Concept Generalization in Visual Representation Learning

Mert Bulent Sariyildiz[1,2,] Yannis Kalantidis[1] Diane Larlus[1] Karteek Alahari[2]
[1] NAVER LABS Europe [2] Inria*

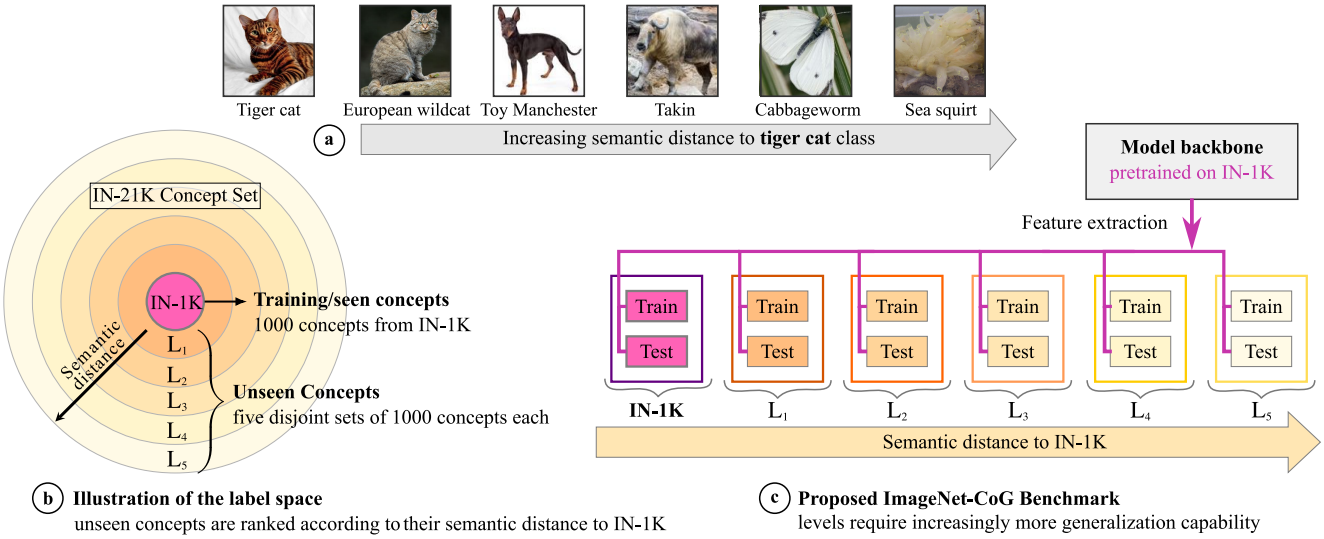


Figure 1: **An overview of our Concept Generalization (CoG) benchmark.** (a) An example of five concepts from the ImageNet-21K dataset [11] (IN-21K), ranked by increasing *semantic* distance (decreasing Lin similarity [34]) to the ImageNet-1K (IN-1K) dataset [47] concept "Tiger cat". (b) We rank the 21K concepts of IN-21K according to their semantic distance to the 1000 concepts of IN-1K and split the ranked list to extract 5 groups of 1000 concepts. We refer to the five IN-1K-sized datasets of increasing semantic distance from IN-1K as *concept generalization levels*, denoted as $L_{1/2/3/4/5}$. (c) The proposed ImageNet-CoG benchmark uses a model trained on IN-1K as a feature extractor and evaluates its concept generalization capabilities by learning linear classifiers for each level of more and more challenging unseen concepts.

Abstract

Measuring concept generalization, i.e., the extent to which models trained on a set of (seen) visual concepts can be leveraged to recognize a new set of (unseen) concepts, is a popular way of evaluating visual representations, especially in a self-supervised learning framework. Nonetheless, the choice of unseen concepts for such an evaluation is usually made arbitrarily, and independently from the seen concepts used to train representations, thus ignoring any semantic relationships between the two. In this paper, we argue that the semantic relationships between seen and unseen concepts affect generalization performance and propose ***ImageNet-CoG****,*[1] *a novel benchmark on the ImageNet-21K (IN-21K) dataset that enables measuring concept generalization in a principled way. Our benchmark leverages expert knowledge that comes from WordNet in order to define a sequence of unseen IN-21K concept sets that are semantically more and more distant from the ImageNet-1K (IN-1K) subset, a ubiquitous training set. This allows us to benchmark visual representations learned on IN-1K out-of-the box. We conduct a large-scale study encompassing 31 convolution and transformer-based models and show how different architectures, levels of supervision, regularization techniques and use of web data impact the concept generalization performance.*

*UGA, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

[1] https://europe.naverlabs.com/cog-benchmark

1. Introduction

There has been an increasing effort to tackle the need for manually-annotated large-scale data in deep models via transfer learning, i.e., by transferring representations learned on resourceful datasets and tasks to problems where annotations are scarce. Prior work has achieved this in various ways, such as, imitating knowledge transfer in low-data regimes [60], exploiting unlabeled data in a self- [22] or weakly- [37] supervised manner.

The quality of the learned visual representations for transfer learning is usually determined by checking whether they are useful for, i.e., generalize to, a wide range of downstream vision tasks. Thus, it is imperative to quantify this generalization, which has several facets, such as generalization to different input distributions (e.g., from synthetic images to natural ones), to new tasks (e.g., from image classification to object detection), or to different semantic concepts (e.g., across different object categories or scene labels). Although the first two facets have received much attention recently [18, 20], we observe that a more principled analysis is needed for the last one.

As also noted by [12, 67], the effectiveness of knowledge transfer between two tasks is closely related to the semantic similarity between the concepts considered in each task. However, assessing this relatedness is not straightforward, as the semantic extent of a concept may depend on the task itself. In practice, models consider an exhaustive list of downstream tasks that cover a wide range of concepts [7, 29] in order to test their transfer learning capabilities. Previous attempts discussing this issue have been limited to intuition [67, 75]. We still know little about the impact of the *semantic relationship* between the concepts seen during training visual representations and those seen during their evaluation (*seen* and *unseen* concepts, respectively).

In this paper, we study the generalization capabilities of visual representations across concepts that exist in a large, popular, and broad ontology, the subset of WordNet [41] used to build ImageNet-21K [11] (IN-21K), while keeping all the other generalization facets fixed. Starting from a set of seen concepts, the concepts from the popular ImageNet-1K [47] (IN-1K) dataset, we leverage semantic similarity metrics based on this ontology crafted by experts to measure the semantic distance between IN-1K and every unseen concept (i.e., any concept from IN-21K that is not in IN-1K). We rank unseen concepts with respect to their distance to IN-1K and define a sequence of five, IN-1K-sized *concept generalization levels*, each consisting of a distinct set of unseen concepts with increasing semantic distance to the seen ones. This results in a large-scale benchmark that consists of five thousand concepts, that we refer to as the **ImageNet Co**ncept **G**eneralization benchmark, or **ImageNet-CoG** in short. The benchmark construction process is illustrated in Fig. 1.

Given a model trained on IN-1K, the evaluation protocol for ImageNet-CoG consists of two phases: it first extracts features for images of IN-1K and of the five concept generalization levels, and then learns individual classifiers, for each level, using a varying amount of samples per concept. By defining the set of seen concepts for our benchmark to be IN-1K classes, we are able to evaluate models trained on IN-1K out-of-the box. We therefore use publicly available pretrained models and analyse a large number of popular models under the prism of concept generalization. Our contributions are as follows.

- We propose a systematic way to study concept generalization, by defining a set of seen concepts along with sets of unseen concepts that are semantically more and more distant from the seen ones.
- We design ImageNet-CoG, a large-scale benchmark, which embodies this systematic way. It is designed to evaluate models pretrained on IN-1K out-of-the-box and draws unseen concepts from the rest of the IN-21K dataset. We measure concept generalization performance on five, IN-1K-sized levels, by learning classifiers with a few or all the training images from the unseen concepts.
- We conduct a large-scale study benchmarking 31 state-of-the-art visual representation learning approaches on ImageNet-CoG and analyse how different architectures, levels of supervision, regularization techniques and additional web data impact the concept generalization performance, uncovering several interesting insights.

2. Related Work

Generalization has been studied under different perspectives such as regularization [52] and augmentation [69] techniques, links to human cognition [16], or developing quantitative metrics to better understand it, e.g., through loss functions [31] or complexity measures [42]. Several dimensions of generalization have also been explored in the context of computer vision, for instance, generalization to different visual distributions of the same concepts (domain adaptation) [10], or generalization across tasks [71]. Generalization across concepts is a crucial part of zero-shot [51] and few-shot [60] learning. We study this particular dimension, concept generalization, whose goal is to transfer knowledge acquired on a set of *seen* concepts, to newly encountered *unseen* concepts as effectively as possible. Different from existing work, we take a systematic approach by considering the semantic similarity between seen and unseen concepts when measuring concept generalization.

Towards a structure of the concept space. One of the first requirements for rigorously evaluating concept generalization is structuring the concept space, in order to analyze the impact of concepts present during pretraining and transfer stages. However, previous work rarely discusses the par-

ticular choices of splits (seen vs. unseen) of their data, and random sampling of concepts remains the most common approach [21, 24, 30, 63]. A handful of methods leverage relations designed by experts. The WordNet graph [41] for instance helps build dataset splits in [15, 67] and a domain-specific ontology is used to test cross-domain generalization [20, 61]. These splits are however based on heuristics, instead of principled mechanisms built on semantic relationship between concepts as we do in this paper.

Transfer learning evaluations. When it comes to evaluating the quality of visual representations, the gold standard is to benchmark models by solving diverse tasks such as classification, detection, segmentation and retrieval on many datasets [4, 7, 13, 18, 22, 29, 73]. The most commonly used datasets are IN-1K [47], Places [76], SUN [64], Pascal-VOC [14], MS-COCO [35]. Such choices, however, are often made independently from the dataset used to train the visual representations, ignoring their semantic relationship.

In summary, semantic relations between pretraining and transfer tasks have been overlooked in evaluating the quality of visual representations. To address this issue, we present a controlled evaluation protocol that factors in such relations.

3. Our ImageNet CoG Benchmark

Transfer learning performance is highly sensitive to the semantic similarity between concepts in the pretraining and the target datasets [12, 67]. Studying this relationship requires carefully constructed evaluation protocols: i) controlling which concepts a model has been exposed to during training (seen concepts), and ii) the semantic distance between these seen concepts and those considered for the transfer task (unseen concepts). As discussed earlier, current evaluation protocols severely fall short on handling these aspects. To fill this gap, we propose *ImageNet Concept Generalization (CoG)*—a benchmark composed of multiple image sets, one for pretraining and several others for transfer, curated in a controlled manner in order to measure the transfer learning performance of visual representations to sets of unseen concepts with increasingly distant semantics from the ones seen during training.

While designing this benchmark, we considered several important points. First, in order to exclusively focus on concept generalization, we need a controlled setup tailored for this specific aspect of generalization. In other words, we need to make sure that the only change between the pretraining and the transfer datasets is the set of concepts. In particular, we need the input image distribution (natural images) and the annotation process (which may determine the statistics of images [57]) to remain constant.

Second, to determine the semantic similarity between two concepts, we need an auxiliary knowledge base that can provide a notion of semantic relatedness between visual concepts. It can be manually defined with expert knowledge, e.g., WordNet [41], or automatically constructed, for instance by a language model, e.g., word2vec [40].

Third, the choice of the pretraining and target datasets is crucial. We need these datasets to have diverse object-level images [2] and to be as less biased as possible, e.g., towards canonical views [39].

Conveniently, the **IN-21K dataset** fulfills all these requirements. We therefore choose it as the source of images and concepts for our benchmark. IN-21K contains 14,197,122 curated images covering 21,841 concepts, all of which are further mapped into synsets from the WordNet ontology, which we use to measure semantic similarity.

In the rest of this section, we first define the disjoint sets of seen and unseen concepts, then present our methodology to build different levels for evaluating concept generalization, and describe the evaluation protocol.

3.1. Seen concepts

We make a natural choice and use the 1000 classes from the ubiquitous IN-1K dataset [47] as the set of our *seen* concepts. IN-1K is a subset of the IN-21K [11]. It consists of 1.28M images and has been used as the standard benchmark for evaluating novel computer vision architectures [23, 50, 53, 58], regularization techniques [49, 59, 69, 74] as well as self- and semi-supervised models [5, 8, 19, 22, 65].

Choosing IN-1K as the seen classes further offers several advantages. Future contributions, following standard practice, could train their models on IN-1K, and then simply evaluate generalization on our benchmark with their pretrained models. It also enables us to benchmark visual representations learned on IN-1K out-of-the box, using publicly available models (as shown in Sec. 4).

3.2. Selecting eligible unseen concepts

We start from the Fall 2011 version of the IN-21K. [11] dataset[2] Since we are interested in concepts that are not seen during training, we explicitly remove the 1000 concepts of IN-1K. We also remove all the concepts that are ancestors of these 1000 in the WordNet [41] hierarchy. For instance, the concept "cat" is discarded since its child concept "tiger cat" is in IN-1K. It was recently shown that a subset of IN-21K categories might exhibit undesirable behavior in downstream computer vision applications [66]. We therefore discard all the concepts under the 'person' sub-tree. In addition, we chose to discard a small set of potentially offensive concepts (see supplementary material for details). We follow IN-1K [47] and keep only concepts that have at least 782

[2]Note that the recently released Winter 2021 ImageNet version shares the same set of images for all the unseen concepts selected in our benchmark with the Fall 2011 one. We refer the reader to the supplementary for further discussion on both the recent Winter 2021 release as well as a newer, blurred version of IN-1K.

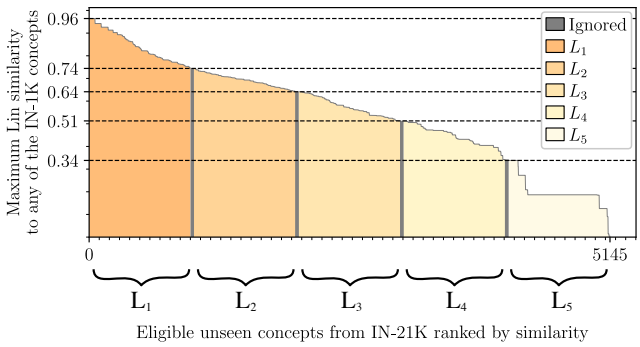


Figure 2: **Concept generalization levels.** We rank all the 5146 eligible IN-21K unseen concepts with respect to their similarity to IN-1K using Eq. (2) and split the ranked list into 5 groups of 1000 concepts each. Each group defines a concept generalization level, each denoted by $L_{1/2/3/4/5}$. Gray-shaded areas correspond to concepts that are ignored.

images, ensuring a relatively balanced benchmark. Finally, we discard concepts that are not leaf nodes in the WordNet subgraph defined by all so-far-eligible concepts. Formally, for any c_1 and c_2 in the set of unseen concepts, we discard c_1 if c_1 is a parent of c_2. These requirements reduce the set of eligible *unseen* IN-21K concepts to 5146 categories.

3.3. Concept generalization levels

Our next step is defining a sequence of unseen concept sets, each with decreasing semantic similarity to the seen concepts in IN-1K. We refer to each one of these as a *concept generalization level*. They allow us to measure concept generalization in a controlled setting, i.e., to consider increasingly difficult transfer learning scenarios.

Recall that IN-21K is built on top of the word ontology WordNet, where distinct concepts or synsets are linked according to their semantic relationships drafted by linguists. This enables the use of existing semantic similarity measures [3] that exploit the graph structure of WordNet to capture the semantic relatedness of pairs of concepts. Following prior work [12, 46], we use Lin similarity [34] to define a concept-to-concept similarity. The Lin similarity between two concepts c_1 and c_2 is given by:

$$\text{sim}_{\text{Lin}}(c_1, c_2) = \frac{2 \times \text{IC}(\text{LCS}(c_1, c_2))}{\text{IC}(c_1) + \text{IC}(c_2)}, \qquad (1)$$

where LCS denotes the lowest common subsumer of two concepts in the WordNet graph, and $\text{IC}(c) = -\log p(c)$ is the information content of a concept with probability $p(c)$ of encountering an instance of concept c in a specific corpus (in our case the subgraph of WordNet including all IN-21K concepts and their parents till the root node of WordNet: 'entity'). Following [44, 45], we define $p(c)$ as the number of concepts that exist under c divided by the total number of concepts in the corpus. An example of five concepts from IN-21K ranked by decreasing Lin similarity to the IN-1K concept "Tiger cat" is shown in Fig. 1(a).

We extend the above formulation to define the asymmetric similarity between the set of seen concepts from IN-1K, $\mathcal{C}_{\text{IN-1K}}$, and any unseen concept c as the maximum similarity between *any* concept from IN-1K and c:

$$\text{sim}_{\text{IN-1K}}(c) = \max_{\tilde{c} \in \mathcal{C}_{\text{IN-1K}}} (\text{sim}_{\text{Lin}}(c, \tilde{c})). \qquad (2)$$

While designing our benchmark, we considered different semantic similarity measures before choosing Lin similarity. We explored other measures defined on the WordNet graph [38], such as the path-based Wu-Palmer [62] and the information content-based Jiang-Conrath [25]. We also considered semantic similarities based on Word2Vec representations [40] of the titles and textual descriptions of the concepts. Our experiments with these alternative measures led to observations similar to the ones presented in Sec. 4 for Lin similarity. We refer the curious reader to the supplementary material for additional results with some of these measures.

With the similarity measure defined, our goal now is to group all eligible unseen concepts into multiple evaluation sets, which are increasingly challenging in terms of generalization. To ensure this, we would like the concepts contained in each consecutive set to be of decreasing semantic similarity to any concept from IN-1K. We achieve this by first ranking all unseen concepts with respect to their similarity to IN-1K using Eq. (2). Then, we split the ranked list into groups of consecutive concepts as shown in Fig. 2; each group corresponds to a concept generalization level.

We design our levels to be comparable to IN-1K [47], and therefore choose 1000 concepts per level. With 5146 eligible unseen concepts, we populate *five* sets. For increased diversity, we utilize the full span of the ranked list and end up with small gaps between levels (see supplementary material for more details). We denote the five concept generalization levels as $L_{1/2/3/4/5}$. Similar to [47], we further limit the maximum number of training images per concept to 1300. This brings the total number of training images per level to 1.10 million, which is close to the 1.28 million training images of IN-1K.

3.4. Evaluation protocol

We now present the protocol for ImageNet-CoG, and summarize the metrics for the different experiments presented in Sec. 4. The benchmark consists of two phases. First, a feature extraction phase, where the model trained on IN-1K is used to extract features, followed by the evaluation phase that is conducted on each level independently. An overview of the benchmark is presented in the gray box.

The ImageNet-CoG benchmark in a nutshell

Prerequisites:

A `model` pretrained on IN-1K

Sets of unseen concepts organized in levels $L_{1/2/3/4/5}$

Phase 1: Feature extraction

Use the `model` to extract image features for all image sets.

Phase 2: Evaluation

For the seen concepts (IN-1K) and for each level of unseen concepts ($L_{1/2/3/4/5}$), separately:

- Learn a linear classifier using all the training data
  < How resilient is my model to the semantic distance between seen and unseen concepts?>
- Learn a linear classifier using $N \in \{1, 2, 4, \ldots, 128\}$ samples per concept.
  < How fast can my model adapt to new concepts?>

3.4.1 Phase 1: Feature extraction

We base our protocol on the assumption that *good* visual representations should generalize to new tasks with minimal effort, i.e., without fine-tuning the backbones. Therefore, our benchmark only uses the pretrained backbones as feature extractors and decouples representation from evaluation. Concretely, we assume a model learned on the training set of IN-1K. We use this model as an encoder to extract features for images of IN-1K and of all the five levels $L_{1/2/3/4/5}$.

We extract features from the layer right before the classifiers from the respective models, following recent findings [27] that suggest that residual connections prevent backbones from overfitting to pretraining tasks. We ℓ_2-normalize the features and extract them offline: no data augmentation is applied when learning the subsequent classifiers.

3.4.2 Phase 2: Evaluation

We learn linear logistic regression classifiers for each level using all available training images. Since each level is by design a dataset approximately as big as IN-1K, we also learn linear classifiers on IN-1K with the same protocol; this allows us to compare performance across seen and unseen concepts. We also evaluate how efficiently models adapt when learning unseen concepts, i.e. how many samples they need to do so, by performing few-shot concept classification.

3.4.3 Metrics and implementation details

We report top-1 accuracy for all the experiments. Absolute accuracy numbers are comparable across IN-1K and each level by construction, since all the levels share the same number of concepts and have training sets of approximately the same size. However, we mostly plot accuracy *relative to a baseline model*, for two reasons: (i) it makes the plots clearer and the differences easier to grasp, (ii) the performance range at each level is slightly different so it helps visualizing the trends better.

To create the train/test split, we randomly select 50 samples as the test set for each concept and use the remaining ones (at least 732, at most 1300) as a training set. We use part of the training data to optimize the hyper-parameters of the logistic regression for each level; see details in Sec. 4.

We use Optuna [1] to optimize the learning rate and weight decay hyper-parameters for every model and every level; we use 20% of the training sets as a validation set to find the best configuration and then re-train using the complete training set. We report results only on the test sets. We repeat the hyper-parameter selection 5 times with different seeds, and report the mean of the final scores; standard deviation is also presented in all figures.

4. Evaluating models on ImageNet-CoG

We now present our large-scale experimental study which analyzes how different CNN-based and transformer-based visual representation models behave on our benchmark, following the evaluation protocol defined in the previous section. For clarity, we only highlight a subset of our experiments and provide additional results in the supplementary material.

4.1. Models

We choose 31 models to benchmark and present the complete list in Tab. 1. To ease comparisons and discussions, we split the models into the following four categories.

Architecture. We consider several architectures including CNN-based (*a*-VGG19 [50], *a*-Inception-v3 [53], ResNet50, *a*-ResNet152 [23]), transformer-based (*a*-DeiT-S [58], *a*-DeiT-S-distilled, *a*-DeiT-B-distilled, *a*-T2T-ViT-t-14 [68]) and neural architecture search (*a*-NAT-M4 [36], *a*-EfficientNet-B1 [54], *a*-EfficientNet-B4 [54]) backbones with varying complexities. We color-code the models in this category into two groups, depending on whether their number of parameters are comparable to ResNet50 (red) or not (orange); If they do, they are also directly comparable to all models from the following categories.

Self-supervision. ResNet50-sized models trained in a self-supervised manner (in blue) include contrastive (*s*-SimCLR-v2 [7, 8], *s*-MoCo-v2 [9, 22], *s*-InfoMin [56], *s*-MoCHi [26], *s*-BYOL [19]), clustering-based (*s*-SwAV [5], *s*-OBoW [17], *s*-DINO [6]), feature de-correlation (*s*-BarlowTwins [72]), and distilled (*s*-CompReSS [28]) models.

Regularization. ResNet50-sized models with label regularization techniques (in purple) applied during the training phase include distillation (*r*-MEAL-v2 [49]), label augmentation (*r*-MixUp [74], *r*-Manifold-MixUp [59], *r*-CutMix [69]

Model	Notes (optionally # param. / amount of extra data)
Reference model: ResNet50 [23]	
ResNet50	Baseline model from the torchvision package (23.5M)
Architecture: Models with different backbone	
a-T2T-ViT-t-14 [68]	Visual transformer (21.1M)
a-DeiT-S [58]	Visual transformer (21.7M)
a-DeiT-S-distilled [58]	Distilled *a*-DeiT-S (21.7M)
a-Inception-v3 [53]	CNN with inception modules (25.1M)
a-NAT-M4 [36]	Neural architecture search model (7.6M)
a-EfficientNet-B1 [54]	Neural architecture search model (6.5M)
a-EfficientNet-B4 [54]	Neural architecture search model (17.5M)
a-DeiT-B-distilled [58]	Bigger version of *a*-DeiT-S-distilled (86.1M)
a-ResNet152 [23]	Bigger version of ResNet50 (58.1M)
a-VGG19 [50]	Simple CNN architecture (139.6M)
Self-supervision: ResNet50 models trained in this framework	
s-SimCLR-v2 [7, 8]	Online instance discrimination (ID)
s-MoCo-v2 [9, 22]	ID with momentum encoder and memory bank
s-BYOL [19]	Negative-free ID with momentum encoder
s-MoCHi [26]	ID with negative pair mining
s-InfoMin [56]	ID with careful positive pair selection
s-OBoW [17]	Online bag-of-visual-words prediction
s-SwAV [5]	Online clustering
s-DINO [6]	Online clustering
s-BarlowTwins [72]	Feature de-correlation using positive pairs
s-CompReSS [28]	Distilled from SimCLR-v1 [7] (with ResNet50x4)
Regularization: ResNet50 models with additional regularization	
r-MixUp [74]	Label-associated data augmentation
r-Manifold-MixUp [59]	Label-associated data augmentation
r-CutMix [69]	Label-associated data augmentation
r-ReLabel [70]	Trained on a "multi-label" version of IN-1K
r-Adv-Robust [48]	Adversarially robust model
r-MEAL-v2 [49]	Distilled ResNet50
Use of web data: ResNet50 models using additional data	
d-MoPro [32]	Trained on WebVision-V1 ($\sim 2\times$)
d-Semi-Sup [65]	Pretrained on YFCC-100M ($\sim 100\times$), fine-tuned on IN-1K
d-Semi-Weakly-Sup [65]	Pretrained on IG-1B ($\sim 1000\times$), fine-tuned on IN-1K
d-CLIP [43]	Trained on WebImageText ($\sim 400\times$)

Table 1: **List of models evaluated on ImageNet-CoG.**

and *r*-ReLabel [70]) and adversarial robustness (*r*-Adv-Robust [48]) models.

Use of web data. Models pretrained using additional web data with noisy labels are color-coded in green. This includes student-teacher models *d*-Semi-Sup [65] and *d*-Semi-Weakly-Sup [65], which are first pretrained on YFCC-100M [55] (100x the size of IN-1K) and IG-1B [37] (1000x) and then fine-tuned on IN-1K. We also consider cross-modal *d*-CLIP [43] pretrained on WebImageText (400x) with textual annotations, and noise tolerant tag prediction model *d*-MoPro pretrained on WebVision-V1 [33] (2x). As it is not clear if YFCC-100M, IG-1B, WebImageText or WebVision-V1 contain images of the unseen concepts we selected in the levels, *models in this category are not directly comparable*.

We use publicly available models provided by the corresponding authors for all these approaches. All the models, with the exception of those in the use-of-web-data category, are only pretrained on IN-1K. We also use the best ResNet-50 backbones released by the authors for all the ResNet-based models. We use the vanilla ResNet50 (the version available in the torchvision package) as a reference point, which makes cross-category comparisons easier. We prefix models' names with the category identifiers for clarity.

4.2. Results

We measure image classification performance on IN-1K and each of the concept generalization levels $L_{1/2/3/4/5}$ of ImageNet-CoG for the 31 models presented above, using a varying number of images per concept. These experiments allow us to study (i) how classification performance changes as we semantically move away from the seen concepts (Sec. 4.2.1), and (ii) how fast models can adapt to unseen concepts (Sec. 4.2.2). We refer the reader to Sec. 3.4 for the justification of our protocol and the choice of metrics.

4.2.1 Generalization to unseen concepts

We report the performance of linear classifiers learnt with all the training data in Fig. 3. In Fig. 3(a) we report top-1 accuracy for all models and levels, while Fig. 3(b)-(e) present performance relative to the baseline ResNet50 across the 4 model categories. Our main observations are as follows.

* ***It is harder to generalize to semantically distant concepts.*** The absolute performance of all models monotonically decreases as we move away semantically from IN-1K. This implies that transfer learning becomes more and more challenging on levels from L_1 to L_5, i.e., as we try to distinguish concepts that are further from the training ones.

* ***Self-supervised models excel at concept generalization.*** Many recent self-supervised models (*s*-DINO, *s*-SwAV, *s*-BYOL, *s*-OBoW and *s*-SimCLR-v2) outperform ResNet50 on all levels. In general, we see that the performance gaps between ResNet50 and self-supervised models progressively shift in favor of the latter (Fig. 3(b)). Surprisingly, from Fig. 3(a) we also see that a ResNet50 trained with *s*-DINO competes with the top-performing models on L_5 *across all categories and model sizes*. This shows that augmentation invariances learned by the model transfer well to images of unseen concepts.

* ***Visual transformers overfit more to seen concepts*** (for models with as many parameters as ResNet50). The top-performing model of the study overall is *a*-DeiT-B-distilled, a large visual transformer. However, for the same number of parameters as ResNet50, we see that the large gains that visual transformers like *a*-DeiT-S and *a*-T2T-ViT-t-14 exhibit on IN-1K are practically lost for unseen concepts (red lines in Fig. 3(e)). In fact, both end up performing slightly worse than ResNet50 on L_5.

* ***Using noisy web data highly improves concept generalization.*** Weakly-supervised models *d*-Semi-Sup, *d*-Semi-Weakly-Sup and *d*-CLIP pretrained with roughly 100x, 1000x, and 400x more data than IN-1K exhibit improved performance over ResNet50 on all levels (Fig. 3(d)). It is worth re-stating, however, that since their datasets are web-based and much larger than IN-1K, we cannot confidently claim that concepts in our levels are indeed *unseen* during training. Results on this model category should therefore be

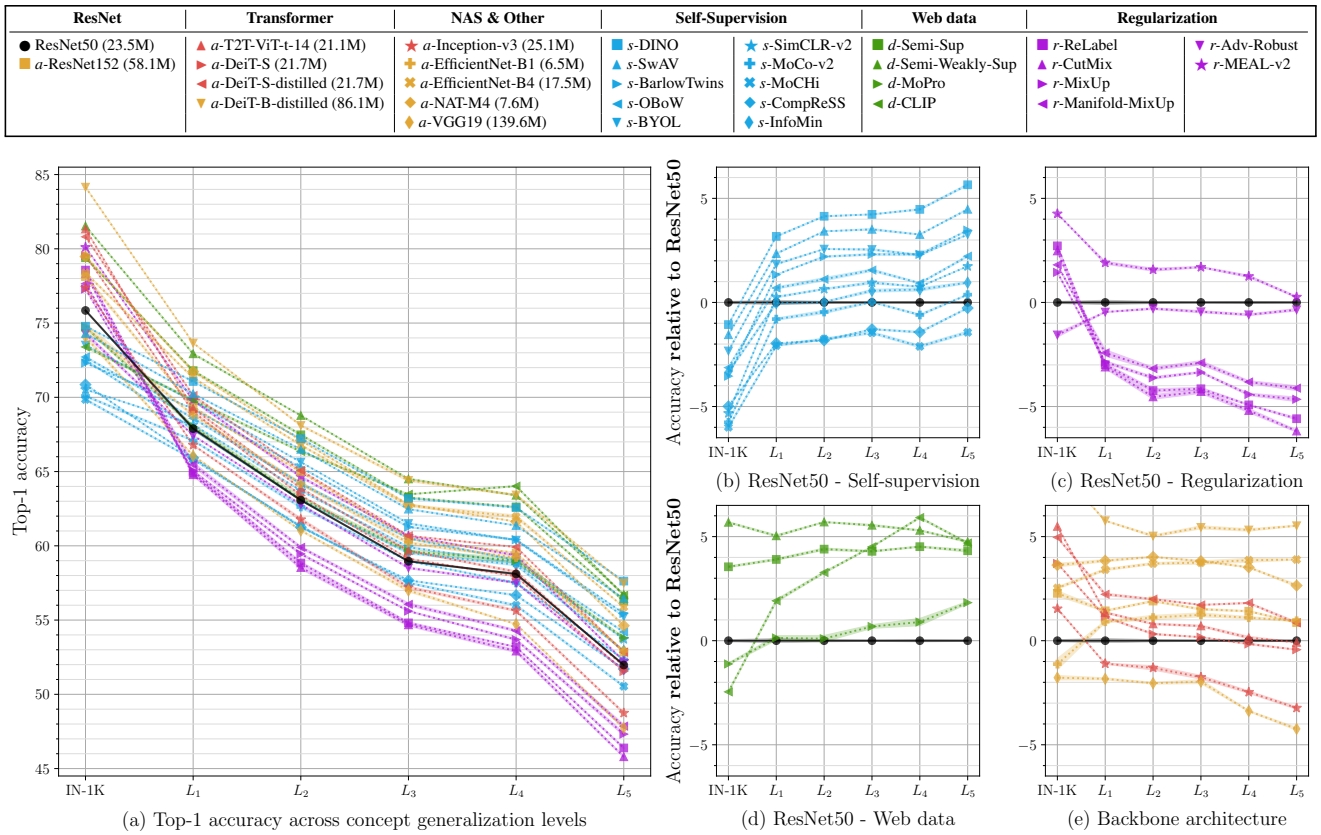


Figure 3: **Linear classification on ImageNet-CoG.** Top-1 accuracies for all the 31 models listed in Tab. 1 after training logistic regression classifiers on IN-1K and each level $L_{1/2/3/4/5}$. (a) Absolute top-1 accuracy on all levels. (b)-(e) accuracy relative to the baseline ResNet50 for all the models, split across the four model categories presented in Sec. 4.1.

taken with a pinch of salt.

* ***Model distillation generally improves concept generalization performance.*** We see that distilled supervised models *r*-MEAL-v2 and *a*-DeiT-S-distilled consistently improve over their undistilled counterparts on all levels (Figs. 3(c) and (e)). However, these gains decrease progressively, and for L_5 performance gains over the baseline are small. It is also worth noting that adversarial training (*r*-Adv-Robust) does not seem to hurt concept generalization.

* ***Neural architecture search (NAS) models seem promising for concept generalization.*** All NAS models we evaluate (*a*-EfficientNet-B1, *a*-EfficientNet-B4 and *a*-NAT-M4) exhibit stable gains over the baseline ResNet50 on all levels (Fig. 3(e)), showing good concept generalization capabilities. Among them, *a*-NAT-M4, a NAS model tailored for transfer learning with only 7.6M parameters achieves particularly impressive performance over all levels including IN-1K.

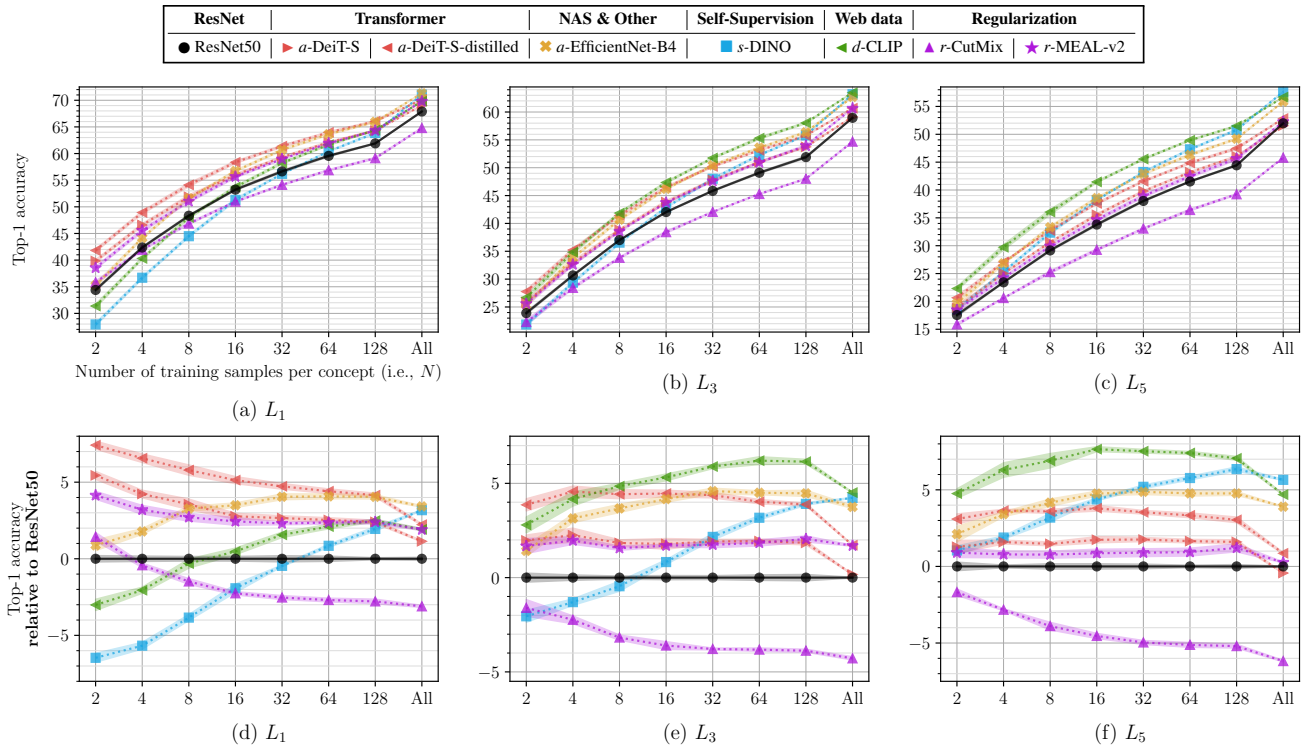
* ***Label-associated augmentation techniques deteriorate concept generalization performance.*** Although methods like *r*-MixUp, *r*-Manifold-MixUp, *r*-ReLabel and *r*-CutMix exhibit strong performance gains over ResNet50 on IN-1K, i.e., for concepts seen during training, Fig. 3(c) shows that such gains do not transfer when generalizing to unseen ones. They appear to overfit more to the seen concepts.

* ***What are the top-performing models overall for concept generalization?*** From Fig. 3(a) we see that better and larger architectures and models using additional data are on top for L_3-L_5. However, it is impressive how *s*-DINO, a contrastive self-supervised model, is among the top methods, outperforming the vast majority of models at the most challenging levels.

4.2.2 How fast can models adapt to unseen concepts?

We now study few-shot classification, i.e., training linear classifiers with $N = \{2, 4, 8, 16, 32, 64, 128\}$ samples per concept. For clarity, we selected a subset of the models and in Fig. 4 we present their performance on L_1, L_3 and L_5. The complete set of results for all models and levels is given in the supplementary material. We discuss the most interesting observations from Fig. 4 below.

* ***Transformer-based models are strong few-shot learners.*** Transformer-based models exhibit consistent gains over



(a) L_1 (b) L_3 (c) L_5

(d) L_1 (e) L_3 (f) L_5

Figure 4: **Few-shot linear classification on ImageNet-CoG.** Top-1 accuracies for a subset of the models listed in Tab. 1 after training logistic regression classifiers on L_1, L_3, L_5 using $N = \{2, 4, 8, 16, 32, 64, 128\}$ training samples per concept. Performance when using all the samples is also shown for reference. (a)-(c): Absolute top-1 accuracy. (d)-(f) accuracy relative to the baseline ResNet50. *The complete set of results for all the 31 models and levels is in the supplementary material.*

ResNet50 on all levels when $N \leq 128$. Despite the fact that performance gains from transformers diminish when using all available images on L_5, they exhibit a consistent 3-4% accuracy gain over ResNet50 for $N \leq 128$ (Fig. 4(f)).

* ***Model Distillation and Neural Architecture Search (NAS) exhibit consistent gains also in low-data regimes.*** The NAS-based *a*-EfficientNet-B4 model exhibits consistently higher performance than ResNet50 on all levels for all N. The same stands for the distilled *r*-MEAL-v2 and *a*-DeiT-S-distilled that are also consistently better than their undistilled counterparts for all N and all levels.

* ***Bigger models and additional web data help at few-shot learning.*** This is an observation from the extended set of figures (see supplementary material). Bigger models have consistent gains in low-data regimes. The same stands for models with additional web data. Moreover, as we go towards semantically dissimilar concepts, *a*-NAT-M4 outperforms all other methods and it even challenges the much bigger *a*-DeiT-B-distilled model.

5. Conclusion

In this paper, we studied concept generalization through the lens of our new ImageNet-CoG benchmark. It is designed to be used out-of-the-box with IN-1K pretrained models. We evaluated a diverse set of 31 methods representative of the recent advances in visual representation learning.

Our extensive analyses show that self-supervised learning produces representations that generalize surprisingly better than any supervised model with the same number of parameters. We see that the current transformer-based models appear to overfit to seen concepts, unlike neural architecture-search-based models. The latter outperform several other supervised learning models with far less parameters.

We also studied how fast models can adapt to unseen concepts by learning classifiers with only a few images per class. In this setting, we verify that visual transformers are strong few-shot learners, and show how distillation and neural architecture search methods achieve consistent gains even in low-data regimes.

We envision ImageNet-CoG to be an easy-to-use evaluation suite to study one of the most important aspects of generalization in a controlled and principled way.

Acknowledgements. This work was supported in part by MIAI@Grenoble Alpes (ANR-19-P3IA-0003), and the ANR grant AVENUE (ANR-18-CE23-0011).

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proc. ICKDDM*, 2019. 5

[2] Tamara Berg and Alexander Berg. Finding iconic images. In *Proc. CVPRW*, 2009. 3

[3] Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *CL*, 32(1), 2006. 4

[4] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised Pre-Training of Image Features on Non-Curated Data. In *Proc. ICCV*, 2019. 3

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proc. NeurIPS*, 2020. 3, 5, 6

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 5, 6

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, 2020. 2, 3, 5, 6

[8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *Proc. NeurIPS*, 2020. 3, 5, 6

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 5, 6

[10] Gabriela Csurka, editor. *Domain Adaptation in Computer Vision Applications*. Advances in Computer Vision and Pattern Recognition. Springer, 2017. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009. 1, 2, 3

[12] Thomas Deselaers and Vittorio Ferrari. Visual and semantic similarity in imagenet. In *Proc. CVPR*, 2011. 2, 3, 4

[13] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *Proc. CVPR*, 2021. 3

[14] Mark Everingham, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2007 Results. 3

[15] Andrea Frome, Greg Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Proc. NeurIPS*, 2013. 3

[16] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Proc. NeurIPS*, 2018. 2

[17] Spyros Gidaris, Andrei Bursuc, Gilles Puy, Nikos Komodakis, Matthieu Cord, and Patrick Pérez. Online bag-of-visual-words generation for unsupervised representation learning. In *Proc. CVPR*, 2021. 5, 6

[18] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proc. ICCV*, 2019. 2, 3

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Proc. NeurIPS*, 2020. 3, 5, 6

[20] Yunhui Guo, Noel Codella, Leonid Karlinsky, John Smith, Tajana Rosing, and Rogerio Feris. A new benchmark for evaluation of cross-domain few-shot learning. In *Proc. ECCV*, 2020. 2, 3

[21] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proc. ICCV*, 2017. 3

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. CVPR*, 2020. 2, 3, 5, 6

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 3, 5, 6

[24] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. In *Proc. NeurIPS*, 2014. 3

[25] Jay Jiang and David Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. ROCLING*, 1997. 4

[26] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *Proc. NeurIPS*, 2020. 5, 6

[27] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting Self-Supervised Visual Representation Learning. In *Proc. CVPR*, 2019. 5

[28] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Compress: Self-supervised learning by compressing representations. In *Proc. NeurIPS*, 2020. 5, 6

[29] Simon Kornblith, Jonathon Shlens, and Quoc Le. Do better imagenet models transfer better? In *Proc. CVPR*, 2019. 2, 3

[30] Christoph Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. CVPR*, 2009. 3

[31] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Proc. NeurIPS*, 2018. 2

[32] Junnan Li, Caiming Xiong, and Steven CH Hoi. Mopro: Webly supervised learning with momentum prototypes. In *Proc. ICLR*, 2020. 6

[33] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017. 6

[34] Dekang Lin. An information-theoretic definition of similarity. In *Proc. ICML*, 1998. 1, 4

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. ECCV*, 2014. 3

[36] Zhichao Lu, Gautam Sreekumar, Erik Goodman, Wolfgang Banzhaf, Kalyanmoy Deb, and Vishnu Naresh Boddeti. Neural architecture transfer. *PAMI*, 2021. 5, 6

[37] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proc. ECCV*, 2018. 2, 6

[38] Lingling Meng, Runqing Huang, and Junzhong Gu. A review of semantic similarity measures in wordnet. *IJHIT*, 6(1), 2013. 4

[39] Elad Mezuman and Yair Weiss. Learning about canonical views from internet image collections. In *Proc. NeurIPS*, 2012. 3

[40] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proc. NeurIPS*, 2013. 3, 4

[41] George A Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11), 1995. 2, 3

[42] Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring generalization in deep learning. In *Proc. NeurIPS*, 2017. 2

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021. 6

[44] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. IJCAI*, 1995. 4

[45] Marcus Rohrbach, Michael Stark, and Bernt Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Proc. CVPR*, 2011. 4

[46] Marcus Rohrbach, Michael Stark, György Szarvas, Iryna Gurevych, and Bernt Schiele. What helps where–and why? Semantic relatedness for knowledge transfer. In *Proc. CVPR*, 2010. 4

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3), 2015. 1, 2, 3, 4

[48] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust ImageNet models transfer better? In *Proc. NeurIPS*, 2020. 6

[49] Zhiqiang Shen and Marios Savvides. Meal v2: Boosting vanilla resnet-50 to 80%+ top-1 accuracy on imagenet without tricks. In *Proc. NeurIPSW*, 2020. 3, 5, 6

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2015. 3, 5, 6

[51] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Proc. NeurIPS*, 2013. 2

[52] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15(1), 2014. 2

[53] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proc. CVPR*, 2016. 3, 5, 6

[54] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proc. ICML*, 2019. 5, 6

[55] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Commun. ACM*, 59(2), 2016. 6

[56] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Proc. NeurIPS*, 2020. 5, 6

[57] Antonio Torralba and Alexei Efros. Unbiased look at dataset bias. In *Proc. CVPR*, 2011. 3

[58] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proc. ICML*, 2021. 3, 5, 6

[59] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *Proc. ICML*, 2019. 3, 5, 6

[60] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Proc. NeurIPS*, 2016. 2

[61] Bram Wallace and Bharath Hariharan. Extending and analyzing self-supervised learning across domains. In *Proc. ECCV*, 2020. 3

[62] Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. *ACL*, 1994. 4

[63] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. *PAMI*, 41(9), 2018. 3

[64] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proc. CVPR*, 2010. 3

[65] Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. 3, 6

[66] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proc. FAT*, 2020. 3

[67] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Proc. NeurIPS*, 2014. 2, 3

[68] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021. 5, 6

[69] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proc. ICCV*, 2019. 2, 3, 5, 6

[70] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *Proc. CVPR*, 2021. 6

[71] Amir Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proc. CVPR*, 2018. 2

[72] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proc. ICML*, 2021. 5, 6

[73] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 3

[74] Hongyi Zhang, Moustapha Cisse, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. ICLR*, 2018. 3, 5, 6

[75] Nanxuan Zhao, Zhirong Wu, Rynson W. H. Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? In *Proc. ICLR*, 2021. 2

[76] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *PAMI*, 2017. 3