation, curvature, 2.5D segmentation, depth estimation, etc.).

4. What tasks do self-supervised models not learn?

After establishing that the representations from fully supervised models are related to the tasks in our task bank, we would like to identify which of these tasks self-supervised models are not learning as well. For example, we expect that a network trained to piece a jigsaw together should be able to detect edges well, a capability which could transfer to object classification.

To form a collection of self-supervised models, we implement and train self-supervised networks with the AlexNet architecture using the ImageNet dataset. We fix as many hyper-parameters as possible to enable an "apples-to-apples" comparison. We consider three self-supervised models: rotation [11], counting [23], and jigsaw puzzle solving task [22]. These tasks represent three different approaches to pretext tasks (learning of orientation, object presence, and object part relationships respectively), and make up the current state-of-the-art in feature transfer to image classification. We train the rotation network to 87% accuarcy on its task, the jigsaw network to 92%, and the counting network to 0.03 loss.

How do these similarities differ for the self-supervised networks? Following the same procedure in the previous section for supervised networks, we can use CCA to estimate the similarity between the representations of self-supervised models and other models. Since we are interested in what is missing, we can plot the differences of similarities. Figure 4 visualizes the difference between CCA scores for the self-supervised and supervised networks conditioned on a task bank. In this plot, the higher (and more red) the difference, the more the self-supervised representation is missing that particular task. We order the tasks on the y-axis according to their transfer classification accuracy with the rotation network, which is the state-of-the-art

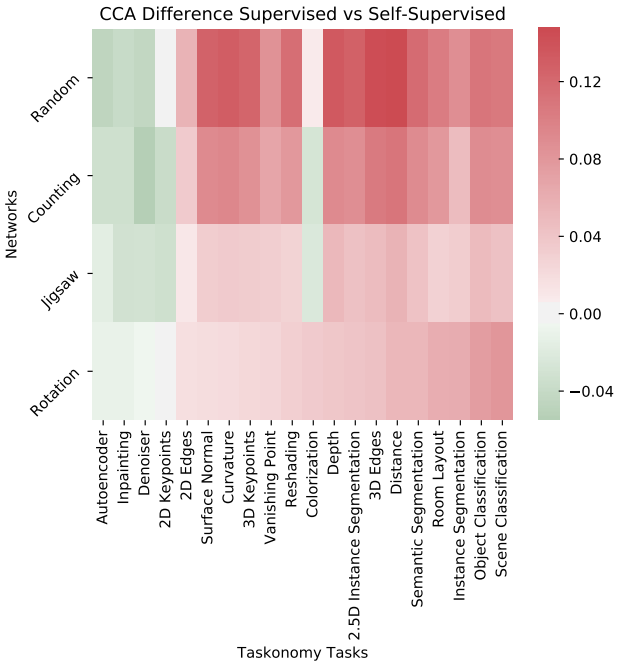

Figure 4: **What's missing from self-supervised learning?** We plot the difference between supervised and self-supervised CCA similarities for visual tasks. Red cells indicate that a self-supervised model correlates less with the task than a supervised model. Likewise, green cells indicate that a self-supervised model is closer to the task than a supervised model. This plot suggests that self-supervised representations are lacking information about object semantics and 3D geometry.

in self-supervised representation learning. Tasks along the x-axis are ordered according to how much they are missing from the rotation network, the strongest of the three self-supervised nets we examine.

These results suggest that semantic tasks such as object and scene classification are missing most from self-supervised networks. After those tasks, tasks requiring an understanding of three-dimensional geometry are most lacking: distance, 3D edge, and depth estimation are notably absent from self-supervised net when compared to the supervised net. Figure 3 suggests that autoencoders, in-painting networks and denoising networks do not contribute much to classification accuracy. The fact that self-supervised models show some correlation with these tasks while supervised models do not suggests that these most self-supervised learning approaches are actually closer in representation space to autoencoders than supervised models. We also observe from the plot that tasks which the self-supervised networks are relatively strongest on early vision tasks such as 2D keypoint estimation and edge estimation. Counting as a task certainly miss the most, which explains

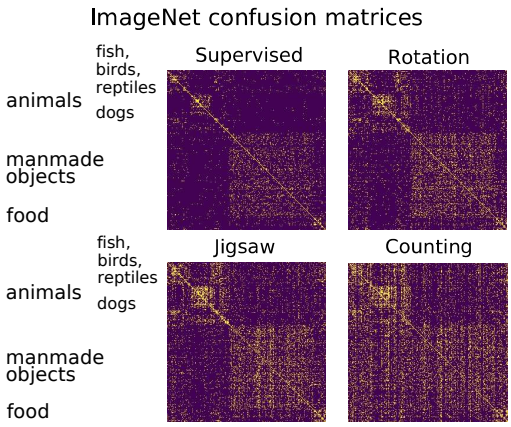

Figure 5: **Comparing Confusions:** We show confusion matrices for all networks. Stronger representations have a more block-diagonal confusion matrix, displaying better hierarchical understanding. Matrices are generated by charting the top 5 predictions for each ImageNet image, to show prediction variation. When the correct label is in the top 5 predictions, the prediction is treated as correct and only the diagonal cell is incremented. Otherwise, all cells corresponding to top 5 predictions are. To facilitate visualization, the matrices are binarized such that cells for any class predicted at least once are activated. 74% of supervised predictions are within block-diagonal sections, compared to 64% for rotation, 58% for jigsaw, and 41% for counting.

it low transfer accuracy.

5. How transferable are self-supervised features?

Self-supervised representation learning attempts to find an input embedding space useful for some downstream task. The common scenario in computer vision tries to find features good for transfer onto image classification using a standard dataset such as ImageNet. There has been intense interest in finding tasks that, while not requiring manual annotation, learn features that yield high classification accuracy when transferred or fine-tuned. Inspired by previous work [39], we study the practical performance of self-supervised representations by using them as input to linear regression models trained on the desired task, in this case, image classification.

To ensure a fair comparison between all self-supervised tasks, we freeze all learned AlexNet weights up to `conv3`, the layer with highest transfer performance reported in every self-supervised task we use, and train one fully connected linear layer on ImageNet classification using the embeddings given by the `conv3` features. Features with better semantic information will allow the regression to better discriminate between classes.

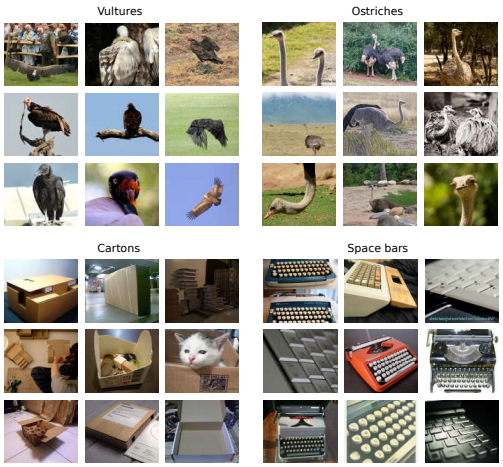

Figure 6: **What makes some classes harder for self-supervision?** Random samples of instances from confusing (top) and less confusing (bottom) classes for self-supervised networks. We observe a greater degree of intra-class variation and deformability in failure modes than in success modes, indicating that self-supervised representations struggle to understand more complex classes.

We examine the results of these linear regression models as well as consider class confusion from a hierarchical perspective, examining whether self-supervised features learn coarse object category discrimination. We also find types of objects that are particularly confusing across all self-supervised representations. We analyze performance of learned features as pretext training progresses, investigating whether getting better at the self-supervised task necessarily means learning better features.

What are common confusions from self-supervision? Previous work [5, 8] shows that many mistakes that supervised neural networks make in classification are confusions within some meaningful semantic class. For example, there are 120 dog breeds in ImageNet, and a discriminative model that confuses them might be forgiven since it still learned some higher-level desirable object understanding.

To analyze this for self-supervised representations, we calculate confusion matrices across all 1,000 ImageNet classes for each network. The classes are ordered by position in the WordNet hierarchy tree, such that tasks under the same node in the tree are nearby in the confusion matrix. We visualize these confusion matrices in Figure 5. These matrices show that the better the representation, the better the hierarchical understanding and the more block-diagonal the confusion matrix. A better self-supervised pretext task may be one that explicitly targets the learning of such fine-grained semantic information.

We also examine the most and least confusing classes for each network in Table 1. More than two-thirds of classes

Network	Class	Performance Gap	Network	Class	Performance Gap	Network	Class	Performance Gap
Rotation	Vulture* *(Bald eagle)*	0.64	**Jigsaw**	Ostrich* *(Meerkat)*	0.82	**Counting**	Cliff dwelling† *(Newt)*	0.88
	Hen* *(Gamecock)*	0.58		Spoonbill* *(Egret)*	0.78		Proboscis monkey* *(Jaguar)*	0.86
	Fig *(Granny Smith apple)*	0.58		Albatross* *(Gray whale)*	0.76		Ostrich* *(Tiger beetle)*	0.84
	English spaniel* *(St. Bernard)*	0.58		Pomeranian* *(Persian cat)*	0.74		Spoonbill* *(Flamingo)*	0.8
	Marimba† *(Carousel)*	0.58		Mountain tent† *(Alp)*	0.74		Three-toed sloth* *(Koala)*	0.8
	Coral reef	-0.02		Promontory	0		Screwdriver†	0.04
	Mortarboard†	-0.04		Overskirt†	-0.02		Spatula†	0.02
	Dishrag†	-0.04		Appenzeller*	-0.02		Hook†	0.02
	Space bar†	-0.06		Coral reef	-0.02		Velvet†	0
	Carton†	-0.06		Space bar†	-0.04		Chiffonier†	0

Table 1: Strongest and weakest classes for each self-supervised representation compared to the fully supervised network. Most common confusions for the weakest classes are listed in *parentheses*. The difference column is equal to supervised accuracy minus self-supervised accuracy, so higher figures indicate weaker classes for the self-supervised networks and vice versa. Classes marked with an * are animals and those with a † are artifacts. $\frac{11}{15}$ of the weakest classes for self-supervised nets are animals, whereas $\frac{11}{15}$ of the strongest classes are artifacts.

where self-supervised nets perform worst belong to the animal subtree of the ImageNet hierarchy. Similarly, more than two-thirds of the best classes for self-supervised nets are artifacts (i.e. manmade objects) which are shallower in the ImageNet hierarchy by approximately 1.1 levels on average. These classes tend to exhibit a lower degree of deformability and variance. This suggests self-supervised networks struggle with highly specific classes that vary significantly in their appearance (in ImageNet, such classes are mostly animals), as visualized in Figure 6. We also note that fine-grained class discrimination may be a harsh metric for evaluating the quality of learned representations.

Should we go bigger with our data? We investigate whether self-supervised learning can take advantage of large unlabeled datasets by comparing the networks' performance on pretext tasks to performance of linear regression layers using representations at various stages of pretext training. In particular, we train using features at epochs corresponding to 20, 40, 60, 80, and 100% of final accuracy, and observe how relative performance on classification changes. For an ideal pretext task which can take advantage of large sources of unlabeled data, improving accuracy would strongly correlate with improving quality of representation, since otherwise the task would saturate the utility of the dataset early. We plot relative performance on the pretext task and on transferred classification for each of the three self-supervised networks in Figure 2. These findings suggest that we would benefit from using datasets orders of magnitude larger than ImageNet, such as [31], to provide models with further semantic understanding.

Answer: Figure 2 suggests that self-supervised representations tend to transfer well, as transferred classification performance improves with pretext training progress, indicating that there may be still be signal left in existing self-supervised techniques. We hypothesize that scaling up datasets even by several orders of magnitude may help further push this boundary. An analysis of classification results shows that poorer performance of self-supervised networks is likely due to missing semantic, hierarchical understanding. Object groups that are most confusing to self-supervised networks are those deeper in the ImageNet class tree, and the misclassifications of stronger representations are more constrained to coarse class groups (e.g. structure, vehicle, bird, fish). This may suggest that, if ImageNet is the goal, a promising direction for future self-supervised tasks could involve a better awareness of object hierarchy. Further, many of these confusing classes exhibit a high degree of deformability and variance, suggesting that self-supervised tasks trained with this in mind may be useful.

Network	Metric	Top three	Bottom three
Rotation	IoU	ocean liner; container ship; lifeboat	jellyfish; parachute; theater curtain
	Spearman	aircraft carrier; mosque; lifeboat	T-shirt; theater curtain; armadillo
	Sharpness	killer whale; geyser; airship	rickshaw; slot machine; ambulance
Jigsaw	IoU	container ship; ocean liner; salamander	Leonberg; koala; jellyfish
	Spearman	school bus; lifeboat; aircraft carrier	Leonberg; vault; T-shirt
	Sharpness	geyser; killer whale; mosquito net	hockey puck; odometer; ambulance
Counting	IoU	ocean liner; container ship; lifeboat	jellyfish; parachute; T-shirt
	Spearman	school bus; rapeseed; ear of corn	jellyfish; stingray; platypus
	Sharpness	nipple; website; geyser	slot machine; football helmet; ambulance
Supervised	Sharpness	indigo bird; house finch; bee eater	cauliflower; mashed potato; carbonara

Table 2: Top and bottom three ImageNet classes for various attention metrics. The supervised net is sharpest on animal classes with high specificity. Self-supervised networks attend most similarly to the supervised net on artifacts and least similarly on animals.

Network	IoU	Spearman's ρ	Sharpness
Rotation	0.139 ± 0.100	-0.080 ± 0.091	0.557 ± 0.157
Jigsaw	0.125 ± 0.102	-0.051 ± 0.087	0.680 ± 0.138
Counting	0.150 ± 0.102	-0.021 ± 0.077	0.535 ± 0.187
Supervised	1.000 ± 0.000	1.000 ± 0.000	0.823 ± 0.093

Table 3: This table compares the attention maps of supervised and self-supervised representations. For all metrics, higher is better. In particular, Spearman rank correlation with the supervised network is statistically insignificant.

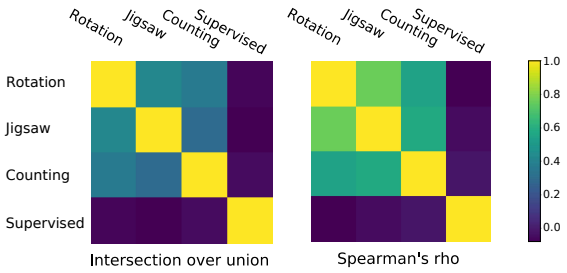

Figure 7: **Calculating Attention Similarity:** Intersection over union and Spearman's rank correlation calculated over all network pairs. Self-supervised networks all attend similarly to each other, but differently from the supervised net.

6. Do self-supervised and supervised models attend to similar spatial regions?

If two visual models attend to different regions in the same image to make their prediction, they have learned different internal representations. By analyzing these the similarities and differences of the learned attention maps, we can gain insight into what self-supervised models are missing in comparison to the supervised counterparts.

What do they attend to? We investigate this question using guided backpropagation [30] and gradient-weighted class activation mapping [28], which provide local and global information on which image regions are important to the model's decision-making. Fig. 8 visualizes the activations of all networks on a random sample of images from ImageNet. Despite extensive data preprocessing across self-supervised tasks, some biases are notable in the attention maps. The jigsaw network does not learn to understand color, since it is heavily regularized against using chromatic aberration to solve puzzles. The counting network learns features which appear to be texture-invariant, but detect repetitive image areas (e.g. row 3 in refrigerator or row 4 in pretzel in Figure 8), for similar reasons. The rotation network, which performs best out of the three, has features most qualitatively similar to the supervised network, but still shows some oversensitivity to edges and other telltale signs of orientation (e.g. row 3 in soup bowl). Attempting to prevent shortcuts in or overfitting to the pretext training stage may prevent neural networks from learning useful image features for downstream tasks such as classification.

Good and bad attention: In Figure 9, we show binarized attention maps for the supervised and self-supervised networks on their strongest and weakest classes (which tend to be animals and artifacts, respectively, as shown in Table 1). What are self-supervised networks paying attention to when they succeed and when they fail? We see that supervised networks are able to focus on very small parts of the image that allow for discrimination between similar classes, whereas self-supervised networks tend to focus on whole objects or object parts. Success modes for the self-supervised networks occur in classes shallower in the WordNet hierarchy , demonstrating that a lack of semantic knowledge hinders self-supervised attention. However, regions highlighted by self-supervised features may be more informative to humans (see e.g. the third row of Figure 9 where, despite misclassification, the entire carton and chiffonier are highlighted).

Quantifying similarity: We select two metrics to numerically analyze attention between neural networks: intersection over union on binarized normalized attention maps (all locations with $>50\%$ of max attention get a value of 1) and Spearman's rank correlation on the 13×13 magnitudes assigned to locations in the output of each net's `conv5` layer. Table 3 shows IoU and Spearman's rank correlation coefficient for all networks. All three self-supervised nets are well within a standard deviation of each other in IoU and rank correlation with the supervised network. That is, none of the self-supervised tasks attends significantly better than the others, and none attends remotely similarly to the supervised network. The rank correlation shows that the self-supervised networks are (if barely) slightly negatively correlated with the supervised networks in how they attend to data. Figure 7 shows the IoU and Spearman for all network pairs; self-supervised networks attend much more similarly to each other than to the supervised network.

Sharpness of attention focus: To investigate the hypothesis that a better network with more semantic information has a more concentrated attention distribution, we would expect a supervised network to be significantly sharper in focus than a self-supervised one. We consider a proxy metric for the non-uniformity of the attention distribution: the sharper a peak this distribution has, the surer a net is in its attention. We define sharpness as the percentage of the image with a relative attention below one half.

Table 3 suggests that the supervised attention is sharper and more consistent than self-supervised attention (the standard error of the measure is around half that of the self-supervised nets). Table 2 shows the classes with highest and lowest values for each of the quantitative metrics used. The supervised network's attention is sharpest on very taxonomically specific animals and most diffuse on generic food items with less distinctive properties.

Answer: Our quantitative and qualitative results demonstrate that self-supervised networks exhibit no significant correlation with the attention of fully-supervised networks, whether in the coarse (binned rank-correlation) or fine (IoU) resolution. Self-supervised features are also unable to confidently attend to salient, semantic image regions. We observe that in many classification cases, the supervised features only focus on uninformative (to humans) and hyper-discriminative image regions while the self-supervised nets

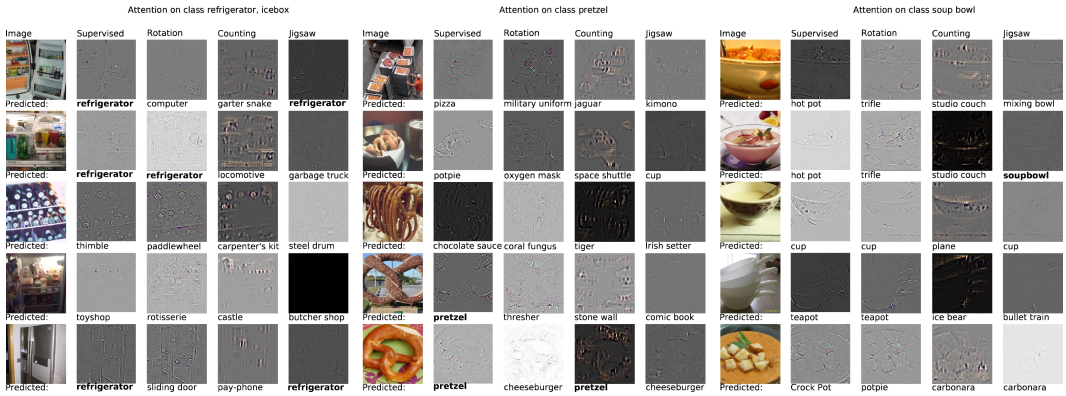

Figure 8: **Comparing Attention Maps:** Guided Grad-CAM maps for supervised and self-supervised networks on randomly selected ImageNet images (with increased contrast for easier visualization). Despite the supervised regression layer, each self-supervised network's attention patterns are biased towards its pretext task. Figure best viewed zoomed in and in color.

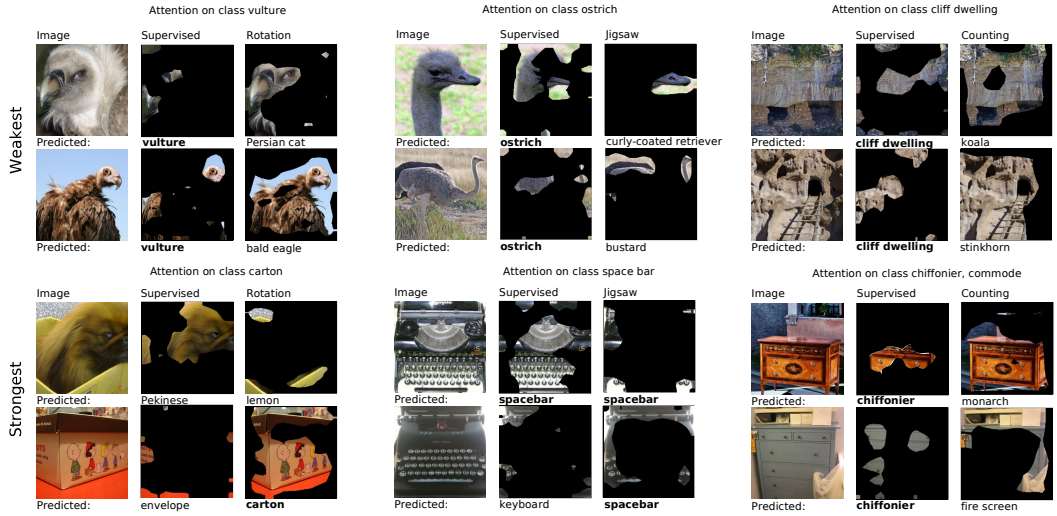

Figure 9: **Attention Regions in Success and Failure:** Binarized Grad-CAM maps for supervised and self-supervised networks on random ImageNet images from top and bottom classes for each network (see Table 1). The first two rows are images from the weakest class for each network, the bottom two from the strongest. Self-supervised representations struggle to focus on regions that allow distinguishing animals and other objects appearing in natural scenes, but have less trouble classifying manmade artifacts. Best viewed zoomed in and in color.

may provide a more natural, interpretable attention mechanism, at the expense of performance. By every metric, self-supervised networks are far more similar to each other in their attention than to a supervised classification network.

7. Discussion

Evaluation of self-supervised visual representations have typically focused on measuring the performance on downstream recognition tasks, such as image classification or object detection. However, this evaluation is fairly limited because it only quantitatively analyzes the end-to-end task performance. Instead, our work suggests that comparison-based evaluations on the hidden representations can provide a refined analysis of self-supervised learning. Our hope is that this analysis will enable us to not only quantify emergent behaviors, but also identify weaknesses and expose possible research directions for the next generation of self-supervised visual learning.

References

[1] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 37–45, 2015. 2

[2] P. Agrawal, R. Girshick, and J. Malik. Analyzing the performance of multilayer neural networks for object recognition. In *European conference on computer vision*, pages 329–344. Springer, 2014. 2

[3] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 1, 3

[4] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, 2017. 2

[5] A. Bilal, A. Jourabloo, M. Ye, X. Liu, and L. Ren. Do convolutional neural networks learn class hierarchy? *IEEE transactions on visualization and computer graphics*, 24(1):152–162, 2018. 5

[6] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[7] V. R. de Sa. Learning classification with unlabeled data. In *Advances in neural information processing systems*, pages 112–119, 1994. 1, 2

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 5

[9] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015. 1, 2

[10] C. Doersch and A. Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2051–2060, 2017. 3

[11] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 1, 2, 4

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[13] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *European conference on computer vision*, pages 340–353. Springer, 2012. 2

[14] A. Kádár, G. Chrupała, and A. Alishahi. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780, 2017. 2

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3

[16] A. Lapedriza, H. Pirsiavash, Z. Bylinskii, and A. Torralba. Are all training examples equally valuable? *arXiv preprint arXiv:1311.6510*, 2013. 2

[17] G. Larsson, M. Maire, and G. Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6874–6883, 2017. 2

[18] K. Lenc and A. Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 991–999, 2015. 2

[19] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015. 2

[20] I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 1

[21] A. Morcos, M. Raghu, and S. Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, pages 5732–5741, 2018. 2, 3

[22] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 1, 2, 4

[23] M. Noroozi, H. Pirsiavash, and P. Favaro. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5898–5906, 2017. 1, 3, 4

[24] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pages 801–816. Springer, 2016. 1, 3

[25] D. Parikh and C. L. Zitnick. Finding the weakest link in person detectors. 2011. 2

[26] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pages 6076–6085, 2017. 1, 2, 3

[27] O. Russakovsky, J. Deng, Z. Huang, A. C. Berg, and L. Fei-Fei. Detecting avocados to zucchinis: what have we done, and where are we going? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2064–2071, 2013. 2

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 1, 2, 7

[29] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[30] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 7

[31] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 1, 2, 6

[32] A. Torralba and A. A. Efros. Unbiased look at dataset bias. 2011. 2, 3

[33] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: Visualizing object detection features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–8, 2013. 2

[34] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016. 2

[35] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 391–408, 2018. 2

[36] J. Walker, A. Gupta, and M. Hebert. Dense optical flow prediction from a static image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2443–2451, 2015. 1, 2

[37] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015. 3

[38] X. Wang, K. He, and A. Gupta. Transitive invariance for self-supervised visual representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1329–1338, 2017. 3

[39] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014. 2, 5

[40] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018. 3

[41] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 1, 2

[42] R. Zhang, P. Isola, and A. A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017. 3

[43] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017. 2

[44] X. Zhu, C. Vondrick, D. Ramanan, and C. C. Fowlkes. Do we need more training data or better models for object detection?. In *BMVC*, volume 3, page 5. Citeseer, 2012. 2