Revisiting Weakly Supervised Pre-Training of Visual Perception Models

Mannat Singh Laura Gustafson Aaron Adcock Vinicius de Freitas Reis Bugra Gedik
Raj Prateek Kosaraju Dhruv Mahajan Ross Girshick Piotr Dollár Laurens van der Maaten
Meta AI

https://github.com/facebookresearch/SWAG

Abstract

Model pre-training is a cornerstone of modern visual recognition systems. Although fully supervised pre-training on datasets like ImageNet is still the de-facto standard, recent studies suggest that large-scale weakly supervised pre-training can outperform fully supervised approaches. This paper revisits weakly-supervised pre-training of models using hashtag supervision with modern versions of residual networks and the largest-ever dataset of images and corresponding hashtags. We study the performance of the resulting models in various transfer-learning settings including zero-shot transfer. We also compare our models with those obtained via large-scale self-supervised learning. We find our weakly-supervised models to be very competitive across all settings, and find they substantially outperform their self-supervised counterparts. We also include an investigation into whether our models learned potentially troubling associations or stereotypes. Overall, our results provide a compelling argument for the use of weakly supervised learning in the development of visual recognition systems. Our models, Supervised Weakly through hashtAGs (SWAG), are available publicly.

1. Introduction

Most modern visual-recognition systems are based on machine-learning models that are pre-trained to perform a task that is different from the downstream task that the system aims to solve. Such pre-training allows the system to leverage (annotated) image or video datasets that are much larger than the datasets available for the downstream task. Arguably the most popular pre-training task is supervised image classification on datasets such as ImageNet and JFT [20, 42, 76], but recent studies have also explored self-supervised [11–14, 27, 29, 31, 51] and weakly supervised [37, 38, 44, 49, 57] tasks for pre-training.

There are trade-offs between these three types of pre-training. Fully supervised pre-training benefits from a strong semantic learning signal for each training example, but does not scale well because manual labeling of training data is time-consuming. By contrast, self-supervised pre-training receives hardly any semantic information on the training examples, but can be scaled to billions of training examples relatively easily [27, 31]. Weakly-supervised approaches fall somewhere in between: for example, hashtags or other text associated with visual data generally provide a noisy semantic learning signal but can be obtained at large scale with relative ease [49, 57].

Following the success of prior work [49], this paper performs an in-depth study of weakly-supervised pre-training using hashtag supervision. We pre-train modern image-recognition models on the largest-ever-dataset of images and associated hashtags, and evaluate the resulting models in a range of transfer-learning experiments. Specifically, we transfer our models to a variety of image-classification tasks and evaluate the performance of the resulting models. We also evaluate the models in *zero-shot transfer* and *few-shot transfer* settings [57]: that is, we evaluate the "off-the-shelf performance" of these models without finetuning them on the target tasks. The overall goal of our study is to shed light on the trade-offs between fully supervised, self supervised, and weakly supervised pre-training. Throughout our experiments, we find the weakly-supervised approach to be very competitive: our best models perform on par with the state-of-the-art on a range of visual-perception tasks despite employing a relatively simple training pipeline.

A potential downside of weakly-supervised pre-training is that models may inherit or amplify harmful associations from the underlying supervisory signal. We perform a series of experiments aimed at assessing the extent to which this happens. Our results do not provide conclusive answers, but they do suggest that the risks involved may not be as large as in language modeling [6, 9]. Overall, we believe our study presents a compelling argument for weakly-supervised pre-training of visual-recognition systems.

2. Related Work

This study is part of a large body of work on pre-training models for visual recognition. This body of work can be subdivided into three key groups.

Fully supervised pre-training was pioneered by [19, 59] and is now the de-facto standard approach to a variety of visual-recognition tasks, including fine-grained image classification [30, 62], object detection [61], image segmentation [32, 70], image captioning [46], visual question answering [40], video classification [21], *etc.* The ImageNet-1K dataset [63] is by far the most commonly used image dataset for pre-training, whereas the Kinetics dataset [39] is often used for pre-training of video-recognition models. Some recent studies have also used the much larger JFT-300M [20] and JFT-3B [76] image datasets, but not much is known publicly about those datasets. The effectiveness of supervised pre-training has been the subject of a number of studies, in particular, [1, 42, 60] analyze the transfer performance of supervised pre-trained models.

Self-supervised pre-training has seen tremendous progress in recent years. Whereas early self-supervised learners such as RotNet [26] or DeepCluster [10] substantially lagged their supervised counterparts in vision pre-training, more recent approaches have become quite competitive. These approaches learn to predict clusters [11], use contrastive learning [13, 31, 51], or use student-teacher architectures in which the teacher is an exponentially moving average of the student [12, 14, 29]. A key advantage of self-supervised pre-training is that is can easily be scaled to billions of training images: several studies have shown that scaling self-supervised learning can lead to substantial performance improvements [27, 31].

Weakly-supervised pre-training has not received nearly as much attention as the other two pre-training paradigms, but has shown very promising performance nonetheless. Whereas early studies that pre-trained models by predicting words [38] or n-grams [44] in image captions were not very competitive because of the limited scale of their training data, recent weakly-supervised pre-training methods are much more competitive on a range of visual-recognition tasks [5, 25, 37, 49, 56, 57]. In particular, ALIGN [37] and CLIP [57] pre-train vision-and-language models on large numbers of images and associated captions, and successfully perform *zero-shot transfer* to new recognition tasks.

Our study builds on [49], which trained convolutional networks on billions of images to predict associated hashtags. Compared to [49], our study: (1) trains larger models with more efficient convolutional and transformer architectures on a much larger dataset, (2) studies the performance of the resulting models in zero-shot transfer settings in addition to standard transfer-learning experiments, (3) performs comparisons of our models with state-of-the-art self-supervised learners, and (4) presents an in-depth study of potential harmful associations that models may adopt from the weak supervision they receive. Despite the conceptual similarities in our approach, our best model achieves an ImageNet-1K validation accuracy that is more than 3% higher than that reported in [49].

3. Pre-Training using Hashtag Supervision

Our weakly supervised pre-training methodology is based on hashtag supervision. We train image-recognition models to predict the hashtags that were assigned to an image by the person who posted the image. Hashtag prediction has great potential as a pre-training task because hashtags were assigned to images to make them *searchable*, *i.e.*, they tend to describe some salient semantic aspects of the image. While hashtag prediction is conceptually similar to image classification, it differs in a few key ways [16, 49, 68]:

1. Hashtag supervision is inherently noisy. Whilst some hashtags describe visual content in the image (*e.g.*, `#cat`), other hashtags may be unrelated to the visual content (*e.g.*, `#repost`). Different hashtags may be used to describe the same visual content, or the same hashtag may be used to describe different visual content. Importantly, hashtags generally do not provide a comprehensive annotation of the visual content of an image, that is, there tend to be many false negatives.
2. Hashtag usage follows a Zipfian distribution [50]; see Figure 1. This implies that the learning signal follows a very different distribution than is common in image-recognition datasets like ImageNet [63], which tend to have a class distribution that is more or less uniform.
3. Hashtag supervision is inherently *multi-label*: a single image generally has multiple hashtags associated with it that all serve as positive classification targets.

Our data pre-processing and model pre-training procedures are designed to (partly) address these issues. We describe them in more detail in Section 3.1 and 3.2, respectively.

3.1. Hashtag Dataset Collection

We follow [49] in constructing a dataset of public Instagram photos and associated hashtags. We adopt the following four steps in constructing the pre-training dataset:

1. Construct a hashtag vocabulary by selecting frequently used hashtags and canonicalizing them.
2. Gather publicly available images that are tagged with at least one of the selected hashtags.
3. Combine the resulting images and associated hashtags into labeled examples that can be used for pre-training.
4. Resample the resulting examples to obtain the desired hashtag distribution.

Next, we describe each of these steps in detail.

Hashtag vocabulary. We select hashtags used more than once in public Instagram posts by US users. Next, we filter

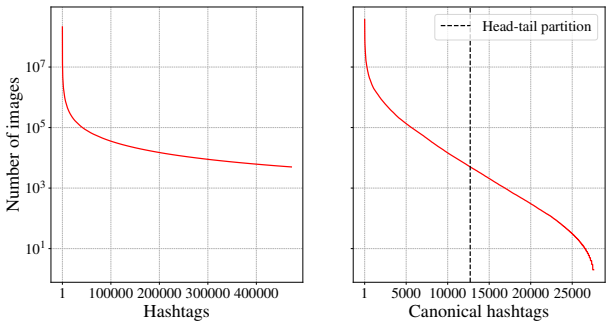

Figure 1. Hashtag distribution of Instagram images. **Left:** Frequency of all hashtags occurring with public images posted by US users. **Right:** Frequency of filtered and canonicalized hashtags occurring with public images by users from all countries. We define the head as the set of canonical hashtags associated with more than 5,000 images; the remaining hashtags form the tail.

out and canonicalize the hashtags using WordNet synsets [22]. More details about this process are in Appendix A. This results in a label set, $\mathcal{C}$, that contains ∼27k *canonical hashtags* that correspond to a set of ∼75k raw hashtags, where multiple hashtags can map to a single canonical hashtag (*e.g.*, `#dog` and `#canine`). We drop the "canonical" qualifier when it is obvious from the context. As the exact images in the dataset may change with time, the number of canonical hashtags varies between 27k and 28k across experiments. The hashtag selection and canonicalization reduces some of the inherent noise in the supervisory signal.

Image collection and labeling. We collect all public Instagram images that have at least one hashtag from our vocabulary.[1] The images were subjected to an array of automated filters designed to remove potentially offensive content. While certainly not perfect, this substantially reduces the issues that plague other large image datasets [8, 55]. We construct a multi-label dataset using these images by converting all hashtags into their corresponding canonical targets (note that a single image may have multiple hashtags). Hashtags that are not in the vocabulary are discarded.

Resampling. We adopt a resampling procedure similar to [49] to generate our final pre-training examples. The resampling procedure aims to down-weight frequent hashtags whilst up-weighting infrequent hashtags in the pre-training task. We do so by resampling according to the inverse square root of the hashtag frequency. Unlike [49], we additionally upsample (with replacement) the long tail of images with at least one infrequent hashtag by ∼100×. Herein, we define infrequent hashtags as those that occur with fewer than 5,000 images (see Figure 1). The resulting resampled dataset comprises 30% tail images and 70% head images (see Appendix A for more details).

[1]We downloaded images from all countries, but excluded images by users from particular countries to comply with applicable regulations.

We note that this means that in a single training epoch, each unique tail image appears multiple times. This implies there is a discrepancy between the number of *unique* images in an epoch and the number of *total samples* processed in that epoch. We label our dataset by the number of unique images in the dataset: our IG-3.6B dataset has ∼3.6 billion unique images. However, a single training epoch over that dataset processes ∼5 billion samples due to our re-sampling procedure. This is different from other datasets we compare with (*e.g.*, JFT-300M) in which the unique number of images equals the total samples processed in an epoch.

3.2. Pre-Training Procedure

In preliminary experiments (Appendix C.1), we studied image-recognition models including ResNeXt [74], RegNetY [58], DenseNet [35], EfficientNet [65], and ViT [20] . We found RegNetY and ViT models to be most competitive, and focus on those in the experiments presented here.

During pre-training, we equip our models with an output linear classifier over $|\mathcal{C}| \approx$ 27k classes. For ViTs we use an additional linear layer with output dimension equal to the input dimension, similar to [20]. Following [49], we use a softmax activation and train the model to minimize the cross-entropy between the predicted probabilities and the target distribution. Each target entry is either $1/K$ or 0 depending on whether the corresponding hashtag is present or not, where K is the number of hashtags for that image.

All our RegNetY models were trained using stochastic gradient descent (SGD) with Nesterov momentum of 0.9. We employed a half-cosine learning rate schedule [48] with a base initial value of 0.1 for a batch size of 256 and a final value of 0. We used a weight decay of 10^{-5}, but disabled weight decay in batch-normalization layers: preliminary experiments suggested that batch-normalization weight decay is effective when pre-training on ImageNet-1k, but significantly degrades results on larger datasets such as IG-3.6B.

Our ViT models were trained using AdamW [47] with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. We used an initial learning rate of $4 \cdot 10^{-4}$, a batch size of 8,192, and a weight decay of 0.1.

Following [28], we scale the initial learning rate linearly with the batch size when doing distributed training. We "warm up" the learning rate for the first 5% of training updates by linearly increasing the learning rate from $1/10$-th of the initial value to the initial value. Similar to [28], we find that performance degrades for batch sizes larger than 8,192 so we did not increase our batch size further.

We trained our models using mixed-precision training on images that were pre-processed to 224×224 resolution using a standard random-resize crop followed by a random horizontal flip. In preliminary experiments, we also evaluated several other training approaches that provide gains in ImageNet-1k pre-training [18, 65], including exponential moving averages [54], mixup [77], label smoothing [52],

AutoAugment [15], and stochastic depth [36]. However, we did not find those approaches to lead to performance improvements; some even deteriorated performance.

We trained our largest model for 2 epochs of the IG-3.6B dataset (10 billion samples seen) using 128 Nvidia V100 32GB GPUs across 16 nodes. The nodes were connected via Ethernet, with 8 GPUs / node connected via NVLink.

4. Experiments

We performed a series of experiments to test the efficacy of our hashtag-based pre-training strategy. We compare our weakly supervised models in transfer-learning experiments with modern supervised (Section 4.2) and self-supervised models (Section 4.3), and with other weakly supervised models in zero-shot transfer (Section 4.4).

4.1. Experimental Setup

In our experiments, we focus on different types of transfer learning to image-classification tasks. Specifically, we study: (1) transfer learning using linear classifiers, (2) transfer learning using finetuning, (3) zero-shot transfer learning, and (4) few-shot transfer learning. We compare the efficacy of our pre-training strategy with that of fully supervised (4.2) and self-supervised (4.3) pre-training strategies.

Datasets. We perform experiments in which we transfer models to ImageNet classification [63] on ImageNet-1k (1.28M training images, 50,000 validation images, 1,000 classes), and ImageNet-5k (6.57M training images, 250,000 validation images, 5,000 classes) as defined in [49, 74]. We also perform experiments in which we transfer pre-trained models to other commonly used image-classification benchmarks, including the iNaturalist 2018 [67], Places365-Standard [79], and Caltech-UCSD Birds-200-2011 (CUB-2011) [69] datasets.

Finetuning. We follow [41] in finetuning our pre-training models for downstream tasks. We finetune the models using SGD with a batch size of 512 and a half-cosine learning rate schedule [48]. The initial value was tuned for every each model-task combination separately via grid-search. We did not use weight decay during finetuning. We finetune RegNetY and ViT B/16 models using an image resolution of 384×384, and ViT L/16 and H/14 models with larger 512×512 and 518×518 resolutions respectively – higher resolutions help these models significantly. For EfficientNets, we use the pre-training resolution for finetuning. For "large" transfer datasets (defined as datasets with $N > 500{,}000$ examples), we finetune for 20,000 parameter updates; for "medium" datasets ($20{,}000 < N \leq 500{,}000$ examples), we finetune for 10,000 steps; and for "small" datasets ($N \leq 20{,}000$ examples), we finetune for 500 steps. We use mixup [77] with $\alpha = 0.1$ during finetuning on all datasets. We used synchronous batch normalization across GPUs, as it improves transfer performance (see appendix).

For ImageNet-1k finetuning, we additionally compute an exponential moving average (EMA) of the parameters during training with a decay rate of 10^{-4} and use the averaged weights for inference [54]. We found this improved the top-1 accuracy for our best RegNetY and ViT models by 0.2%. Lastly, we finetuned ViTs for 28 epochs on ImageNet-1k since the longer schedule helped improve performance.

During evaluation, we resize the smaller side of the image to the final resolution and then take a center crop of the same size (*e.g.*, resize smaller side to 224 then 224×224 center crop). This differs from standard practice [66] but gives a boost of 0.1% to 0.5% on the ImageNet-1k dataset.

4.2. Comparison with Supervised Pre-Training

We compare our weakly supervised RegNetY and ViT models with state-of-the-art supervised EfficientNets [72, 73] and ViTs [20, 76] in transfer-learning experiments on five datasets: (1) ImageNet-1k , (2) ImageNet-5k, (3) iNaturalist, (4) Places365, and (5) CUB-2011. We finetune all models (see 4.1) on the training split of the transfer dataset and measure the classification accuracy of the finetuned models on the validation or test split.

Table 1 presents an overview of the results of these experiments. For each model, the table shows the pre-training dataset used, the image resolution used during pre-training and finetuning, the inference throughput of the model, the number of FLOPs and parameters in the finetuned model, and the test accuracy on the transfer datasets. We do not report results for an approach when its pre-trained model and pre-training dataset are not publicly available. In the table, accuracies that we adopted from the original paper are *italicized*. For the ImageNet-1k dataset, we report both results reported in the original papers and results we obtained when we reproduced the model. We **boldface** the best result and underline the second-best result for each dataset. Table 1 groups models into supervised and weakly supervised. In this grouping, we consider pre-training on JFT datasets to be supervised pre-training but we acknowledge that little is known on how these datasets were collected: [76] refers to the JFT-3B dataset as "weakly labeled" and "noisy" but also states that semi-automatic annotation was used to collect it. This suggests that JFT datasets were manually curated and annotated, which is why we consider them as supervised.[2]

The results in Table 1 show that our weakly-supervised models are very competitive: they achieve the best or second-best accuracy on all five transfer datasets. We note that models pre-trained on IN-1k datasets observe 5% of the CUB test data during pre-training [49] as a result of which their performance is overestimated. This makes the strong performance of our weakly-supervised models (which do

[2] Although our system-level evaluations hamper exact comparisons, our results suggest that the weakly supervised IG-3.6B dataset provides the same amount of supervisory signal as the supervised JFT-300M dataset.

Model	Pre-training	Resolution		IN-1k Accuracy		Classification accuracy				Throughput	FLOPs	Params
		Pre.	Fine.	Report.	Reprod.	IN-5k	iNat.	Places	CUB	(images/sec.)	(B)	(M)
Supervised pre-training[†]												
EfficientNet L2 [73]	JFT 300M[‡]	475	800	*88.4*	88.3	–	–	–	–	108	479.9	480.3
EfficientNet L2 [73]	JFT 300M[‡]	475	–	*88.2*	88.0	**61.8**	**86.5**	59.4	91.2[§]	293	172.6	480.3
EfficientNet B7 [73]	JFT 300M[‡]	600	–	*86.9*	86.7	56.7	82.0	59.2	90.6[§]	652	38.4	66.3
EfficientNet B6 [73]	JFT 300M[‡]	528	–	*86.4*	86.3	55.4	79.9	58.8	89.1[§]	849	19.5	43.0
EfficientNet B8 [72]	IN-1k	672	–	*85.5*	85.2	54.8	81.3	58.6	89.3[§]	480	63.7	87.4
EfficientNet B7 [72]	IN-1k	600	–	*85.2*	85.0	54.4	80.6	58.7	88.9[§]	652	38.4	66.3
EfficientNet B6 [72]	IN-1k	528	–	*84.8*	84.7	53.6	79.1	58.5	88.5[§]	849	19.5	43.0
ViT G/14 [76]	JFT 3B	224	518	***90.5***	–	–	–	–	–	56	2826.1	1846.3
ViT L/16 [76]	JFT 3B	224	384	*88.5*	–	–	–	–	–	567	191.5	304.7
ViT H/14 [20]	JFT 300M	224	518	<u>88.6</u>	–	–	–	–	–	116	1018.8	633.5
ViT L/16 [20]	JFT 300M	224	512	*87.8*	–	–	–	–	–	255	362.9	305.2
ViT L/16 [20]	IN-21k	224	384	*85.2*	85.2	–	81.7	59.0	91.3[§]	567	191.5	304.7
ViT B/16 [20]	IN-21k	224	384	*84.0*	84.2	–	79.8	58.2	90.8[§]	1,161	55.6	86.9
ViT L/32 [20]	IN-21k	224	384	*81.3*	81.5	–	74.6	57.7	88.7[§]	1,439	54.4	306.6
Weakly supervised pre-training												
ViT H/14	IG 3.6B	224	518	<u>88.6</u>		<u>60.9</u>	<u>86.0</u>	**60.7**	**91.7**	116	1018.8	633.5
ViT L/16	IG 3.6B	224	512	88.1		59.0	84.2	**60.7**	<u>91.6</u>	255	362.9	305.2
ViT B/16	IG 3.6B	224	384	85.3		54.5	79.9	59.1	89.8	1,161	55.6	86.9
RegNetY 128GF	IG 3.6B	224	384	88.2		<u>60.9</u>	85.7	60.1	90.8	307	375.2	644.8
RegNetY 32GF	IG 3.6B	224	384	86.8		58.5	82.9	59.6	89.5	976	95.1	145.0
RegNetY 16GF	IG 3.6B	224	384	86.0		57.2	81.4	59.2	88.3	1,401	47.0	83.6

Table 1. Transfer-learning accuracy of models pre-trained on the specified pre-training dataset followed by finetuning and testing on five transfer datasets. Accuracies that were adopted from the original papers are *italicized*. The best result on each dataset is **boldfaced**; the second-best result is <u>underlined</u>. Our weakly-supervised pre-trained models achieve the best or second-best performance on all five transfer datasets. [†]It is unknown how much manual curation was performed to annotate the JFT datasets. [‡]IN-1k is used as supervised pre-training data; JFT 300M is used without labels. [§]Model was pre-trained on IN-1k training set, which overlaps with the CUB-2011 test set.

not see test data during training) particularly noteworthy.

To provide more insight into the classification accuracy and throughput trade-off, we plot one as a function of the other in Figure 2. Comparing ViT and RegNetY models trained on the same IG-3.6B dataset, we observe that vision transformers obtain the highest classification accuracies. In terms of accuracy-throughput tradeoff, RegNetYs outperform at small to medium model sizes. The RegNetY 128GF model performs quite similarly on accuracy and throughput to the semi-supervised EfficientNet L2 model, but at smaller size scales, RegNetYs provide a better tradeoff.

4.3. Comparison with Self-Supervised Pre-Training

Our experiments so far suggest that the ability to scale up weakly-supervised pretraining to billions of images can offset the lower amount of learning signal obtained per training example. This raises the question if we need weak supervision at all, or whether modern *self-supervised* learners [10–14, 26, 27, 29, 31, 51] may suffice. Self-supervised learning scales even more easily than weakly-supervised learning, and prior work has demonstrated the potential of self-supervised pre-training at scale [27, 31].

We perform transfer-learning experiments on ImageNet-1k that compare our weakly-supervised learner with SimCLR v2 [13], SEER [27], and BEiT [3]. The comparison with SEER is of particular interest: because it is trained on a similar collection[3] of Instagram images, we can readily compare both learning paradigms on the same data distribution. We perform experiments in two transfer-learning settings: (1) a setting in which a linear classifier is attached on top of the pre-trained model and the resulting full model is finetuned and (2) a setting that initializes this linear classifier using the zero-shot transfer approach described in Section 4.4 (without Platt scaling) before finetuning the full model. Following prior work [13, 27], we vary the amount of labeled ImageNet examples used for finetuning to 1%, 10%, and 100% of the original ImageNet-1k training set. We report results using images of size 224×224 pixels.

The results of our experiments are presented in Table 2. Results for SimCLRv2, SEER and BEiT were adopted from [3, 13, 27]; small differences in experimental setup may exist. Our results show that weakly-supervised learning substantially outperforms current self-supervised learners, particularly in low-shot transfer settings, likely because our weakly-supervised learners receive more learning signal per sample. Moreover, our results show that weakly-supervised learners benefit from zero-shot initialization in low-shot transfer settings. We note that our observations may change if self-supervised learners are scaled further.

4.4. Zero-Shot Transfer

Another potential advantage of weakly-supervised models is that they have observed a large variety of training targets during pre-training. This may help them recognize new

[3]The data distribution used in [27] and in our study may not be exactly the same, as we use the data resampling approach described in Section 3.1.

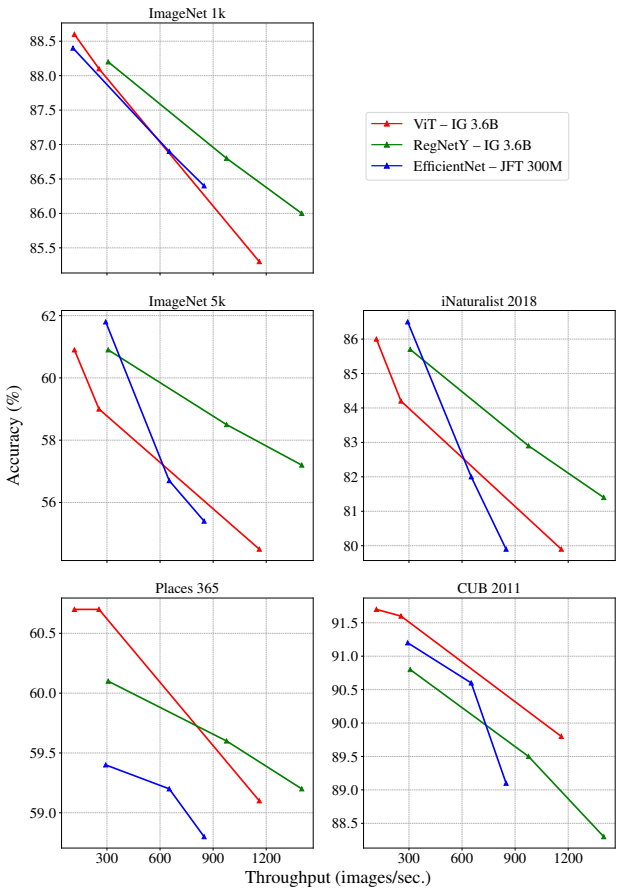

Figure 2. Transfer-learning accuracy as a function of throughput of pre-trained models that were finetuned on five datasets (please refer to Table 1 for full results). ViTs and EfficientNets achieve the highest top-line accuracies, but RegNetY models perform better in the high-throughput regime.

visual concepts quickly. We test the ability of our models to learn and recognize new visual concepts rapidly in *zero-shot transfer* learning setting.[4] In this setting, we use the output layer of the pre-trained model directly without any finetuning. We can do this because we trained on 27k hashtags derived from WordNet [22], allowing us to define a mapping between hashtags and class labels for datasets, like ImageNet-1k, also built on WordNet. We use the same image resolution as pre-training, *viz.*, 224×224 pixels.

Platt scaling. In our zero-shot transfer experiments, we consider a transductive learning setting [24] in which all test examples are available simultaneously at test time. This allows us to train a Platt scaler [53] on the test data that corrects for differences in the distribution of hashtags (which are Zipfian) and the distribution of classes in the target task

[4]Some prior work refers to this learning setting as zero-shot learning [37, 57]. We find this term confusing because it differs from classical zero-shot learning [43]. Hence, we adopt the term *zero-shot transfer*.

Model	Approach	Pre-training	Transfer	Accuracy 1%	10%	100%
Self-supervised pre-training						
RN152w3 + SK	SimCLRv2† [13]	IN-1k	Finetune	*74.9*	*80.1*	*83.1*
RegNetY 128GF	SEER [27]	IG 1B	Finetune	*57.5*	*76.7*	*83.8*
RegNetY 256GF	SEER [27]	IG 1B	Finetune	*60.5*	*77.9*	*84.2*
ViT L/16	BEiT [3]	IN-1k	Finetune	–	–	*85.2*
Weakly supervised pre-training						
RegNetY 128GF	Ours	IG 3.6B	ZS-Init.+Ft.	**82.0**	**84.5**	87.8
RegNetY 32GF	Ours	IG 3.6B	ZS-Init.+Ft.	79.4	82.0	86.5
RegNetY 16GF	Ours	IG 3.6B	ZS-Init.+Ft.	77.6	80.8	85.7
RegNetY 128GF	Ours	IG 3.6B	Finetune	79.2	84.1	**87.9**
RegNetY 32GF	Ours	IG 3.6B	Finetune	74.8	81.7	86.3
RegNetY 16GF	Ours	IG 3.6B	Finetune	72.3	80.4	85.3

Table 2. Transfer accuracy of models on the ImageNet-1k dataset as a function of the percentage of ImageNet-1k training examples used for transfer learning. Transfer learning is performed using either standard finetuning, or zero-shot (ZS) transfer initialization followed by finetuning. The best result in each setting is **boldfaced**; the second-best result is underlined. Accuracies that are adopted from the original paper are *emphasized*. Our weakly supervised pre-trained models outperform models pre-trained with modern self-supervised learners, in particular, in the few-shot regime. †During finetuning, SimCLRv2 accessed 100% of the ImageNet training images but $k\%$ of the labels, whereas SEER and our method accessed $k\%$ of the training data.

(which is uniform). The Platt scaler is parameterized by a weight vector $\mathbf{w} \in \mathbb{R}^C$ and bias vector $\mathbf{b} \in \mathbb{R}^C$, where C is the number of classes. Given a probability vector $\mathbf{p} \in \Delta_C$ with Δ_C the C-simplex, the Platt scaler computes a new output $\mathbf{p}' = \text{softmax}\left(\text{diag}(\mathbf{w})\mathbf{p} + \mathbf{b}\right)$. The Platt scaler is trained to minimize the cross-entropy loss between the test distribution of $\mathbf{p}'$ and a uniform distribution over the C classes. Note that this does not use the test labels; it only encourages the predictions to be uniform over classes.

Mapping from hashtags to ImageNet classes. Because the targets in both the ImageNet and IG-3.6B datasets are English nouns, we can construct a many-to-many mapping between Instagram hashtags and ImageNet classes. To do so, we first map both hashtags and ImageNet classes to WordNet synsets, and then map hashtags to ImageNet classes based on their similarity in WordNet [22]. We use the resulting many-to-many mapping between hashtags and classes to aggregate hashtag-prediction scores over ImageNet classes. We experiment with three different aggregation methods and use the method that we found to work best for each model; see appendix for details.

Results. The results of our zero-transfer results are presented in Table 3. The table presents top-1 classification accuracies on four ImageNet-like test sets for our models with and without Platt scaling. We compare the performance of our models with that of CLIP [57] and ALIGN [37]. These experiments are *system-level* comparisons in which many factors are different: For example, CLIP was trained on

a dataset of 400 million images and captions that appears more curated than ours, it was finetuned at a higher resolution, and it performs zero-shot transfer via prompt engineering [9] which is known to improve recognition accuracy [57]. ALIGN uses a different image-recognition model (*viz.*, EfficientNet) and was trained on 1 billion pairs of web images and corresponding alt-texts [37].

Table 3 presents our results with zero-shot transfer on four ImageNet-like datasets. The results show that our weakly supervised models perform very well out-of-the-box: without ever seeing an ImageNet image, our best model achieves an ImageNet top-1 accuracy of 75.3%. The results also show that Platt scaling is essential to obtain good zero-shot transfer performance with our model, as it corrects for differences in the distribution of hashtags and ImageNet classes. Finally, we find that our ViT models underperform our RegNetY models in the zero-shot transfer setting. This is unsurprising considering that ViTs also underperformed RegNetYs on ImageNet-1k finetuning at an image resolution of 224×224 pixels.

Comparing our models with CLIP [57], we observe that the CLIP ViT L/14 model slightly outperforms our model in zero-shot transfer to the IN-1k dataset; whereas the smaller RN50×64 CLIP model underperforms it. On some datasets, the ALIGN [37] model performs even slightly better. However, the results are not fully consistent: our models do obtain the best performance on the ImageNet-v2 dataset [60]. Because these experiments perform system-level comparisons, it is difficult to articulate what drives these differences in performance. Nonetheless, our results provide further evidence that weakly-supervised approaches like ours, CLIP, and ALIGN provide a promising path towards the development of open-world visual-recognition models [33].

Model	Platt	Classification accuracy			
		IN-1k	ReaL-IN	IN-v2	Obj. Net
Visual n-grams [45]	N/A	*35.2*	–	–	–
CLIP RN50×64 [57]	N/A	*73.6*	–	–	–
CLIP ViT L/14 [57]	N/A	*76.2*	–	*70.1*	***72.3***
ALIGN [37]	N/A	***76.4***	–	*70.1*	–
RegNetY 128GF	Yes	75.3	**79.5**	**71.1**	64.3
RegNetY 32GF	Yes	73.6	78.3	69.1	49.9
RegNetY 16GF	Yes	72.5	77.6	67.9	45.1
RegNetY 128GF	No	65.1	69.7	60.2	54.2
RegNetY 32GF	No	62.2	67.5	57.3	59.1
RegNetY 16GF	No	60.7	66.3	55.6	54.8
ViT H/14	Yes	72.3	76.5	66.5	60.0
ViT L/16	Yes	71.6	76.0	65.7	57.3
ViT B/16	Yes	67.7	73.0	61.9	43.0
ViT H/14	No	62.8	67.3	57.7	52.4
ViT L/16	No	62.1	66.6	56.3	51.1
ViT B/16	No	58.4	63.6	52.3	48.9

Table 3. Zero-shot transfer accuracy of models on four datasets with WordNet-based classes: (1) the ImageNet-1k dataset, (2) the ReaL ImageNet [7] dataset, (3) the ImageNet v2 [60] dataset, and (4) the ObjectNet [4] dataset. The best result on each dataset is **boldfaced**; the second-best result is underlined. Accuracies that are adopted from the original paper are *italicized*. When using Platt scaling, our weakly-supervised RegNetY models work very well out-of-the-box. They achieve 75.3% zero-shot transfer accuracy on ImageNet-1k, and outperform CLIP [57] and ALIGN [37] on the ImageNet v2 [60] dataset.

5. Broader Impact

A potential downside of weakly-supervised training of models on uncurated web data is that they may learn harmful associations that reflect offensive stereotypes [6, 9]. Moreover, the models may not work equally well for different user groups; for example, they do not work as well in non-English speaking countries [17] because we used English hashtags as the basis for training our models. We performed a series of experiments to better understand: (1) the associations our hashtag-prediction models learn with photos of people with varying characteristics, and (2) how well those models perform on photos taken in non-English speaking countries. We summarize the results of those experiments here and refer to the appendix for further details.

Analyzing associations in hashtag predictions. We performed experiments analyzing the associations our RegNetY 128GF hashtag-prediction models make for photos that contain people with different apparent skin tone, apparent age, apparent gender, and apparent race. The experiments were performed using: (1) a proprietary dataset that contains $178{,}448$ Instagram photos that were annotated using the Fitzpatrick skin tone scale [23] and (2) the UTK Faces dataset, which provides apparent age, apparent gender, and apparent race labels [78].

We find that the model has learned several associations between hashtags and skin tone; see the appendix for details. For example, `#redhead` is more commonly predicted for photos of people with a light skin tone, whereas `#black` is more often predicted for people with a dark skin tone. Similarly, some hashtag predictions correlate with the apparent age of people in photos; see the appendix for details. For example, our models more commonly predict `#baby` or `#kid` for photos that contain people who are $1-10$ years old, and more commonly predict `#elder` for the $80-90$ years age group. When analyzing our model for gender stereotypes, we found that our model's hashtag predictions associate men with `#football` and `#basketball` more frequently. By contrast, our model associates photos containing women more frequently with `#makeup` and `#bikini`; see the appendix for details.

The most troubling associations we observed stem from an analysis of model predictions for photos that contain people with different apparent race. In particular, some of our experiments suggest that our model may associate photos that contain Black people with `#mugshot` and `#prison` more frequently; see the appendix. However, it is unclear whether these observations are due to our model making

incorrect or biased predictions for photos in the evaluation dataset, or whether they are due to the evaluation dataset containing a problematically biased image distribution. In particular, a more detailed analysis uncovered the presence of a troubling bias in the evaluation dataset (rather than in our model): we found that the UTK Faces dataset [78] contains a substantial number of mug shots that disproportionally portray Black individuals.

Overall, our results suggest that while our hashtag-prediction models appear to make fewer troubling predictions than language models [6, 9], careful analyses and adaptations would be needed before hashtag predictions from our model can be used in real-world scenarios. Motivated by this observation, we do not release the final hashtag-prediction layer of our models as part of this study.

Analyzing hashtag prediction fairness. We also analyzed how well our hashtag-prediction models work on photos taken across the world. We repeated the analysis of [17] on the Dollar Street dataset and performed analyses on a proprietary dataset that contains millions of images with known country of origin. Akin to [17], we observe large accuracy differences of our model on Dollar Street photos from different countries. Our analysis on the much larger and more carefully collected proprietary dataset confirms this result but suggests that the effect sizes are much smaller than reported in [17]; see the appendix for details. Specifically, we find that the range of per-country accuracies is in a relatively tight range of $\sim 5\%$ *i.e.*, our model achieves per-country recognition accuracies between 65% and 70% for all 15 countries in the dataset. Overall, our results suggest more work is needed to train models that perform equally across the world. In future work, we plan to train multilingual hashtag models [64] as this may lead to models that achieve equal recognition accuracies across countries.

6. Discussion

In this paper, we have presented an in-depth study of fully supervised, self-supervised, and weakly-supervised pre-training for image recognition. Combined with related work [25, 37, 49, 56, 57], our results provide a compelling argument for the use of weakly-supervised pre-training in the development of systems for visual perception. However, our study also uncovers limitations of this line of research.

In particular, we find it is increasingly difficult to perform systematic, controlled experiments comparing different approaches and techniques. There are a variety of reasons for this, including the use of proprietary data that was collected via opaque processes[5], the diversity of model architectures used, the complexity of training recipes, the heterogeneity of hardware and software platforms used, the vast compute resources required, and the fact that not all studies publish pre-trained models. Together, this creates an environment in which researchers cannot perform controlled studies that test the effect of one variable, keeping all other variables fixed. Instead, they can only perform *system-level* comparisons, as we did in this study. Such comparisons provide signal on the potential of various approaches, but they do not produce conclusive results. This problem is exacerbated by the fact that the signal we are measuring is small, as recognition accuracies on commonly used evaluation datasets appear saturated. To create a thriving research community focused on large-scale learning of vision systems, it is imperative that we address these issues.

A second limitation of this line of research is the strong focus on recognition accuracy and inference speed as the main measures of merit. While recognition accuracy and inference speed are obviously important, they are not the only measures that matter for the quality of a visual-perception system. Other measures include the recognition accuracy experienced by different groups of users and the prevalence of predictions that reinforce harmful stereotypes. We presented an initial study of such measures in Section 5 but this foray is not completely conclusive or sufficient. In particular, we found there are no well-established evaluation datasets and experimental protocols that facilitate the rigorous analyses. To make matters worse, the presence of harmful stereotypes in some commonly used vision datasets (such as the association between Black people and mug shots we found in the UTK Faces dataset [78]) appears to be unknown. In order to make hashtag-prediction systems like ours ready for real-world deployment, it is essential that we improve the quality of our analyses, and that we address any issues that those analyses may surface.

To conclude, we emphasize that we remain convinced about the potential of weakly-supervised learning approaches. If we resolve the aforementioned issues, we believe such approaches may improve visual-perception systems in the same way that large-scale language models have improved natural language understanding, machine translation, and speech recognition.

Acknowledgements

We would like to thank Ishan Misra, Priya Goyal, Benjamin Lefaudeux, Min Xu and Vinayak Tantia for discussions and feedback, and Haowei Lu and Yingxin Kang for help with the data loader implementation. We thank Deepti Ghadiyaram, Anmol Kalia, and Katayoun Zand for their work on the internal datasets with apparent skin tone and country annotations. We thank Phoebe Helander, Adina Williams, Maximilian Nickel and Emily Dinan for helpful feedback on the Broader Impact analysis. Lastly, we thank Brian O'Horo for support with the training infrastructure.

[5] We acknowledge that, although the data we use in our experiments is public, it is hard for others to collect that data. However, unlike other studies, we did strive to be comprehensive in describing our data-collection procedure, as we aim to maximize what the reader can learn from our study.

References

[1] Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. Exploring the limits of large scale pre-training. In *arXiv:2110.02095*, 2021. 2

[2] A. Adcock, V. Reis, M. Singh, Z. Yan, L. van der Maaten, K. Zhang, S. Motwani, J. Guerin, N. Goyal, I. Misra, L. Gustafson, C. Changhan, and P. Goyal. Classy vision. https://github.com/facebookresearch/ClassyVision, 2019. 14

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 5, 6

[4] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 7, 15

[5] Josh Beal, Hao-Yu Wu, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. Billion-scale pretraining with vision transformers for multi-task visual representations. In *arXiv:2108.05887*, 2021. 2

[6] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021. 1, 7, 8

[7] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aaron van den Oord. Are we done with ImageNet?, 2020. 7, 15

[8] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. In *arXiv:2110.01963*, 2021. 3

[9] Tom B. Brown, Ben Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen M. Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing*, 2020. 1, 7, 8

[10] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *arXiv:1807.05520*, 2018. 2, 5

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 1, 2, 5

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *arXiv:2104.14294*, 2021. 1, 2, 5

[13] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 1, 2, 5, 6

[14] X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 5

[15] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. In *arXiv:1805.09501*, 2018. 4

[16] E. Denton, J. Weston, M. Paluri, L. Bourdev, and R. Fergus. User conditional hashtag prediction for images. In *Proc. KDD*, pages 1731–1740, 2015. 2

[17] T. DeVries, I. Misra, C. Wang, and L.J.P. van der Maaten. Does object recognition work for everyone? In *CVPR Workshop on Computer Vision for Global Challenges*, 2019. 7, 8, 15

[18] Piotr Dollár, Mannat Singh, and Ross Girshick. Fast and accurate model scaling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 924–932, 2021. 3, 13

[19] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, pages 647–655, 2014. 2

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 3, 4, 5, 13, 15

[21] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *arXiv:2104.11227*, 2021. 2

[22] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998. 3, 6, 12, 14

[23] T. B. Fitzpatrick. Soleil et peau. *Journal de Médecine Esthétique*, 2:33–34, 1975. 7, 16

[24] A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *UAI*, pages 148–155, 1998. 6

[25] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*, pages 12046–12055, 2019. 2, 8

[26] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *arXiv:1803.07728*, 2018. 2, 5

[27] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, and Piotr Bojanowski. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021. 1, 2, 5, 6

[28] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 3

[29] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel 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 *arXiv:2006.07733*, 2020. 1, 2, 5

[30] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, Changhu Wang, and Alan Yuille. Transfg: A transformer architecture for fine-grained recognition. In *arXiv:2103.07976*, 2021. 2

[31] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2019. 1, 2, 5

[32] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 2

[33] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. In *arXiv:1709.01450*, 2017. 7

[34] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 13

[35] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 3, 13

[36] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. Deep networks with stochastic depth. In *arXiv:1603.09382*, 2016. 4

[37] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *arXiv:2102.05918*, 2021. 1, 2, 6, 7, 8

[38] A. Joulin, L.J.P. van der Maaten, A. Jabri, and N. Vasilache. Learning visual features from large weakly supervised data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–84, 2016. 1, 2

[39] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. In *arXiv:1705.06950*, 2017. 2

[40] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. In *arXiv:1704.03162*, 2017. 2

[41] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6(2):8, 2019. 4

[42] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *arXiv:1805.08974*, 2018. 1, 2

[43] C.H. Lampert. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009. 6

[44] A. Li, A. Jabri, A. Joulin, and L.J.P. van der Maaten. Learning visual n-grams from web data. In *International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[45] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4183–4192, 2017. 7

[46] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 2

[47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3, 13

[48] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 3, 4

[49] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018. 1, 2, 3, 4, 8, 12, 13, 14

[50] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. 2

[51] I. Misra and L.J.P. van der Maaten. Self-supervised learning of pretext-invariant representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 5

[52] R. Müller, S. Kornblith, and G.E. Hinton. When does label smoothing help? In *NeurIPS*, 2019. 3

[53] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999. 6

[54] B.T. Polyak and A.B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. 3, 4

[55] V.U. Prabhu and A. Birhane. Large datasets: A Pyrrhic win for computer vision? In *arXiv:2006.16923*, 2020. 3

[56] Filip Radenovic, Animesh Sinha, Albert Gordo, Tamara Berg, and Dhruv Mahajan. Large-scale attribute-object compositions. In *arXiv:2105.11373*, 2021. 2, 8

[57] 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 *arXiv:2103.00020*, 2021. 1, 2, 6, 7, 8

[58] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020. 3, 13, 14

[59] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 512–519, 2014. 2

[60] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 2, 7, 15

[61] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 2

[62] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *arXiv:2104.10972*, 2021. 2

[63] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2, 4

[64] Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, Joao Carreira, Phil Blunsom, and Andrew Zisserman. Visual grounding in video for unsupervised word translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[65] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 3, 13

[66] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *arXiv preprint arXiv:1906.06423*, 2019. 4

[67] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4

[68] A. Veit, M. Nickel, S. Belongie, and L.J.P. van der Maaten. Separating self-expression and visual content in hashtag supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5919–5927, 2018. 2

[69] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 4

[70] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *arXiv 2012.00759*, 2020. 2

[71] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 14

[72] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020. 4, 5, 15

[73] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 4, 5, 13, 15

[74] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 3, 4, 13

[75] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of AISTATS*, 2017. 16

[76] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021. 1, 2, 4, 5, 13, 14, 15

[77] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3, 4

[78] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 7, 8, 17

[79] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 4