Distributionally Robust Classification on a Data Budget

Anonymous authors
Paper under double-blind review

Abstract

Real world uses of deep learning require predictable model behavior under distribution shifts. Models such as CLIP show emergent natural distributional robustness comparable to humans, but may require hundreds of millions of training samples. Can we train robust learners in a domain where data is limited? To rigorously address this question, we introduce JANuS (Joint Annotations and Names Set), a collection of four new training datasets with images, labels, and corresponding captions, and perform a series of careful controlled investigations of factors contributing to robustness in image classification. Using JANuS as a testbed, we show that standard ResNet-50 trained with the cross-entropy loss on 2.4 Mn image samples can attain comparable robustness to a CLIP ResNet-50 trained on 400 Mn samples. To our knowledge, this is the first result showing near state-of-the-art distributional robustness on a very limited data budget.

1 Introduction

Motivation. A *natural distribution shift* is defined as evaluation data which differs from the data on which a model was trained due to natural factors. Real world uses of deep image classifiers require predictable model behavior under such shifts. Unfortunately, the majority of standard deep computer vision models for image classification perform significantly worse under natural shifts Hendrycks & Dietterich (2019); Miller et al. (2021), in contrast with human vision (Recht et al., 2019).

Vision-Language (VL) models such as CLIP, introduced in Radford et al. (2021), showed emergent natural distributional robustness comparable to humans across a wide range of shifts of ImageNet, at the cost of base accuracy. Jia et al. (2021) showed CLIP-like models can be carefully fine-tuned to be robust as well as achieve high base accuracy. However, VL models require massive amounts of data for training; in some cases, orders of magnitude higher than standard supervised models (Pham et al., 2021). These results raise challenging questions: are data scaling laws at work for robust computer vision, similar to those discovered in NLP? Does robustness only emerge when models are trained on massive datasets? And is vision-language pre-training necessary for robustness?

Radford et al. (2021) argue that VL pre-training in CLIP offers unique advantages when compared to conventional large-data model training techniques. By contrast, Fang et al. (2022) and Nguyen et al. (2022) argue that VL robustness is a consequence of the training data diversity and quantity, with vision-language pretraining playing little role.

In most real-world applications, data is limited, and unlikely to be accompanied by informative natural language captions. For example, the PCam medical imaging dataset from Veeling et al. (2019) has only $\sim 320,000$ images. Nevertheless, distributional robustness is of paramount importance in the setting.

What can be done to train robust models in data-limited settings, without access to informative captions? Can we take advantage of other attributes of model training which have largely been disregarded in the literature, such as architecture, model size, and image resolution?

Our Contributions. To clearly delineate the potential sources of emergent distributional robustness, we evaluate a vast suite of existing models trained on diverse data budgets. We also supplement with dozens of new models trained from scratch. Our key contributions are as follows:

1. For the first time, we show that it *is* possible to train highly robust and accurate models, using conventional cross-entropy (CE) loss, even when both data and model size are limited (See Fig. 1).

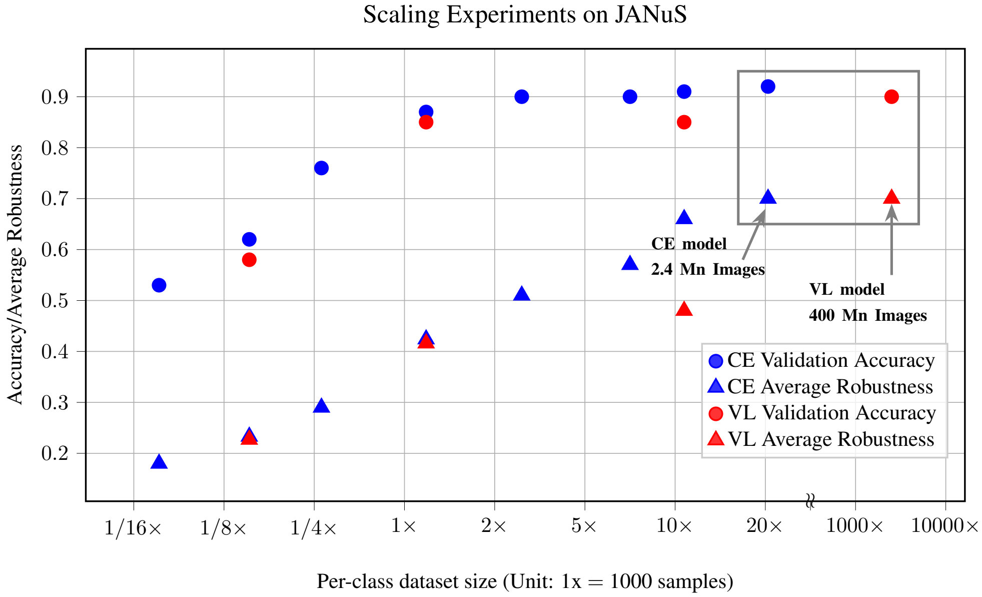

Figure 1: ***Under a data budget, standard CE-loss models outperform VL-loss models in both accuracy and robustness.*** *We train ResNet-50 models using both CE-loss and VL-loss across a wide range of data scales, and find that accuracy of VL-loss and CE-loss models is extremely similar at small scales. For scaling 4X and above, CE-loss models exhibit superior robustness; the CE-loss model trained on just 2.4 Mn JANuS samples has comparable robustness, as well as comparable accuracy, to the CLIP ResNet-50 trained on 400 Mn samples. (See Tab. 2 for information on the JANuS dataset, which we create and use to train these models.) Image best viewed in color.*

2. We introduce JANuS (Joint Annotations and Names Set), a new class-balanced dataset with images, labels, and captions. To our knowledge, this is the first such dataset of its kind. (See Tab. 2).
3. We conduct the largest (to date) robustness analysis of image classification models (numbering over 650), including many recent architectures, and show that even with relatively modest model and data scaling (compared to Brown et al. (2020), Radford et al. (2021)), one can train robust models on both large and fine-grained label sets. (See Fig. 2 and Fig. 3)
4. We outline useful heuristics to improve distributional robustness when data budgets are limited. (See Sec. 7).
5. In order to enable future research and reproducibility, we release our code, our dataset, and a complete description of our results (see supplemental attachments).

2 Related Work

Our paper follows a series of recent works studying robustness under distribution shift in the context of image classification (Recht et al., 2019; Taori et al., 2020; Miller et al., 2021; Fang et al., 2022; Nguyen et al., 2022). This line of inquiry into distributional robustness focused on the linear fit between in-distribution and out-of-distribution accuracy found between common image classification datasets (such as ImageNet) and their distribution shifts. In contrast to most of these earlier papers, our analysis takes place in a realistic setting where models are trained on a wide range of datasets. Therefore, following the results in Nguyen et al. (2022), we do not use linear fit measures in our analysis, instead relying on average out-of-distribution robustness.

Jia et al. (2021); Pham et al. (2021) showed that human-level distributional robustness is possible even as base accuracy approaches state-of-the-art, as long as sufficient data is available for training. The gains are not limited to CLIP; other VL-loss (vision-language loss) functions also achieve strong distributional robustness (Yu et al., 2022; Wang et al., 2022b). We discuss some of these alternate approaches in Appendix Sec. A, while noting that none of these exhibit superior robustness than CLIP.

The internals of CLIP differ from those of typical models in several important ways: the choice of loss function, the training dataset, and the use of natural language captions as labels. However, it is still an open question as to which of these differences lead to CLIP's extraordinary robustness. Recent works have addressed this question, and have reached various interesting conclusions. Fang et al. (2022) argue that intrinsic diversity of training image data is the main source of the distributional robustness gains of VL models in the zero-shot setting, with factors such as language supervision contributing little to no distributional robustness. However, in a different (transfer learning) setting, Santurkar et al. (2022) argue that, given a sufficiently large pretraining dataset and descriptive, low variability captions, contrastively trained VL models are more robust than self-supervised image-only models trained with the SIMCLR-loss. We conduct controlled comparisons between vision-language classifiers and conventional classifiers, and find that when controlling for data quantity and diversity, high accuracy VL-loss models are actually *less* robust than high accuracy CE-loss models.

Nguyen et al. (2022) is an important precursor to our work. Their extensive experiments on vision-language models in the low accuracy regime showed that controlling for the pretraining dataset was essential for understanding distributional robustness. We extend this understanding, and show that model architecture, size, image resolution, and even the label set selected for the classification problem can all have substantial effects on robustness. Finally, unlike Nguyen et al. (2022), all our results are shown in both low and high accuracy regimes, and across different test sets.

In their paper investigating the role of language on robustness, Fang et al. (2022) introduced ImageNet-Captions, which added Flickr-captions to nearly 450,000 ImageNet images. We extend this work by introducing JANuS, which add over 50,000 new human-supervised samples to 100 classes in ImageNet-Captions in order to rebalance the classes, as it has been shown that CE-loss models often struggle with imbalanced classes (Phan & Yamamoto, 2020).

3 Preliminaries

Training Datasets and Distribution Shifts. Our principal tool for measuring robustness in this paper is model accuracy on natural distribution shifts. We focus on the ImageNet dataset which has extensively been studied in the literature on distributional robustness. Following Radford et al. (2021), we focus on the following four distribution shifts: Imagenet-Sketch (IN*-s), Imagenet-R (IN*-r), Imagenet-A (IN*-a), and Imagenet-V2 (IN*-v2) for our evaluation. Additional details on our pretraining datasets and distribution shifts are in Appendix Sec. B and Sec. C.

Definition of data budgets. Traditionally, the problem of image classification has been conceptually divided between pre-train and fine-tune approaches and fully supervised approaches. Both approaches require large amounts of labeled data; this motivates our focus on data budgets. What data budget is reasonable is, of course, dependent on the problem. Since our analysis focuses on ImageNet, we define our data budget in multiples of the approximate per-class size of the original ImageNet dataset. For example, for a 100-class label set, a budget of 1 million samples would be a 10x data budget. Throughout this paper, we provide scaling experiments which cover a wide range of data budgets.

Metrics for distributional robustness. Our primary metric is *average robustness* (abbv: Avg. Rob.), which is the average test-set accuracy of a model on all distribution shifts; in our case, four. Although this measure is easy to interpret, it can conceal substantial performance differences between shifts; therefore, we also include *shift-specific accuracy* in Appendix Sec. B.

Another metric we reference is *effective robustness*, introduced by Taori et al. (2020), primarily to situate our work within the existing literature. This metric is defined in Taori et al. (2020) as a graphical tool to describe how robust a model is on natural distribution shifts. Humans have been shown to be perfectly robust; therefore, a graph of base-versus-shift test accuracy follows the $y = x$ trendline; for neural networks this trend-line typically is parallel to and is generally below $y = x$.

Finally, we include *Effective Robustness Ratio* (abbv: E.R.R.), from Feuer et al. (2022) in our appendix tables. This is defined as the ratio of average robustness over base task accuracy. We find that this is an effective measure when we limit our comparisons to models with roughly similar base accuracy.

Glossary. For ease of understanding, we provide a glossary of common terms and abbreviations.

Loss functions. We examine models trained with two types of losses. *VL-loss* refers to the InfoNCE loss used by CLIP (Radford et al., 2021). *CE-loss* is the typical cross-entropy loss used to train the vast majority of models for image classification.

Label types. CE-loss models use *integer* labels (referring to discretely labelled classes), and VL-loss models use *caption* labels. We refer to human-annotated labels (whenever available) as *ground-truth*. We refer to labels generated by automated processes as either *synthetic* or *subset-matched* (defined below in Sec. 5).

Data filtration. We define *data filtration* as any strategy which sub-selects image-caption pairs.

4 Experimental Design

The models used in our robustness evaluation experiments are drawn from three sources; the pytorch-image-models (timm) Wightman (2019) repository, the open-clip repository Ilharco et al. (2021), and models trained by us. We provide details about all studied models in the Appendix; see Sec. E. We do not consider models which are not publicly available.

We divide our study into one *1000-class* evaluation and two *100-class* evaluations. The 1000-class evaluation is ImageNet-1k (IN1000-Val, IN1000-V2, etc), and the 100-class evaluations are both subsets of ImageNet.

We evaluate on both 1000-class and 100-class label set sizes because we found dramatic differences in model performance depending on the task; models trained on many-class problems become more accurate and robust when the label set size is reduced to a subset of those classes at inference time, and the improvements are not necessarily proportionate (Tab. 1.). Hence, considering both label set sizes offers a more complete picture of model robustness.

Model Name	Average Validation Accuracy	Average Robust Accuracy
CLIP-RN50 (1000-class)	0.5985	0.4306
CLIP-RN50 (Avg. 100-class)	0.8517	0.7182
SWSL-RN50 (1000-class)	0.8362	0.6857
SWSL-RN50 (Avg. 100-class)	.9524	0.7612

Table 1: ***Zero-shot model robustness is affected by the difficulty of the task.*** *Both quantity and quality of labels alters model accuracy and robustness under shift; it also changes the comparative performance of VL-loss and CE-loss models. In this table, we transform the 1000-class IN1000 label set into ten 100-class label sets, and find that the resulting predictions are far more accurate and robust, particularly those of the VL-loss model. This finding motivates our choice to study model robustness on multiple label set sizes.*

Within the 100-class label set size, we consider a broad-scope classification problem (IN100-Val, IN100-V2, etc), as well as a fine-grained classification problem (IN100-Dogs, IN100-Dogs-V2, etc), a 100-class subset of ImageNet which consists entirely of dog breeds.

We report IN100-Val results separately for pretrained models with large amounts of data and our own models trained on various data budgets. The exact class indices for each shift can be found in the Appendix (Sec. I).

5 JANuS: A Benchmark Dataset for Robust Model Training

Challenges of robust model training. Training robust models from scratch on IN1000 presents resource challenges for researchers, particularly vision-language models. Prior works such as Fang et al. (2022) have cited low-accuracy results and relied on the linear fit hypothesis to project those results to high accuracy regimes. However, as observed in Nguyen et al. (2022), researchers cannot rely on trends observed in low-accuracy regimes to persist in high-accuracy regimes unless the training dataset and loss function are fixed in advance. Following from our observations about the effect of label set size on model performance, we postulate that a 100-class, broad scope problem is ideal for comparative studies of robustness in high-accuracy regimes. However, no training dataset exists which is designed for 100-class ImageNet problems and is sufficiently diverse to train high accuracy, high-robustness models with both VL-loss and CE-loss objectives.

To resolve this challenge, we introduce JANuS (Joint Annotations and Names Set), a collection of four new training datasets with images, labels and corresponding captions. Each dataset in JANuS builds upon an existing dataset by

Dataset	G.T. Label	Machine Label	Caption Source	Supervised	Filtered	Balanced
ImageNet-100 (IN100)	✓	✓	Flickr, BLIP	Yes	Human	Yes
OpenImages-100 (OI100)	✓	✓	Flickr, BLIP, annotated	Yes	None	No
LAION-100 (LAION100)	X	✓	alt-text	No	CLIP	No
YFCC-100 (YFCC100)	X	✓	Flickr	No	Algo	No

Table 2: ***The JANuS dataset allows for controlled comparisons between VL-loss and CE-loss models in a high accuracy regime.*** *The experiments in Fig. 1 were conducted using a combination of the four main datasets in JANuS, which are described here.* ***G.T. Lbl.*** *indicates the presence of human-annotated ground truth labels in the dataset.* ***Machine Lbl.*** *indicates availability of synthetic labels; labeling strategies are detailed in D.* ***Caption Src.*** *lists the sources for captions in the dataset.* ***Supervised*** *indicates when ground truth labels exist for the dataset. CE-loss models benefit most from supervised data.* ***Filtered*** *indicates when the dataset contents were processed in some way prior to inclusion in the dataset. VL-loss models struggle on unfiltered data.* ***Balanced*** *indicates whether the dataset is approximately class-balanced.*

selecting or adding data from a known data source. Data sources for which ground truth labels exist are filtered by class. For unsupervised data sources, we use a technique called *subset matching* to prefilter JANuS. A detailed explanation of this technique can be found in Sec. H.1. The primary advantage of JANuS over its constituent datasets is that every sample has descriptive captions *as well as* class labels (either as human annotated or synthetic labels), and is compatible with IN100 classification. This allows for JANuS to be used to fairly compare both image and image-text training approaches while controlling for dataset size and quality. We propose that JANuS be used as both a standard benchmark and a source of high quality training data. The constituent datasets are the following:

1. **ImageNet-100 (IN100):** The 100 largest ImageNet-Captions classes from Fang et al. (2022), followed by class rebalancing by addition of over 50,000 new image samples annotated with human-authored ground-truth labels.
2. **OpenImages-100 (OI100):** A subset of the OpenImages dataset, Kuznetsova et al. (2018), with restored original Flickr-captions, and new BLIP-captions; samples selected by mapping human-labeled OpenImages-100 classnames to ImageNet-100 classnames.
3. **LAION-100 (LAION100):** A subset of the unlabeled LAION dataset, Schuhmann et al. (2021), with samples selected via subset matching on ImageNet-100 classes.
4. **YFCC-100 (YFCC100):** A subset of the unlabeled YFCC dataset, Thomee et al. (2016), with samples selected via subset matching on ImageNet-100 classes.

We compare some of the key properties of each component of JANuS in Tab. 2. Detailed information on the process used to create JANuS, as well as the composition of each subset, is available in Sec. D.

Performance Variations in JANuS. In order to ensure that the baseline performance of VL-loss and CE-loss models is comparable on IN100 and the standard ImageNet despite the newly added images, we train a VL (using the standard "A photo of a $CLASSNAME" prompt) and CE-loss model from scratch on IN100, and compare it to a CE-loss model trained for 256 epochs on the same 100-class subset of ImageNet. Controlling for size, we find that our dataset performs slightly worse than the baseline, but considerably better than that subset of ImageNet-captions alone.

Despite the fact that OI100 is a slightly larger than IN100, we find that models trained on OI100 perform worse on IN100-Val than models trained on IN100. We find that the extreme class imbalance shown in Fig. D.1 is the cause of most, but not all, of the decrease in accuracy (See models (in100-sup, oi100-sup-int, oi100-sup-int-classbal) in the Appendix, Table Sec. H).

VL-loss class imbalances (detected by searching for exact-match classnames in caption strings) are also present in the other web-scraped datasets in JANuS, LAION and YFCC; this may contribute to the lower performance of VL-loss models on long-tailed classification.

Training on JANuS. In order to minimize differences in model architecture, we train two families of models: A ResNet-50 for CE-loss models, and a VL-loss model with a ResNet-50 vision backbone. The only difference in the two architectures is that for CE models, we append a ResNet-50 with a 1000-class linear head; we allow this since, as noted in Radford et al. (2021); Santurkar et al. (2022), this does not affect CLIP performance. To control for dataset size, we train models on various subsets of JANuS and measure base accuracy and distributional robustness.

We train with mixed precision, at a batch size of 256, and do not use gradient clipping. We use the AMP library to implement the training process. Model hyperparameters are chosen via grid search. Models are typically distributed across a single node with 4 NVIDIA GPUs; our largest models were trained on 16 NVIDIA GPUs. We train non-JANuS models for 32 or 64 epochs unless otherwise specified. All JANuS models were trained for 256 epochs.

Following Santurkar et al. (2022), we use SimCLR augmentations (resize, crop, flip, jitter, blur, grayscale) rather than CLIP augmentations (resize and crop) for model training. We share our code for reproducibility.

Evaluation for models trained on JANuS. Following Nguyen et al. (2022), we measure performance on IN100-Val, regardless of the choice of pretraining dataset. We report best accuracy scores, with ”best” being determined by the model's peak performance on IN100-Val rather than shifts. For ImageNet-R and ImageNet-A, which are subsets of ImageNet, we evaluate only the 35 shared classes.

6 Controlled ablation studies

In order to better understand which factors are most decisive in distributional robustness, in this section, we group and evaluate the models according to various factors which are thought to contribute strongly to it; specifically, we compare across VL-loss and CE-loss groups, utilizing ViT from Dosovitskiy et al. (2021) and convolution-based architectures, number of parameters, and size of training dataset. We also provide Spearman rank correlations for each feature in Tab. 3. We establish ordinal rankings based on the ordering we would expect based on the current "folk wisdom" about robustness: VL-loss > CE-loss, ViT > CNN; larger model size, and more training data leads to greater robustness.

Comparing VL-loss and CE-loss models. Large VL-loss models such as those of Radford et al. (2021); Pham et al. (2021) have been conventionally presented as robust generalist models which can handle arbitrary (open vocabulary) classification tasks.

Setting aside the appeal of their versatility, is it actually the case that VL-loss models performing some classification task are more robust than fully trained CE-loss models on that same task? We propose to examine this question from the perspective of model inference, as well as model training. First, we compare the performance of pretrained VL-loss and CE-loss models on three classification tasks, controlling for dataset size. Second, we train VL-loss and CE-loss models from scratch on JANuS and evaluate them on IN100, again controlling for dataset size.

Model Attr.	IN1000	IN100	IN100-Dogs
VL-loss	-.059	.025	-.092
ViT	.38	.384	.358
More training data	.328	.352	.188
Model size	.453	**.439**	**.378**
Image resolution	**.541**	.319	.252

Table 3: ***Large ViT-based architectures are positively correlated with model robustness; VL-loss has mixed, weak correlations.*** *We calculate the Spearman's rank correlation of each model attribute we consider for the 650 models in our study. Negative scores indicate negative correlation between the named attribute and average robustness.* ***Boldface*** *indicates the attribute with the strongest correlation. Choosing a ViT over a convolution-based architecture correlates positively with model robustness, as does increasing number of model parameters.*

Model inference comparison. For dataset sizes below 400 million samples, we find no reliable evidence that VL-loss models are more robust than CE-loss models in absolute terms on IN1000 or IN100-Dogs; see Fig. 3 (R). We also note that VL-loss models have lower base accuracy on these problems.

VL-loss models do show a robustness advantage on IN100. This is in part because smaller label sets are easier to disambiguate using natural language; in Sec. F we provide per-class accuracies for a VL-loss and CE-loss ResNet-50 trained on many samples, and note that several of the classes on IN1000 where VL models substantially underperform have identical natural language descriptions, making classification impossible; in IN1000, OpenAI's classnames include two classes labeled "missile" and two classes labeled "sunglasses", reflecting ambiguities in the underlying problem (Radford et al., 2021; Beyer et al., 2020). It is also the case that vision-language models have significantly higher parameter counts than standard computer vision models, due to their multimodal architecture.

In Fig. 3, we observe that increasing the parameter count produces larger gains on IN100 than on our other evaluations; increased model size could explain VL-loss's strong performance on this benchmark. We postulate that the difference in robustness is attributable, at least in part, to VL-loss's general difficulty with fine-grained class distinctions.

We train a VL-loss model on 10 million YFCC samples, filtering out all samples which contain a matching term with an ImageNet class, and find that the resulting model achieves just 3% accuracy on IN100-Dogs, but achieves 33% accuracy on IN100; in the absence of ground truth labels, the model 'guesses better', in essence, when classes are dissimilar. See Tab. 4.

Model	Dataset	**IN1000 Val. Acc.**	**IN100 Val. Acc.**	**IN100-Dogs Val. Acc.**
ResNet-50	YFCC-10Mn-N.I.	.127	.329	.034
ResNet-50	YFCC-15Mn	.324	.741	.086

Table 4: ***VL-loss models trained on web-scraped caption labels learn classes unevenly.*** *VL-loss models learn a lot about broad distinctions between classes from captions, and little about fine-grained class boundaries. This finding holds even when we remove all samples which match with any term in the OpenAI ImageNet classnames from the YFCC-15Mn dataset (YFCC-10Mn-N.I.). Robustness scores can be found in our main results table in the supplementary attachment.*

Above 400 million samples, our investigation is limited by the fact that relatively few public models have been trained on such huge datasets; our largest CE-loss models were trained on half the data of the largest VL-loss models, and they have few other architectural features in common. Limited evidence, however, indicates that VL-loss models have a robustness advantage at massive data scales.

Model training comparison. Ground-truth labels have been shown to improve base accuracy of VL-loss models. Fang et al. (2022) found that a ResNet-50 VL-loss model trained on ImageNet-1k with ground truth labels ("A photo of the CLASSNAME") achieved accuracy and robustness parity with a CE-loss ResNet-50 for IN1000 classification. In Fig. 1, we show that this is also the case for ResNet-50 models trained and evaluated on IN100. However, these models have low average robustness. When we attempt to increase the average robustness of the VL-loss model by augmenting the training dataset, we find that on IN100, the VL-loss models are consistently less robust than the CE-loss models when we control for dataset size.

Overall, we conclude that for the vast majority of real-world problems and data budgets, CE-loss will offer more robust performance than VL-loss.

ViT and convolution-based architectures. Another important component we consider when evaluating model robustness is the architecture.

The 650 models in our analysis include 385 convolution-based architectures and 204 vision transformers; despite the relative overrepresentation of convolution architectures in the study, of the 100 timm models with the highest average robustness on IN1000, 60 are ViTs and 40 are convolution-based architectures. On IN100, the split is 70 / 30, and on IN100-Dogs, 72 / 28.

Comparing the top 100 most robust models for each problem, we find that ViTs are, on average, substantially more robust, and the advantage grows at massive data scales (see Fig. 2 (L)). When we select the two largest models which are identical in all other respects, we see the same result in Tab. 5. The evidence in our study addresses datasets at least the size of ImageNet; for datasets substantially smaller than ImageNet, we recommend supplementing the dataset with ImageNet classes to improve the model's generalization ability.

Effects of scaling model parameters. In our analysis, 438 of the models have fewer than 50 million parameters, 126 have between 50 and 100 million, and 86 have over 100 million parameters. Of the top 100 most robust models, 13 have fewer than 50 million, 33 have between 50 and 100 million, and 54 have over 100 million parameters.

Model	Param. Count.	Val. Acc.	Avg. Rob.
CAIT-m48-448	356m	.863	.630
ResNet-V2-101x3-bitm-448	388m	.854	.573

Table 5: ***ViTs are more robust when controlling for other factors.*** *We select two of the largest models in our study, a CAIT (Touvron et al., 2021) vision transformer and a Big Transfer ResNet (Kolesnikov et al., 2019), with approximately similar parameter counts, identical input image resolution and identical training data (ImageNet-1k). The ViT is substantially more robust.*

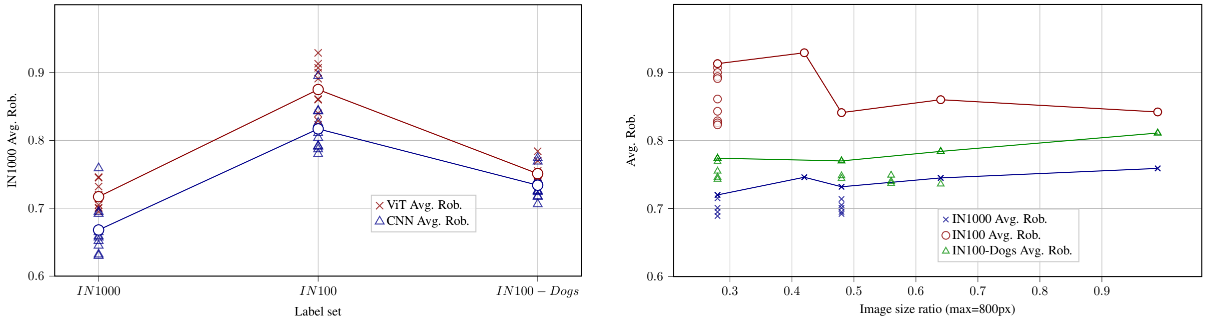

Figure 2: ***(L) ViT and CNN comparative robustness. (R) Comparative robustness as image resolution increases.*** *(L) We compare average robustness of ViT and CNN architectures across all three label sets; individual marks on the graph represent the average robustness of the ten most robust models for each evaluation. Trend lines follow the group average. The ViT models have an advantage on all evaluations. (R) We evaluate average robustness of models in the study when grouped by input image resolution, and find a weak positive association on IN100-Dogs and IN1000. The trend lines follow the best performing model in each group.*

In Fig. 3 (L), we compare model performance on our three evaluation metrics, grouped by parameter count. We find that average robustness improves reliably with model size across all evaluations, although the gains are most significant on IN100.

Effects of scaling input image resolution. In Fig. 2 (R), we plot average robustness against image resolution, expressed as the ratio of actual model resolution to maximum model resolution in the study (800px). We find that increasing input image resolution leads to gains in robustness on IN1000 and IN100-Dogs, but that these effects are smaller than choice of architecture and number of model parameters.

Effects of scaling data quantity. Researchers in robustness such as Fang et al. (2022) have argued that diverse, and presumably large, training distributions account for the strong robustness of VL-loss models, and that among CE-loss models, factors other than data have little impact on robustness, except insofar as they increase base accuracy (Taori et al., 2020). However, these studies were conducted prior to the surge in popularity of ViT-based architectures, which show greater robustness than convolution-based architectures. (See Sec. 6, ViT vs CNN).

In Tab. 6, we contrast the marginal robustness gain of going from a VGG-16 from Simonyan & Zisserman (2014) to the best model trained on ImageNet-1k alone, ImageNet-21k, and *any amount of data*. We find that the average robustness gain from scaling data alone ranges from 3.6% to 12.1%, depending on the label set. While these gains are substantial, they are nevertheless much smaller than the combined effect of other factors; a large and diverse dataset is a necessary, but far from sufficient, condition for optimal robustness.

100-class problems. On 100-class problems, we find that the robustness advantage of large-data models is most prominent when there are relatively few classes and their visual differences are relatively large. On fine-grained problems, the situation is actually reversed; it is the small-data models which have the advantage (in base accuracy).

Controlled experiments in data scaling. To better understand how and when robustness emerges during training, we conduct scaling experiments on JANuS which control for architecture, model size, image resolution and data diversity.

Model	IN1000 Avg. Rob.	IN100 Avg. Rob.	IN100-Dogs Avg. Rob.
Best (Any Data)	.759	.929	.811
Best (14 Mn)	.745	.861	.784
Best (1.2 Mn)	.682	.760	.749
VGG-16 (1.2 Mn)	.266	.402	.433

Table 6: ***The impact of massive data is limited for most image classification problems.*** *The best-performing model on an arbitrary data budget shows only minor improvements in average robustness compared to the best model trained on ImageNet-21k; gains are largest on IN100. The combined improvement from going from a VGG-16 (Simonyan & Zisserman, 2014) baseline to the most robust model trained on ImageNet-1k alone contributes far more to model robustness than training data in isolation.*

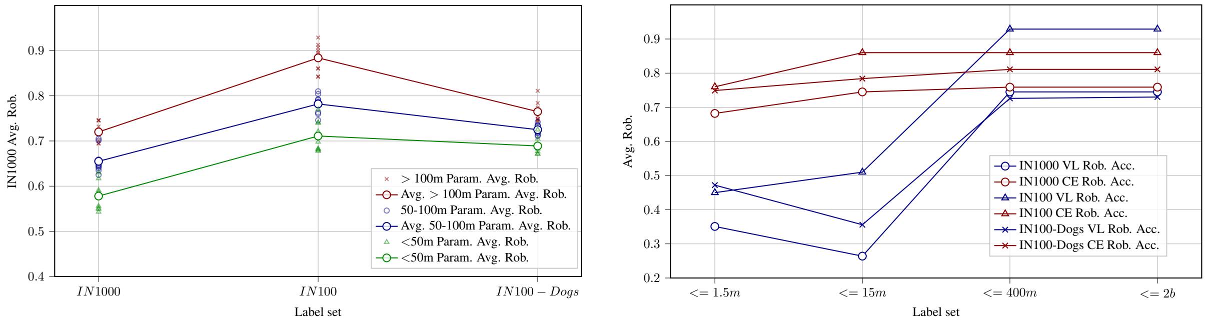

Figure 3: ***(L) Comparative robustness as parameter count increases. (R) Comparative robustness of VL-loss and CE-loss models on arbitrary data budgets.*** *(L) When we compare the average robustness of models in the study by their parameter count, we see reliable improvements as model size increases; larger models are more robust on all label sets. (R) We compare the most robust VL-loss and CE-loss models for every tier of dataset size across three different evaluation metrics. Models trained on fewer than 1.5m samples are trained exclusively on supervised data; larger models are trained on a mix of supervised and semi-supervised data. VL-loss models are more robust on IN100. CE-loss models are more robust on IN1000 and IN100-Dogs, and when less data is available.*

In this setting, we find that increasing the quantity of training data steadily increases robustness, up to and including 20x scaling, the largest amount of data we evaluated on JANuS (see Fig. 1).

Taken together, we conclude that in lower data regimes, scaling data reliably improves robustness, but that the benefits tail off sharply as the scaling multiple grows.

Image diversity in pretraining datasets. Massive training datasets have a inherent property of being very diverse, specifically displaying large intra-class variance. Naturally, distributional robustness being an effect of generalization, we therefore study the effect of image diversity on robustness.

We compare two recent computer vision architectures with three ResNet-50 architectures trained on different quantities of per-class data, and find that architecture is the key factor in determining how robustly models generalize between classes (see Tab. 7.) Since the specifics of CLIP's training data are unknown, we model average class frequency in CLIP's dataset using term matching on the public CC12m dataset, averaging across all classes. (Changpinyo et al., 2021)

The size of the combined JANuS dataset is very similar to the size of ImageNet-1k. The label set size of JANuS, however, is 1/10th that of ImageNet-1k. When we control for model architecture, the difference in robustness is stark; the model trained on JANuS has a large advantage compared to the model trained on ImageNet-1k.

More recent architectures such as VOLO (Yuan et al., 2021) and ConvNeXT (Liu et al., 2022), however, which are pretrained on ImageNet-1k, achieve robustness comparable to the JANuS model, despite seeing far fewer *in-class* examples. Selecting a robust architecture, then, offers the additional benefit of requiring fewer per-class examples during training.

Model	Dataset Size	Approx. Class Size	Val. Acc.	Avg. Rob.
ConvNext-L	1.2m	1,200	.973	.662
volo-d4	1.2m	1,200	.977	.707
ResNet-50	1.2m	1,200	.956	.504
ResNet-50 (JANuS+yfcc)	2.4m	24,000	.927	.701
ResNet-50 (CLIP)	400m	82,400	.9	.707

Table 7: ***Recent architectures perform better under a data budget.*** *We compare two recent computer vision architectures (ConvNeXT and VOLO) with three ResNet-50 architectures, trained on different quantities of per-class data. When comparing between ResNet-50 models, data scaling improves model robustness; however, modern architectures achieve comparable robustness with far fewer per-class examples.*

7 Discussion and Useful Heuristics

Our detailed ablation studies in Section 6 demonstrate the effects of various model and data choices on distributional robustness.

We conclude with a summary of takeaway points, along with a list of suggested useful heuristics for training robust models under various data budgets and problem sizes.

1. For **few-class problems** where either the classes themselves or their *visual properties (color, shape, a type of <SUPERCLASS>) are easily disambiguated using text alone*, the most robust and most efficient approach is to **use a zero-shot VL model.** On such problems, even a small ResNet-50 CLIP model performs quite well, and the larger CLIP models are consistently the most robust, at the cost of almost no loss in base accuracy.
2. However, for **fine-grained classification problems**, problems with **ambiguous class names**, and **many-class** problems, the best approach is to train a **CE-loss model** with a **large ViT-based robust architecture** at high image resolution. To choose an appropriate robust architecture for any particular label set, we refer to our table of complete results, which can be found in the supplementary materials.
3. **Transformer architectures, such as ViTs, benefit from data scaling** even when data is not in the target label set. Therefore, on low data budgets, it is best to conduct some **pretraining**. One practical approach with small dataset sizes is to **fine-tune** an existing pretrained model. Another approach is to train the model on a large label set, supplementing in-class training images with labeled images from ImageNet-21k, and then zero out unneeded class logits during inference.

Future work. As computer vision models and datasets grow in size, and multimodal generative models such as OFA from Wang et al. (2022a) introduce and solve new, complex problems, the task of developing a prescriptive set of "scaling laws" for emergent distributional robustness will only increase in importance (Cherti et al., 2022). Equally important will be comparing the behavior of models on distribution shifts for datasets other than ImageNet. Finally, a comprehensive understanding of model performance on long-tailed classification problems (such as iNaturalist) will shed more light on the robustness profile of models in the real world.

References

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *ICLR*, 2022. 26

Lucas Beyer, Olivier J. H'enaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *NeurIPS Workshop: ImageNet Past, Present, Future*, 2020. 7

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher 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. *NeurIPS*, 2020. 2

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 9, 16

Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. *NIPS*, 2017. 26

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. *ArXiv*, abs/2212.07143, 2022. 10

Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Arxiv*, 2021. 26

Stéphane d'Ascoli, Hugo Touvron, Matthew L. Leavitt, Ari S. Morcos, Giulio Biroli, and Levent Sagun. Convit: improving vision transformers with soft convolutional inductive biases. *Journal of Statistical Mechanics: Theory and Experiment*, 2021. 26

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 15, 31

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 6

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (CLIP). In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *ICML, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research, 2022. 1, 2, 3, 4, 5, 7, 8, 16, 20, 26, 31

Benjamin Feuer, Ameya Joshi, and Chinmay Hegde. A meta-analysis of distributionally-robust models. *ICML PODS Workshop on Distribution Shift*, 2022. 3

Ben Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. *ICCV*, 2021. 26

Kai Han, Yunhe Wang, Qiulin Zhang, Wei Zhang, Chunjing Xu, and Tong Zhang. Model rubik's cube: Twisting resolution, depth and width for tinynets. *NeurIPS*, 2020. 26

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019. 1

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021a. 16

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021b. 16

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. *CVPR*, 2016. 26

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. URL `https://doi.org/10.5281/zenodo.5143773`. 4, 26

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1, 2

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *ECCV*, 2019. 8, 14, 26

Alina Kuznetsova, Hassan Rom, Neil Gordon Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4. *International Journal of Computer Vision*, 2018. 5

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4. *International Journal of Computer Vision*, 2020. 31

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022a. 26, 31

Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *NeurIPS*, 2022b. 26

Zhuang Liu, Hanzi Mao, Chaozheng Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CVPR*, 2022. 9

Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *International Workshop on Computational Aspects of Deep Learning at 17th European Conference on Computer Vision (CADL2022)*, 2022. 26

John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. *ICML*, 2021. 1, 2, 16

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *NeurIPS*, 2022. 1, 2, 3, 4, 6, 16

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning. *ArXiv*, abs/2111.10050, 2021. 1, 2, 6

Trong Huy Phan and Kazuma Yamamoto. Resolving class imbalance in object detection with weighted cross entropy losses. *arXiv*, 2020. 3

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 26

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 *ICML*, 2021. 1, 2, 3, 5, 6, 7, 14, 26

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 1, 2, 16

Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a caption worth a thousand images? a controlled study for representation learning. *ICLR*, 2022. 3, 5, 6

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *Data Centric AI NeurIPS Workshop*, 2021. 5, 17, 31

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 8, 9

Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. *CVPR*, 2021. 26

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ICML*, 2019. 26

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *NeurIPS*, 2020. 2, 3, 8, 16

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl S. Ni, Douglas N. Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. *Commun. ACM*, 59:64–73, 2016. 5, 17, 31

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. *ICCV*, 2021. 8, 26

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. *ECCV*, 2022. 26

Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. *CVPR*, 2021. 26

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. In *MICCAI*, 2019. 1

Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary Chase Lipton. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019a. 16

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *CVPR*, 2019b. 26

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052, 2022a. 10

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *ArXiv*, 2022b. 2

Ross Wightman. Pytorch image models. *GitHub repository*, 2019. doi: 10.5281/zenodo.4414861. 4, 26

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *ICML*, 2022. 14

Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. *CVPR*, 2018. 26

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. *CVPR*, 2020. 26

Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. *ICCV*, 2021. 26

Ismet Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Kumar Mahajan. Billion-scale semi-supervised learning for image classification. *Arxiv*, 2019. 26

Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. *CVPR*, 2017. 26

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022. 2

Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 9

Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, Sercan Ö. Arik, and Tomas Pfister. Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. In *AAAI Conference on Artificial Intelligence*, 2021. 26

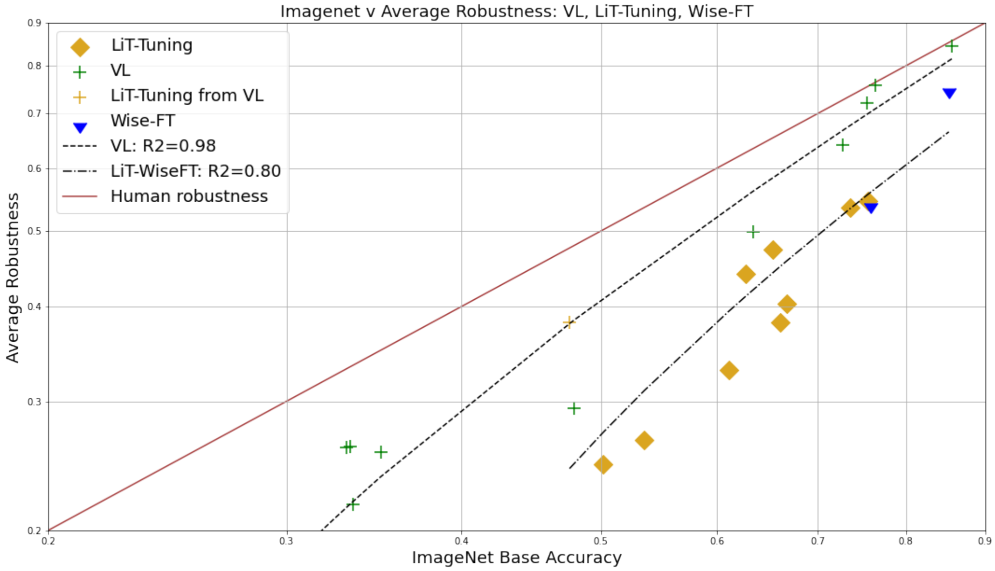

Figure 4: **Wise-FT, optimized to balance id/ood accuracy, fits the LiT-tuned effective robustness line.** Both Wise-FT and LiT-tuning exhibit lower effective robustness than conventional vision-language pretraining.

A Transfer learning in vision-language models

Another approach to robust classification in VL is using some form of transfer learning instead of training from scratch. The robustness advantages of transfer learning are well understood in conventional computer vision (see Kolesnikov et al. (2019)), and many recent model releases include variants which are pretrained on ImageNet-21k. Such models generally exhibit improved robustness when compared to models trained on ImageNet-1k alone (See main table in supplemental attachments).

There are a few prominent strategies for transfer learning in VL-loss models as well; we catalog them below and discuss their strengths and weaknesses.

Fine-tuning VL models. Unfortunately, the unique robustness properties of VL-loss models are not conserved when the image tower alone is fine-tuned. As reported in Radford et al. (2021), fine-tuning the VL-loss vision tower using a CE-loss objective improves base accuracy but degrades robustness. This effect grows stronger the longer the model is fine tuned, making fine-tuning the image-tower an inefficient solution for problems where robustness is a consideration.

A similar effect takes place if both vision and language towers are fine-tuned on ground-truth caption data; after 4 epochs of fine tuning on IN1000, a ViT-L-14 CLIP base accuracy improves from .76 to .83; however, average robustness declines from .72 to .69. (See main table in supplemental attachments).

Wise-FT, introduced by Wortsman et al. (2022) is a fine-tuning method which interpolates the weights of zero-shot CLIP with its fine-tuned counterparts. For certain distribution shifts, it is possible to find a 'sweet spot' where both i.d. and o.o.d. accuracy increase. However, Wise-FT models lose zero-shot capability, and are still not as robust as VL-loss models with the same base accuracy.4

LiT-tuning. LiT-tuning, or locked-image text-tuning, is an alternate approach to vision-language training in which a pretrained image tower is aligned with an untrained language model. LiT-tuned models are somewhat more data-efficient than VL models trained from scratch, but they, too, are not as robust as VL-loss models with the same base accuracy. (See 4).

Additionally, we observe the following;

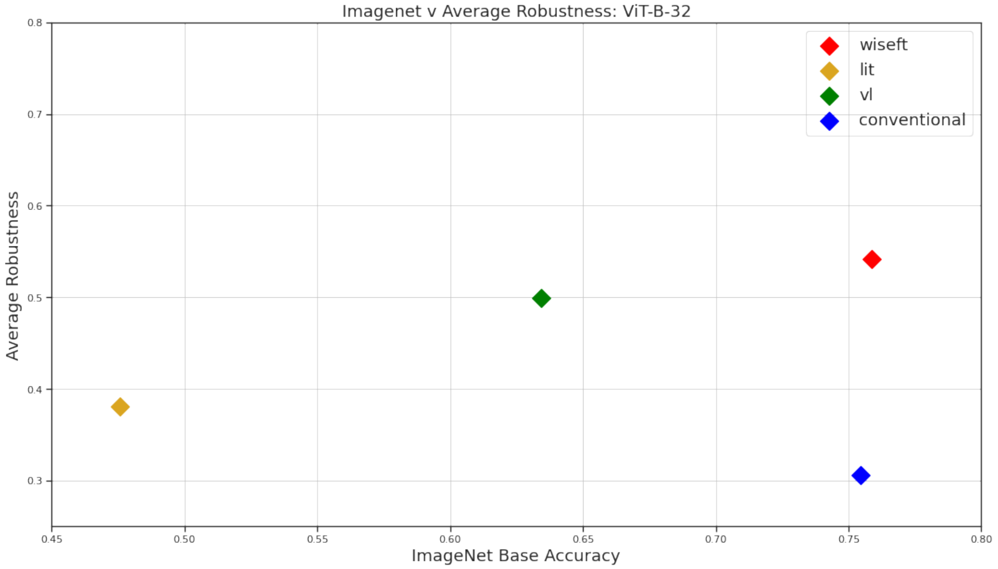

Figure 5: **LiT-tuning on a VL-trained image tower reduces accuracy without altering effective robustness, suggesting that VL pretraining is at least as robust as LiT-tuning.** *Wise-FT tuning greatly increases base accuracy and slightly improves effective robustness, at the cost of zero-shot capability. CE from-scratch training matches Wise-FT accuracy, but sacrifices effective robustness and zero-shot.*

1. Like Wise-FT, LiT-tuning produces models whose i.d. / o.o.d accuracy trade-off fits a line between that of traditional models and VL models – more robust than the former, less robust than the latter. The only exception we found was when we LiT-tuned the vision tower of a ViT trained on the CLIP objective – in this case, LiT-tuning decreased base accuracy while holding effective robustness constant (the near-opposite effect of Wise-FT)
2. LiT-tuning offers negative benefit for fully trained VL models, suggesting that it can only hope to approach, rather than exceed, the accuracy of its baselines (See 5)
3. LiT-tuning performance tends to closely correlate to the base accuracy of the underlying vision model
4. Intriguingly, we find that this is true regardless of the specific dataset used for LiT-tuning – LiT-tuned models trained on small amounts of data are able to recover accuracy on out-of-distribution tasks even when very little data from that distribution shift appears in the pretraining data
5. These experiments suggest that some degree of effective robustness is "locked away" in many vision models, but is lost during the training process, but that certain techniques are able to increase effective robustness disproportionate to the loss in base accuracy, pushing the model 'above the line' we would normally expect. Furthermore, if the distribution shift of interest is known and well-defined, it is possible to select a tuning to optimize for that shift

B Distribution Shifts

ImageNet is a large-scale visual ontology of images built upon the backbone of the WordNet structure. ImageNet aims to populate the majority of the 80,000 synsets of WordNet with an average of 500–1000 clean and full resolution images, making it a roughly class-balanced, fully supervised dataset. Deng et al. (2009)

ImageNet-21k, the largest version of ImageNet, contains 14,197,087 images in 21,841 classes.

There now exist a wide range of distribution shifts on ImageNet. These are novel test datasets designed to overcome some of the limitations of the original benchmark. While they cannot remedy issues with the labeling scheme, these datasets do provide challenging new contexts in which to analyze classifier performance.

ImageNet-V2 was designed to duplicate, as closely as possible, the original ImageNet test set. It was intended to answer the question of whether ImageNet-trained classifiers could successfully generalize even to the most mild of distribution shifts.Recht et al. (2019)

Imagenet-Sketch is a distribution shift covering sketches, paintings, drawings and illustrations of ImageNet classes. This test set is very large and comprehensive.Wang et al. (2019a)

Imagenet-R is a 200-class subset of ImageNet-2012 focused on renditions of everyday objects, defined broadly as drawings, paintings, photographs of food art, etc.Hendrycks et al. (2021a)

Imagenet-A is a 200-class subset of ImageNet-2012 which was algorithmically selected – the natural distribution shift captured here is the set of ImageNet-class images which most often fool a RN50. This test is challenging, and tends to include a lot of images with challenges such as occlusion, changes in angle or position, and changes in brightness.Hendrycks et al. (2021b)

B.1 Different shifts respond to different interventions

Recent works such as Fang et al. (2022) demonstrate the power of effective robustness as an explanatory tool for performance differences in VL models; Miller et al. (2021) showed that there exists a strong correlation between most models trained on random subsets of a data distribution, and the fully trained model. However, these authors also caution that it has significant limitations – Taori et al. (2020) and Nguyen et al. (2022) show that models trained on more (or different data) can significantly change the effective robustness line of a particular model, and also that these changes were shift-specific, with stronger fits on shifts like ImageNet-V2 and weaker fits on shifts like ImageNet-A.

We find that ImageNet-V2 responds more to model architecture than other shifts, with the handful of non-ResNet models we evaluated outperforming nearly all other models, regardless of training objective.

ImageNet-R and ImageNet-Sketch both showed high sensitivity to the training data, with the CC12M and LAION-15m distributions considerably outperforming even the best YFCC-trained models. These types of shifts are particularly amenable to subset matching strategies.11, 9

On ImageNet-A, CE models significantly underperformed compared to VL models regardless of the data, and all models significantly underperformed compared to the ViT-L CLIP.10

We also note that there is no readily apparent logit-scaled linear trend in these distribution shifts when one considers models trained on a wide range of different datasets, underscoring the importance of a well-chosen baseline for comparison.

We find that different shifts tend to disadvantage different kinds of models, which makes improving on all of them simultaneously very challenging. The fact that ViT-L CLIP was able to do is both impressive and, given the vital importance of the underlying data distribution in such measures, a mystery which is unlikely to ever be solved. Even the massive public datasets such as LAION are unable to match the performance of the dataset CLIP was trained on, although other factors might possibly have played a role.

A standardized benchmark of distribution shifts on ImageNet would be a welcome contribution to this area of research.

C Pretraining Datasets

Today, many SOTA models are pretrained on web-scale unsupervised data. We utilized three such datasets in our experiments. We observe that one major challenge of conducting research on unsupervised datasets is that the links provided as part of the dataset fail more and more over time, leading to each group getting a different version of the dataset. Therefore, to the extent possible, we report the details of each dataset in the appendix, and encourage other researchers working with these datasets to do the same.

CC-12M is a lightly supervised web-scale dataset created by Google. The image-caption pairs in CC-12M were filtered and selected for the purposes of training models to caption images.Changpinyo et al. (2021) Our version of CC12M contained 9703885 image-caption pairs.

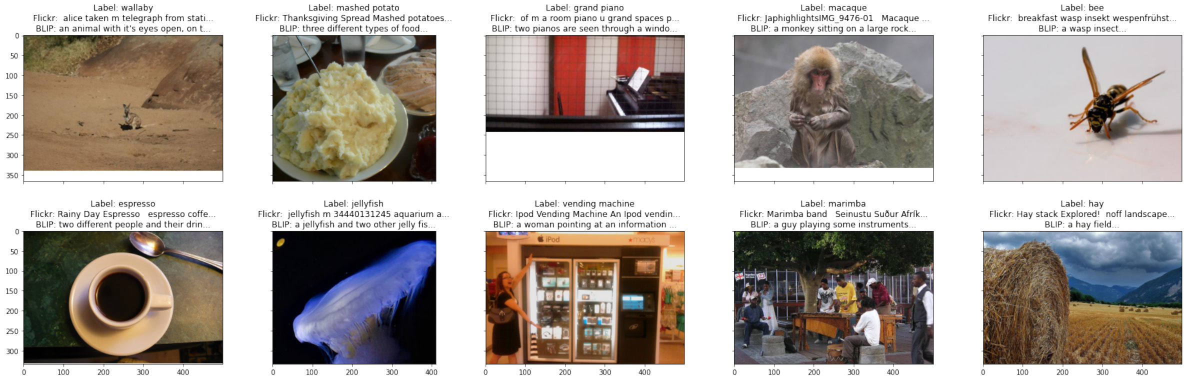
Label: wallaby
Flickr: alice taken m telegraph from stati...
BLIP: an animal with it's eyes open, on t...
Label: mashed potato
Flickr: Thanksgiving Spread Mashed potatoes...
BLIP: three different types of food...
Label: grand piano
Flickr: of m a room piano u grand spaces p...
BLIP: two pianos are seen through a windo...
Label: macaque
Flickr: JaphighlightsIMG_9476-01 Macaque ...
BLIP: a monkey sitting on a large rock...
Label: bee
Flickr: breakfast wasp insekt wespenfrühst...
BLIP: a wasp insect...
Label: espresso
Flickr: Rainy Day Espresso espresso coffe...
BLIP: two different people and their drin...
Label: jellyfish
Flickr: jellyfish m 34440131245 aquarium a...
BLIP: a jellyfish and two other jelly fis...
Label: vending machine
Flickr: Ipod Vending Machine An Ipod vendin...
BLIP: a woman pointing at an information ...
Label: marimba
Flickr: Marimba band Seinustu Suður Afrík...
BLIP: a guy playing some instruments...
Label: hay
Flickr: Hay stack Explored! noff landscape...
BLIP: a hay field...

Figure 6: **ImageNet-100 samples from JANuS.**

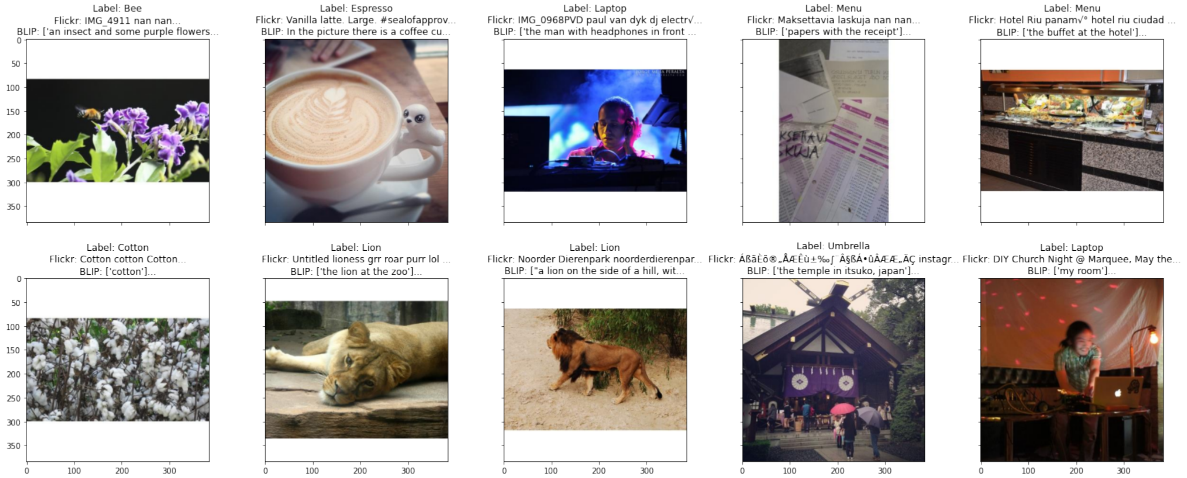
Label: Bee
Flickr: IMG_4911 nan nan...
BLIP: ['an insect and some purple flowers...
Label: Espresso
Flickr: Vanilla latte. Large. #sealofapprov...
BLIP: In the picture there is a coffee cu...
Label: Laptop
Flickr: IMG_0968PVD paul van dyk dj electr√...
BLIP: ['the man with headphones in front ...
Label: Menu
Flickr: Maksettavia laskuja nan nan...
BLIP: ['papers with the receipt']...
Label: Menu
Flickr: Hotel Riu panam√° hotel riu ciudad ...
BLIP: ['the buffet at the hotel']...
Label: Cotton
Flickr: Cotton cotton Cotton...
BLIP: ['cotton']...
Label: Lion
Flickr: Untitled lioness grr roar purr lol ...
BLIP: ['the lion at the zoo']...
Label: Lion
Flickr: Noorder Dierenpark noorderdierenpar...
BLIP: ["a lion on the side of a hill, wit...
Label: Umbrella
Flickr: ÁßãÈõ®„ÅÆÊù±‰∫¨Â§ßÁ•ûÂÆÆ„ÄÇ instagr...
BLIP: ['the temple in itsuko, japan']...
Label: Laptop
Flickr: DIY Church Night @ Marquee, May the...
BLIP: ['my room']...

Figure 7: **OpenImages-100 samples from JANuS.**

YFCC-15M is a subset of YFCC-100M, which is 100M image-metadata pairs taken from Yahoo-Flickr in 2016. The subset was selected by OpenAI. This dataset contains images and metadata, which includes a "title" and a "description" field. These fields are combined and processed in various ways by researchers in order to generate captions for models to train on.Thomee et al. (2016) Our version of YFCC contained 14825134 image-caption pairs.

LAION is a 5B image-caption dataset recently created by LAION.ai. It is the first publicly available dataset which matches the scale of the datasets used by the large companies to train their best models.Schuhmann et al. (2021) The subset of LAION we refer to as LAION-15m contained 13775512 image-caption pairs.

D Details on JANuS

The most important new contribution in JANuS is ImageNet-100. To the best of our knowledge, ImageNet-100 is the only version of ImageNet which duplicates the original distribution's class balance and supervision properties (ImageNet is not perfectly class balanced, but it does not contain any long-tail classes; all classes in ImageNet have at least 750 samples), while also being fully captioned with original web-scraped labels. We find that both VL and CE models trained on relatively small amounts of data can achieve high base accuracy on some JANuS subsets, making it possible for the first time to compare model distributional robustness while controlling for base accuracy.

In Table8, we discuss in detail the supervision strategy used for JANuS, with a per-class breakdown of each class.

An overview of the supervision process follows;

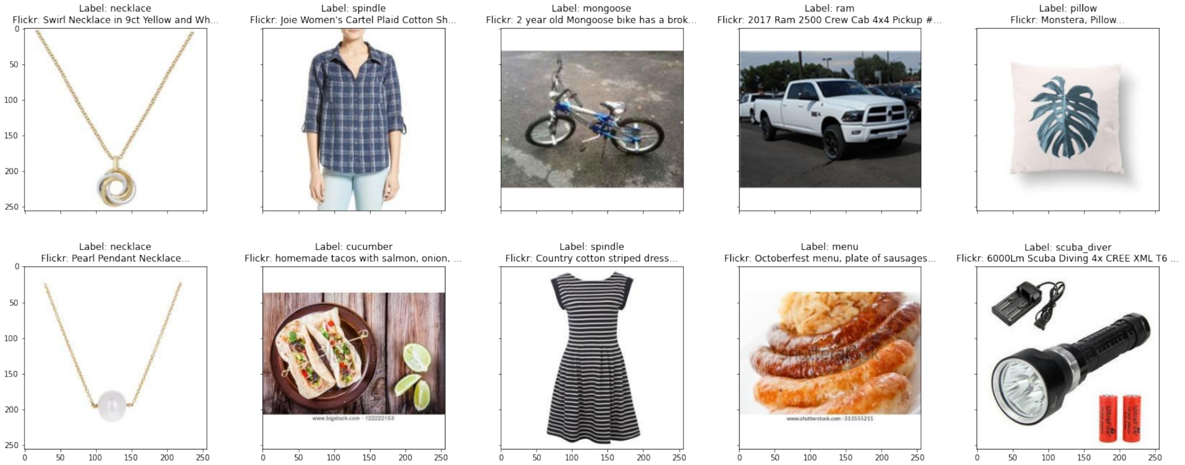
Label: necklace
Flickr: Swirl Necklace in 9ct Yellow and Wh...
Label: spindle
Flickr: Joie Women's Cartel Plaid Cotton Sh...
Label: mongoose
Flickr: 2 year old Mongoose bike has a brok...
Label: ram
Flickr: 2017 Ram 2500 Crew Cab 4x4 Pickup #...
Label: pillow
Flickr: Monstera, Pillow...
Label: necklace
Flickr: Pearl Pendant Necklace...
Label: cucumber
Flickr: homemade tacos with salmon, onion, ...
Label: spindle
Flickr: Country cotton striped dress...
Label: menu
Flickr: Octoberfest menu, plate of sausages...
Label: scuba_diver
Flickr: 6000Lm Scuba Diving 4x CREE XML T6 ...
0
50
100
150
200
250

Figure 8: **LAION-100 samples from JANuS.**

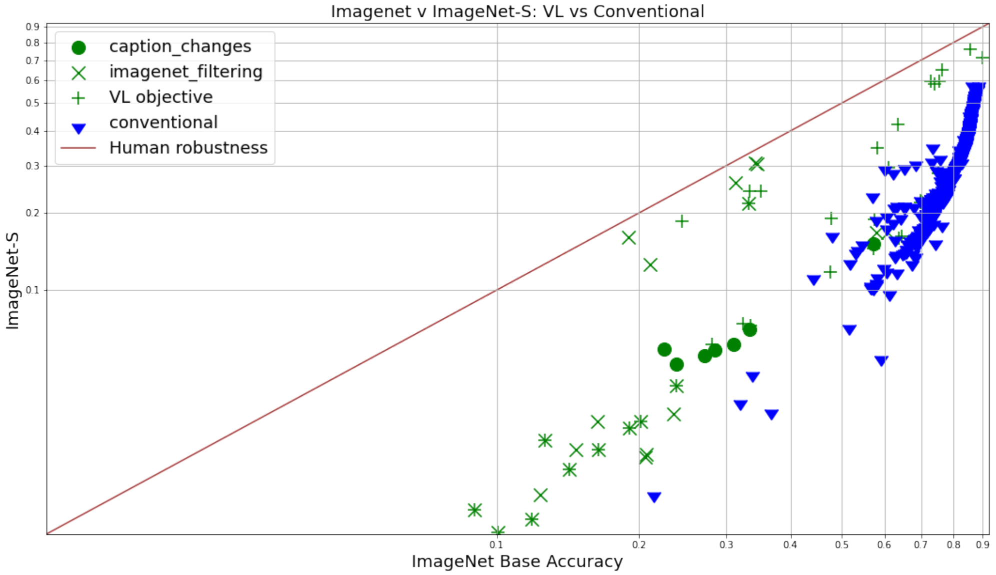

Figure 9: **Non-linearities in ImageNet-Sketch.** *ImageNet-sketch performance is not linear, with only the very largest VL models showing a reliable improvement over CEly trained models, when controlling for dataset size.*

- All samples were supervised by the authors of the paper
- Samples were sourced from flickr using the available API, sorted by 'interesting', with safesearch enabled, searching only samples with Creative Commons licenses
- Additional filtering terms were passed to the API in order to eliminate commonly encountered confounds in the search terms
- After the search term was selected, items were downloaded in bulk
- All downloaded samples were then individually tagged by the researchers as either "IN-class" or "out-of-class", using reference photographs from each class as a baseline comparison

We found that classes varied widely along several vectors;

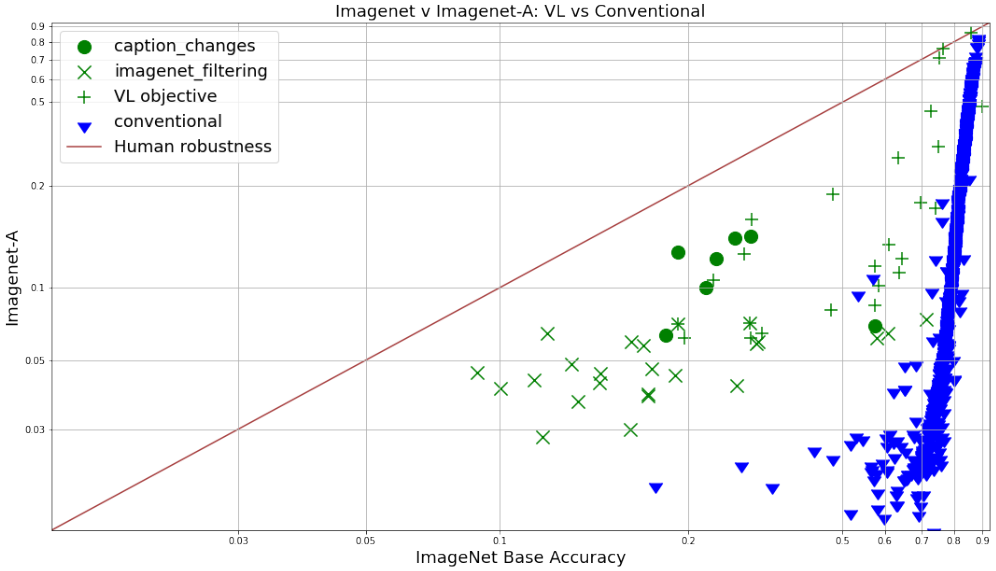

Figure 10: **ImageNet-A is learnable by all models at extremely high base accuracy.** *Although VL models seem to learn ImageNet-A faster than CE models, CE models reach near-parity with VL models when base accuracy gets very high.*

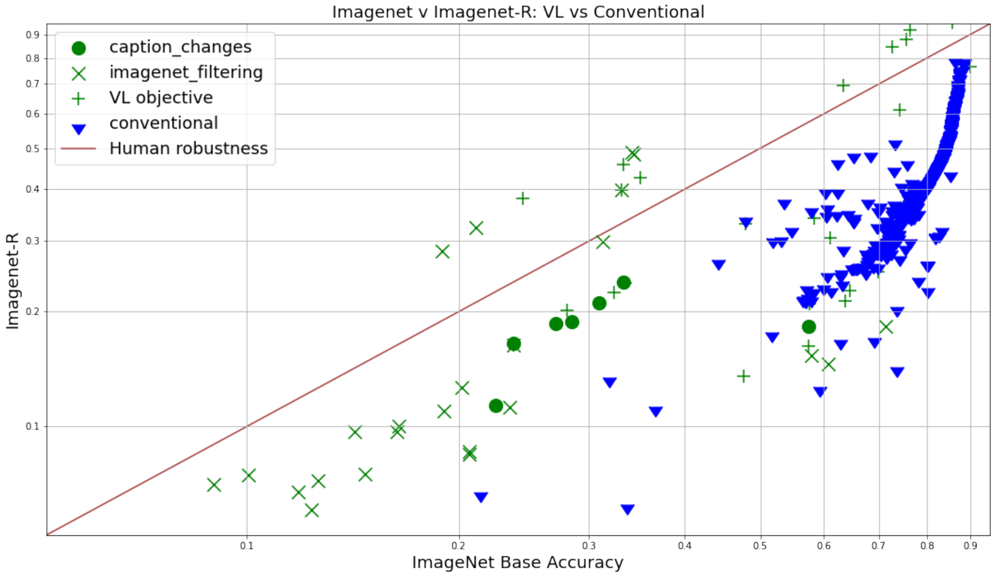

Figure 11: **VL performance on ImageNet-R outstrips base accuracy.** *On ImageNet-R, which is a 200-class subset of ImageNet, VL models are able to achieve higher accuracy than on ImageNet itself. VL continues to outperform CE models on this dataset, even at very high accuracies.*

- Some classes had far greater availability than others (ranging from 450,000 to 283 available samples)

- Some classes were much cleaner than others (ranging from 100 percent clean to around 25 percent)

- Some classes tended to be the 'subject' of photographs, such as dog breed, while others, such as mashed potato, tended to be featured as secondary items in the background of a photograph of something else

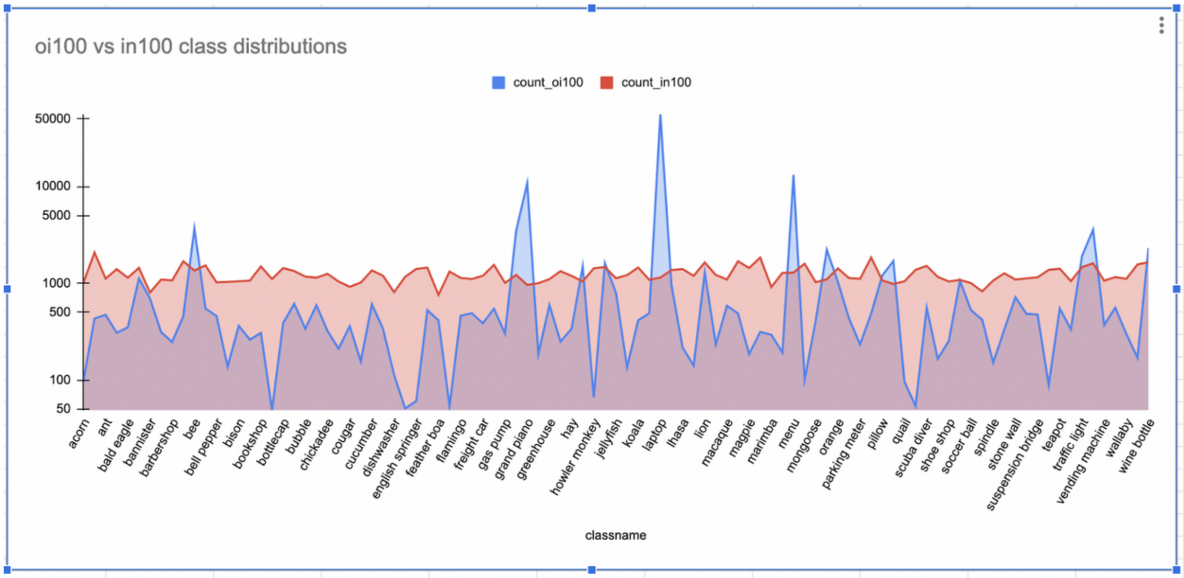

Figure 12: This log-scale figure shows the extreme class imbalance of the *unfiltered* OI100 dataset, compared to the *prefiltered* IN100 dataset; certain classes which are very common in web-scraped images, such as laptops, are overrrepresented, while others are not represented at all. The OI100 class imbalance is produced by a difference in dataset labeling strategies. VL-loss class imbalances (detected by searching for exact-match classnames in caption strings), which are present in the other web-scraped datasets in JANuS, LAION and YFCC, co-occur with comparatively low accuracy scores on fine-grained classification tasks.

D.1 Dataset Construction

The 100 classes in JANuS were selected randomly from a subset of all classes with more than 600 captions available in ImageNet-Captions (Fang et al., 2022). The list of classes selected is available in Section I. We note that this approach introduces a potential bias in class selection, since it may be that captions were still available for those images ten years after ImageNet was originally constructed for some reason that correlates with properties we are interested in studying; however, we feel that the risk of this is outweighed by the many benefits of having such a dataset available for study.

Table 8: **JANuS Supervision: Search Terms and Sample Quality** *Since many of the findings in our paper highlight the importance of both the amount and type of label noise, this table records statistics pertaining to our filtration process for the new samples in IN100. In the search term field, a - symbol indicates that all samples which included that word in the title, tags or description were NOT matched. Boolean OR, AND, and "" symbols behave as they typically do.*

in1k classname	search term	good samples	total samples	avail. samples	pct. good
lion	lion	962	1000	450000	0.96
wine bottle	wine bottle	925	1000	29500	0.93
book shop	bookstore	816	984	83000	0.83
parking meter	parking meter	377	1000	9500	0.38
african elephant	african elephant	885	1000	44000	0.89
bagel	bagel	699	988	20500	0.71
tarantula	tarantula	667	981	9000	0.68
ice cream	ice cream	741	984	154500	0.75
fig	fig	517	1000	46000	0.52
shoe shop	shopping shoes	425	1000	13000	0.43
french bulldog	french bulldog	887	996	7500	0.89
hen	hen	412	1000	73000	0.41
guacamole	guacamole	683	998	6500	0.68
broccoli	broccoli	679	997	19000	0.68
howler monkey	howler monkey	817	847	9000	0.96
scuba diver	scuba diver	827	1000	15000	0.83
spindle	"spindle wool, spindle -wool thread"	311	867	867	0.36
lhasa	lhasa dog	719	1000	2500	0.72
traffic light	stoplight	622	991	5500	0.63
lionfish	lionfish	552	897	6500	0.62
popsicle	popsicle -animal -sticks -animals -insect -insects -icicle -garden -sticks -icicles -gardens -toes -label -labels	638	943	7500	0.68
lampshade	lampshade	446	807	6500	0.55
spiderweb	spiderweb -spiderman -halloween -pumpkin -butterfly -pleiades -nebula -stars	832	996	17500	0.84
lifeboat	lifeboat	572	1000	13000	0.57
cucumber	cucumber -sea -spider -beetle -flower -spiral	730	999	26500	0.73
english springer	english springer spaniel	772	993	3500	0.78
macaw	macaw	972	1000	13500	0.97
mailbox	mailbox	900	1000	36500	0.9
peacock	peacock -butterfly	966	999	72000	0.97
bee	bumblebee OR wasp OR hornet -jet -airplane -helicopter -navy -aircraft -comic -RIAT -military -Helicopter -Helicopters -helicopters -aviation -Hudson -car -basketball -sports -Transformers -cosplay -disfrazado -costume -transformer AND flower	686	761	110000	0.9
dungeness crab	dungeness AND crab -restaurant -breakfast -lunch -dinner -shack -creels -traps -cannery	474	1000	1500	0.47

banana	banana -plant -blossom -flower -seed -seedlings -tree -spider -leaf -abstract -bay -band -festival -doll -sexy -sexiest -bread -soup -puree -smoothie -car -plantation -cake -cream -monkey -pudding -zoo -republic -boxes -buying -selling -vendor -bridge -scone -moon	793	995	65000	0.8
corn	corncob	354	1000	1000	0.35
lemon	lemon -plant -blossom -flower -seed -seedlings -tree -spider -leaf -abstract -bay -band -festival -doll -sexy -sexiest -bread -soup -puree -smoothie -car -plantation -scent -fresh -cleaner -butterfly -grove -shots -car -sunrise -paint -graffiti -origami -cake -cream -pudding -boxes -buying -selling -vendor -bridge -scone -don -lime	693	1000	65000	0.69
marimba	marimba instrument	127	283	283	0.45
orange	orange food fruit -plant -blossom -flower -seed -seedlings -tree -spider -leaf -abstract -bay -band -festival -doll -sexy -sexiest -bread -soup -puree -smoothie -car -plantation -cake -cream -monkey -pudding -zoo -republic -boxes -buying -selling -vendor -bridge -scone -moon -cupcake -cake -sales -seller -pancakes -crepes -crep -crepe -pancake -cookie -flavored -juice -soda -pop -beach -island -cove -grove -street -drive -tea -curd -marmalade -bars -cabs -chicken -cheesecake -pie -milk	744	1000	4000	0.74
bell pepper	bell pepper vegetable -plant -blossom -flower -seed -seedlings -tree -spider -leaf -abstract -bay -band -festival -doll -sexy -sexiest -bread -soup -puree -smoothie -car -plantation -cake -cream -monkey -pudding -zoo -republic -boxes -buying -selling -vendor -bridge -scone -moon -cupcake -cake -sales -seller -pancakes -crepes -crep -crepe -pancake -cookie -flavored -juice -soda -pop -beach -island -cove -grove -street -drive -tea -curd -marmalade -bars -cabs -chicken -cheesecake -pie -milk -market -spice	392	505	505	0.78
espresso	espresso coffee -maker -machine -beans -building -exterior -window	828	1000	22000	0.83
mashed potato	mashed potato	635	996	10000	0.64
stingray	stingray water -dolphin -shark -cruise -boat -scuba -fish	600	983	2000	0.61
flagpole	flagpole -lighthouse -church -bank -station	614	991	7000	0.62

teapot	teapot -tea -flower -tower -building -dome -art -fashion -vase -store -stores -shop -shops -Sagittarius -project365 -fountain -candle -mug -teacup -keg -vessel -amphora -urn -coffeepot	660	997	10500	0.66
umbrella	umbrella	911	1000	126000	0.91
beer bottle	beer bottle -house -door -brewery -glass -cap	909	1003	19000	0.91
barn	barn -swallow -owl -bird	980	1000	115000	0.98
christmas stocking	christmas stocking fireplace	317	779	779	0.41
magpie	magpie -screenshots -moth -butterfly -coprinopsis -thieving -mushroom	736	983	25500	0.75
mitten	mitten glove	800	995	1500	0.8
ram	ram sheep -Church -window -Window -church -school -dance -parade -festival -celebration -festivities -community -fair -ewe -fox -lamb -bird -cat -dog -Dodge	742	1000	3000	0.74
warthog	warthog animal -zebra -cheetah -leopard -giraffe -gazelle -hippo -rhino -donkey -armadillo -elephant -crocodile -lion -leopard -impala -cat -monkey -bird	946	997	2500	0.95
goose	geese	474	500	69000	0.95
bubble	soap bubble -dancer -dance -fairy -tree -leaf -leaves -flowers -water -toy -art -abstract -museum -dog -cat -butterfly -food -wine -beer -chocolate -Chocolate	414	500	5000	0.83
cougar	cougar animal -warthog -mascot -zebra -cheetah -leopard -giraffe -gazelle -hippo -rhino -donkey -armadillo -elephant -crocodile -lion -leopard -impala -cat -monkey -bird -lake -Lake -river -River -blonde -Blonde -woman -girl -milf -bear -cliff -Cliffs -cliffs -military -wallaby -horse -jet -print	297	500	1000	0.59
daisy	daisy flower	500	500	52000	1
menu	menu	431	500	92000	0.86
bald eagle	bald eagle	475	500	33500	0.95
necklace	necklace jewelry -brooch -pendant -creation -earring -earrings -bracele -ring -Engraver -bauble -anklet	478	500	12500	0.96
chickadee	chickadee bird -Goldfinch -goldfinch -robin -thrush -jay -cardinal -woodpecker -wren -hawk -raven -titmouse -nuthatch	494	500	9000	0.99
stone wall	"""stone wall"""	424	500	32000	0.85
flamingo	flamingo bird	476	500	38500	0.95
gas pump	gas station	348	500	41000	0.7
vulture	vulture bird -hawk -crow -eagle	489	500	15500	0.98
pizza	"""pizza pie"" -Fest -festival -summit -experience -party -band -moon -parade -Parade -harvard -mosaic -montage"	305	500	1000	0.61

wallaby	wallaby -warthog -mascot -zebra -cheetah -leopard -giraffe -gazelle -hippo -rhino -donkey -armadillo -elephant -crocodile -lion -leopard -impala -cat -monkey -bird -koala -sports -kangaroo -soccer -football -food -church -hills -stadium -tribute -grass -rugby -apartment -car	369	500	10000	0.74
hay	haystack field -hole -trail -poster -sign	360	500	1000	0.72
grand piano	"kawai grand piano, steinway grand piano"	312	455	455	0.69
laptop	laptop	443	500	98000	0.89
dishwasher	dishwasher appliance	191	268	268	0.71
cricket	cricket -batting -sports -team -match	337	500	44000	0.67
sea slug	nudibranch	468	500	12500	0.94
mongoose	mongoose -bike -bicycle -park -tree -joe -rocket -military -airplane -toy -car	379	500	5000	0.76
siamese cat	siamese cat -bangkok -flower -snake -campaign -wat -costume -cosplay -festival	416	500	13000	0.83
freight car	freight car	491	500	70500	0.98
vending machine	"""vending machine"""	411	500	13000	0.82
bottlecap	bottlecap -tab	448	500	3500	0.9
acorn	acorn -woodpecker -fairy -squirrel -weevil -travel -squash -street	352	500	25000	0.7
feather boa	feather boa	135	500	2000	0.27
macaque	macaque	485	500	14500	0.97
bolete	boletus	444	500	3500	0.89
border terrier	"""border terrier"""	422	500	1500	0.84
barbell	barbells	352	500	1000	0.7
fly	housefly	398	500	1500	0.8
suspension bridge	suspension bridge	432	500	33500	0.86
jellyfish	jellyfish	477	500	46500	0.95
barbershop	barbershop -quartet -singers	430	500	9000	0.86
koala	koala	458	500	32500	0.92
bannister	bannister staircase	174	183	183	0.95
pillow	pillow -talk -fight -cat -dog -moss -sky -cloud -sky	420	500	34500	0.84
bib	baby bib -shower -food	406	500	1500	0.81
junco	junco bird -finch -sparrow -thrush -cardinal -woodpecker -jay	475	500	7000	0.95
chainlink fence	chainlink fence	375	500	3500	0.75
soccer ball	"""soccer ball"" -match -game -milky -beach -Lewes"	349	500	2500	0.7
stupa	stupa	418	500	23500	0.84
quail	quail bird -finch -sparrow -thrush -cardinal -woodpecker -jay -partridge -rabbit -hawk -avocet -deer -dog -wolf -coyote -gopher -eagle -vole -molerat -butterfly	396	500	11000	0.79
padlock	padlock	378	500	9500	0.76
great white shark	"""great white shark"""	309	500	2000	0.62

totem pole	"""totem pole"" wood"	383	500	1000	0.77
ant	ant insect	447	500	18000	0.89
bison	bison	429	500	41500	0.86
greenhouse	greenhouse	407	500	82000	0.81

Adding BLIP Captions to JANuS

Since we could not find human-authored captions for ImageNet, we used BLIP Li et al. (2022a) to generate descriptive captions on ImageNet-100. BLIP often uses word fragments to describe objects, so we used a spell checker as a simple intervention to improve the quality of BLIP captions. Finally, because BLIP's vocabulary does not include many of the specialized classes available in ImageNet, we augmented the BLIP captions with Flickr image titles, the form of text which is most commonly available for an image. We used top p=0.9, max length=40, min length=5, repetition penalty=1.1.

We repeated the process for OpenImages-100. However, we used human-authored captions sourced from Pont-Tuset et al. (2020) instead of BLIP whenever available; around 16,000 out of the 135,000 OpenImages-100 samples had human-authored captions.

E Model Training Details

F Classwise Shifts

F.1 Per class accuracies for CLIP RN50 and SWSL RN50

In the supplementary files, we provide per-class confusion matrices on IN1000 for CLIP ResNet-50, trained on 400 Mn samples, as well as a semi weakly supervised ResNet-50 trained by Facebook on 1 Bn samples.Yalniz et al. (2019)

In addition to classnames which are literally identical (there are two instances of the class "missile" and two instances of the class "sunglasses" in the OpenAI classnames for IN1000), we find that the model struggles to disambiguate short words with similar starting token strings, such as "quail", "quilt" and "quill", and classes that start with common (and contextually misleading) words, such as "night snake".

G Details on models in study

Our meta-analysis made extensive use of the popular timm Wightman (2019) computer vision library, including models from Zhang et al. (2021); Bao et al. (2022); Kolesnikov et al. (2019); Srinivas et al. (2021); Touvron et al. (2021); Xu et al. (2021); Dai et al. (2021); d'Ascoli et al. (2021); Touvron et al. (2022); Huang et al. (2016); Yu et al. (2017); Chen et al. (2017); Maaz et al. (2022); Li et al. (2022b); Xie et al. (2020); Tan & Le (2019); Wu et al. (2018); Han et al. (2020); Wang et al. (2019b); Vaswani et al. (2021); Graham et al. (2021); for the complete list, please refer to the timm repository. In our supplementary results spreadsheet, the name field for each model is the same as that model's name in the timm repository – where model names have been modified between the time our evalutions took place and the time the paper was completed, we note the new names in the updated name column. CE-loss models evaluated in the study which can be cross-referenced by looking them up on timm.

Pretrained VL-loss model weights are taken from the open-clip Ilharco et al. (2021) repository, and can be cross-referenced on the repository's github page. The remaining models were trained by us for this study, and are described in the name field of the spreadsheet.

H Abbreviated JANuS Results

H.1 Subset matching strategies

For unsupervised web-scraped captioned datasets (such as LAION and YFCC), ground-truth class labels do not exist. Therefore, we must choose a strategy to assign class labels to samples in such datasets. VL-loss models use captions as labels. There is no easy way for CE-loss models to directly use captions as labels. To facilitate this, we propose a strategy we call *subset matching*, a modification of the "substring matching" technique proposed by Fang et al. (2022).

This strategy, illustrated in detail in 13, labels samples as follows. First, construct a dict of integers and "matching terms". A matching term is a string judged to be a good text representation of an image class, such as the string 'elephant' for an image of an elephant. Our standard choice of matching terms is based on Radford et al. (2021).

If a sample caption contains a matching term, then the corresponding integer class label is applied. If the sample caption contains multiple matching terms, then we apply one of three strategies, which we label strict, multi-class (mc) and single-class (sc) matching, explained in detail in H.1; we use single-class matching whenever possible, since it usually performs best. If the sample caption contains no matching terms for any class, then no label is applied and the image is dropped from the training set. Otherwise, the caption is replaced with the corresponding integer-valued label.

A **subset matching strategy** is an algorithmic method for applying machine labels to images, based on caption labels. All of these methods share in common the same underlying approach, as seen in 13.

In this section, we fully define and describe some important variations on the basic subset matching strategy as

Table 9: **Captioning strategy can affect model distributional robustness and accuracy.** *VL models perform better on OpenImages when flickr-captions are replaced with synthetic captions (BLIP+Title captions), but the same captioning method provides no benefits on ImageNet-100.*

Model	Technique	Val. Acc. (IN100)	Avg. Rob.	E.R.R.
ImageNet-100 BLIP-Caption	VL	0.574 ± 0.014	0.216 ± 0.008	0.376
ImageNet-100 Flickr-Caption	VL	0.574 ± 0.014	0.217 ± 0.008	0.378
OpenImages-100 BLIP+Human-Caption	VL	0.283 ± 0.012	0.131 ± 0.008	0.464
OpenImages-100 Flickr+Human-Caption	VL	0.225 ± 0.012	0.11 ± 0.008	0.489
OpenImages-100 Flickr-Caption	VL	0.197 ± 0.012	0.095 ± 0.008	0.482

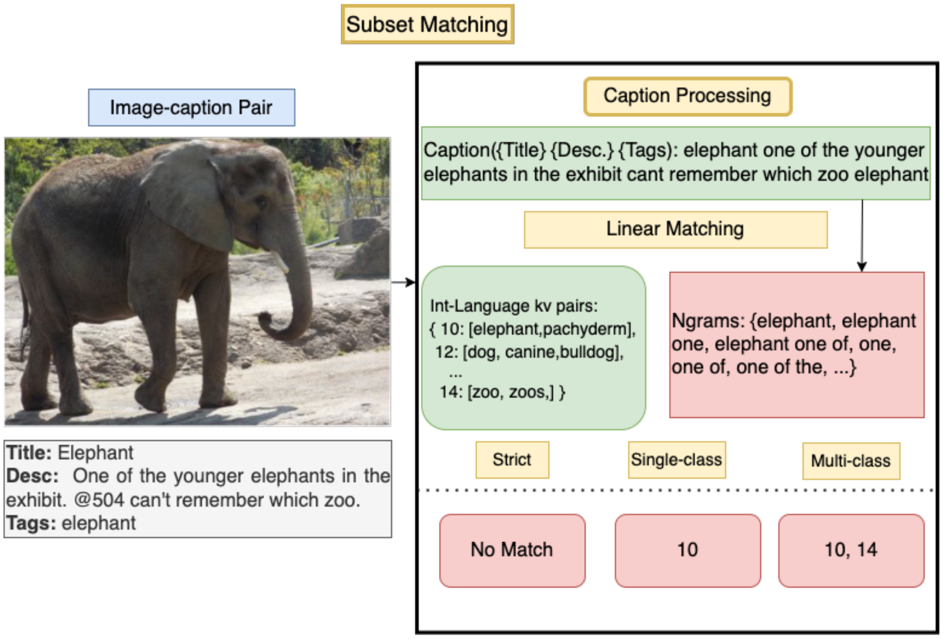
Subset Matching
Image-caption Pair
Title: Elephant
Desc: One of the younger elephants in the exhibit. @504 can't remember which zoo.
Tags: elephant
Caption Processing
Caption({Title}{Desc.}{Tags}: elephant one of the younger elephants in the exhibit cant remember which zoo elephant
Linear Matching
Int-Language kv pairs:
{ 10: [elephant,pachyderm],
12: [dog, canine,bulldog],
...
14: [zoo, zoos,] }
Ngrams: {elephant, elephant one, elephant one of, one, one of, one of the, ...}
Strict
Single-class
Multi-class
No Match
10
10, 14

Figure 13: **Subset matching; an overview.** *Subset matching is a simple labeling strategy for unsupervised image-caption pairs. The caption is processed and converted to n-grams which are then matched against a database of terms which point to integer-label classes.*

described in the main paper. All of our subset matching experiments utilized one of these three strategies.

Strict: Strict subset matching means that the model only applies the label to the image if the caption contains term(s) which map to exactly one class.

Strict subset matching was generally the most accurate method on ImageNet – we believe this is because of the ImageNet dataset filtration strategy, in which label selection is strongly dependent on caption contents.

It was generally less accurate method than single-class on OpenImages, where labels and caption contents are independently derived.

We find that strict matching performance tends to degrade when the pool of matching terms grow; it also tends to punish synthetic captions, which use a smaller vocabulary than web-scraped or annotated captions.

Single class: In single-class subset matching, the model greedily takes the first matching term to be the true class and ignores all subsequent matching terms.

As a general matter, we found that single-class matching struck the best balance between dataset utilization and accuracy, and we used this method for most of our experiments.

Multi class: In multi-class subset matching, we match up to 25 classes per sample (if we see multiple terms for a single class, we ignore those additional terms, and we do not attempt to rank classes by frequency).

The cross-entropy loss of the model is then given by the sum of the loss on each class; in other words, we reward the model for applying a high probability on each label assigned to the sample and for applying a low probability to each label which was not assigned to the sample.

This approach, while intriguing, was challenging because we only had one ground-truth label for each image; therefore, multi-class matching was always less accurate than single-class matching in direct comparison.

Since our cross-entropy model used a softmax loss, we found that model error tended to be high as the number of matched classes grew. We also found empirically that images which actually required multiple labels were not particularly common in our dataset. Perhaps for these reasons, this approach performed worse than single-class matching in most experiments.

Additional term definitions.

Label accuracy. On datasets for which supervised ground-truth labels exist, we report label accuracy as the count of machine-generated labels which match ground truth labels, divided by the total number of samples in the dataset.

Dataset utilization. Dataset utilization (Ds. Util) of a model on a dataset is the ratio of correctly labeled samples to total number of samples (including correct, incorrect and unlabeled samples). We use this metric to judge how useful a labeling strategy is; ground truth labels have a utilization of 100%; automated labeling methods gives typically significantly less utilization.

Abbreviated results table.

Please see Tab. H.1

I Class Frequency Counts for IN100 subset matching distributions, openai labels, mc matching

I.1 ImageNet-100

'orange': 1820, 'lion': 1788, 'barn': 1695, 'macaw': 1684, 'umbrella': 1583, 'banana': 1500, 'mitten': 1500, 'warthog': 1488, 'magpie': 1438, 'lemon': 1437, 'koala': 1435, 'espresso': 1400, 'bagel': 1376, 'howler monkey': 1337, 'tarantula': 1331, 'broccoli': 1299, 'fig': 1295, 'ice cream': 1285, 'cucumber': 1272, 'goose': 1231, 'daisy': 1224, 'junco': 1207, 'chickadee': 1193, 'teapot': 1175, 'french bulldog': 1166, 'vulture': 1150, 'stingray': 1142, 'guacamole': 1134, 'flamingo': 1126, 'lifeboat': 1120, 'ant': 1114, 'suspension bridge': 1109, 'greenhouse': 1100, 'lhasa apso': 1093, 'wallaby': 1073, 'stupa': 1073, 'bald eagle': 1063, 'lionfish': 1057, 'fly': 1055, 'english springer spaniel': 1051, 'necklace': 1048, 'bison': 1047, 'barbell': 1042, 'mailbox': 1041, 'quail': 1037, 'macaque': 1032, 'padlock': 1026, 'hen': 1024, 'pizza': 995, 'pillow': 995, 'acorn': 993, 'vending machine': 976, 'bottle cap': 969, 'stone wall': 968, 'popsicle': 955, 'spider web': 949, 'totem pole': 934, 'spindle': 920, 'bookstore': 903, 'bubble': 893, 'border terrier': 889, 'mongoose': 888, 'corn': 874, 'parking meter': 866, 'flagpole': 864, 'dungeness crab': 862, 'marimba': 862, 'peafowl': 848, 'bee': 840, 'bell pepper': 821, 'menu': 758, 'wine bottle': 734, 'great white shark': 733, 'jellyfish': 703, 'dishwasher': 701, 'soccer ball': 700, 'beer bottle': 663, 'grand piano': 600, 'bolete': 576, 'hay': 547, 'gas pump': 541, 'christmas stocking': 534, 'traffic light': 479, 'cougar': 471, 'scuba diver': 470, 'feather boa': 435, 'african bush elephant': 408, 'siamese cat': 358, 'lampshade': 352, 'barbershop': 349, 'baby bib': 258, 'freight car': 119, 'laptop computer': 46, 'sea slug': 37, 'shoe store': 32, 'cricket insect': 19, 'baluster handrail': 2

I.2 YFCC-100

'grand piano': 62610, 'orange': 37182, 'fly': 30889, 'lion': 16043, 'bee': 14164, 'pizza': 12084, 'barn': 11854, 'goose': 11200, 'ice cream': 10556, 'greenhouse': 9479, 'menu': 7463, 'umbrella': 7164, 'banana': 6933, 'bubble': 6838, 'corn': 6835, 'cougar': 6619, 'lemon': 6439, 'daisy': 5386, 'scuba diver': 5044, 'cricket insect': 4702, 'laptop computer': 4601, 'ant': 4465, 'hay': 4427, 'peafowl': 4140, 'pillow': 4137, 'flamingo': 3735, 'bookstore': 3668, 'necklace': 3201, 'bald eagle': 2912, 'ram adult male sheep': 2617, 'jellyfish': 2482, 'vulture': 2462, 'suspension bridge': 2439, 'espresso': 2189, 'mailbox': 2125, 'bison': 2073, 'flagpole': 2012, 'fig': 1973, 'hen': 1896, 'cucumber': 1815, 'bagel': 1746, 'koala': 1592, 'magpie': 1366, 'stone wall': 1337, 'spider web': 1296, 'acorn': 1277, 'popsicle': 1226, 'baluster handrail': 1182, 'vending machine': 1118, 'broccoli': 1114, 'junco': 1113, 'quail': 1108, 'stupa': 1043, 'feather boa': 1018, 'stingray': 971, 'macaw': 961, 'wallaby': 942, 'sea slug': 832, 'chickadee': 783, 'lifeboat': 781, 'baby bib': 774, 'mitten': 748, 'teapot': 728, 'macaque': 661, 'traffic light': 638, 'mashed potatoes': 625, 'african bush elephant': 600, 'tarantula': 593, 'barbershop': 537, 'gas pump': 520, 'padlock': 517, 'beer bottle': 433, 'warthog': 430, 'mongoose': 407, 'siamese cat': 395, 'guacamole': 393, 'parking meter': 381, 'spindle': 379, 'wine bottle': 370, 'dishwasher': 361, 'lampshade': 358, 'lhasa apso': 356, 'howler monkey': 314, 'lionfish': 296, 'shoe store': 285, 'soccer ball': 260, 'marimba': 168,

Model Name	**Val**	**V2**	**Sketch**	**R**	**A**	**AvgRob**	**EffRob**	**Label Acc.**	**DS Util.**
in100-384-res	0.877	0.805	0.322	0.356	0.176	0.415	0.473	1	1
in100-RN50x4	0.874	0.805	0.336	0.369	0.19	0.425	0.486	1	1
in100-sup	0.87	0.791	0.373	0.378	0.153	0.42375	0.487	1	1
in100-jpeg10	0.809	0.728	0.341	0.345	0.131	0.386	0.478	1	1
in100-cliplabel	0.813	0.717	0.278	0.328	0.127	0.363	0.446	0.9	1
in100-size-sbm-ttd	0.801	0.7	0.267	0.311	0.124	0.351	0.438	0.89	1
in100-sbm-ttd	0.754	0.674	0.285	0.331	0.123	0.353	0.468	0.89	0.72
in100-sbm-tags	0.723	0.636	0.251	0.297	0.109	0.323	0.447	0.87	0.58
in100-sbm-title	0.686	0.603	0.237	0.301	0.107	0.312	0.455	0.94	0.49
in100-gtcaps	0.849	0.768	0.37	0.373	0.17	0.421	0.495	1	1
in100-gtcaps-tokscramble	0.837	0.765	0.372	0.399	0.162	0.425	0.507	1	1
in100-jpeg10	0.75	0.682	0.311	0.352	0.144	0.372	0.496	1	1
in100-gtcaps-vitl14	0.715	0.617	0.164	0.205	0.116	0.276	0.385	1	1
in100-ttd	0.587	0.487	0.162	0.173	0.085	0.227	0.386	0.89	0.72
in100-ttd-tokstrip	0.585	0.475	0.145	0.19	0.081	0.223	0.381	0.89	0.72
in100-blipcap	0.405	0.351	0.138	0.165	0.07	0.181	0.447	0.61	0.28
in100-classname-only	0.236	0.218	0.122	0.1	0.05	0.123	0.521	1	1
oi100-sup-int-classbal	0.812	0.734	0.39	0.399	0.167	0.423	0.520	1	1
oi100-sup-int	0.667	0.595	0.316	0.399	0.156	0.367	0.549	1	1
oi100-cliplabel	0.631	0.553	0.273	0.343	0.134	0.326	0.516	0.9	1
oi100-submat-ttd	0.369	0.304	0.109	0.2	0.104	0.17925	0.486	0.48	0.08
oi100-gtcaps	0.694	0.644	0.35	0.423	0.177	0.399	0.574	1	1
oi100-ttd	0.26	0.22	0.065	0.121	0.066	0.118	0.454	0.53	0.11
oi100-blipcap+annotcap	0.343	0.291	0.09	0.174	0.055	0.152	0.443	0.46	0.14
oi100-blipcap	0.298	0.28	0.095	0.151	0.065	0.148	0.495	0.42	0.12
JANuS-gt+swinlabels-1.1m	0.908	0.863	0.678	0.731	0.349	0.655	0.721	N/A	N/A
JANuS-gt+submat-1.1m	0.871	0.817	0.625	0.659	0.276	0.594	0.682	N/A	N/A
JANuS-gt+ttd-1.1m	0.846	0.757	0.447	0.506	0.204	0.478	0.566	N/A	N/A
JANuS-ofa-1.1m	0.67	0.587	0.392	0.453	0.147	0.395	0.589	N/A	N/A
JANuS+yfcc15m-int-cliplabels-2.4m	0.927	0.877	0.7	0.78	0.449	0.702	0.757	N/A	N/A

’freight car’: 114, ’great white shark’: 104, ’christmas stocking’: 100, ’dungeness crab’: 97, ’french bulldog’: 94, ’bottle cap’: 85, ’bolete’: 80, ’chain link fence’: 51, ’barbell’: 23, ’english springer spaniel’: 22, ’border terrier’: 19

I.3 LAION-100

’spindle’: 68186, ’necklace’: 59079, ’orange’: 52221, ’pillow’: 41020, ’laptop computer’: 23319, ’lion’: 14246, ’lemon’: 12769, ’ram adult male sheep’: 11550, ’bubble’: 11079, ’barn’: 9909, ’pizza’: 9597, ’daisy’: 8331, ’umbrella’: 8323, ’banana’: 7690, ’corn’: 7140, ’menu’: 6788, ’cougar’: 6714, ’ice cream’: 6539, ’cricket insect’: 5696, ’peafowl’: 4662, ’espresso’: 4227, ’flamingo’: 4029, ’goose’: 3532, ’soccer ball’: 3532, ’barbershop’: 2963, ’dishwasher’: 2853, ’bald eagle’: 2678, ’fig’: 2635, ’greenhouse’: 2460, ’broccoli’: 2348, ’teapot’: 2298, ’acorn’: 2164, ’cucumber’: 2053, ’hay’: 2023, ’wine bottle’: 1824, ’scuba diver’: 1818, ’bison’: 1736, ’lampshade’: 1497, ’mitten’: 1457, ’french bulldog’: 1435, ’stone wall’: 1402, ’koala’: 1394, ’bee’: 1296, ’mailbox’: 1199, ’padlock’: 1126, ’stingray’: 1115, ’bookstore’: 1069, ’spider web’: 976, ’macaw’: 964, ’barbell’: 913, ’christmas stocking’: 887, ’traffic light’: 825, ’vending machine’: 808, ’popsicle’: 780, ’quail’: 768, ’chickadee’: 744, ’bagel’: 714, ’baluster handrail’: 713, ’jellyfish’: 706, ’bottle cap’: 648, ’beer bottle’: 603, ’flagpole’: 589, ’bell pepper’: 553, ’grand piano’: 544, ’guacamole’: 520, ’magpie’: 481, ’suspension bridge’: 477, ’african bush elephant’: 459, ’baby bib’: 451, ’wallaby’: 423, ’stupa’: 399, ’macaque’: 350, ’gas pump’: 335, ’great white shark’: 333, ’mongoose’: 308, ’junco’: 302, ’siamese cat’: 291, ’marimba’: 289, ’hen’: 272, ’tarantula’: 257, ’lifeboat’: 236, ’lionfish’: 205, ’totem pole’: 199, ’english springer spaniel’: 192, ’warthog’: 186, ’shoe store’: 166, ’border terrier’: 145, ’vulture’: 118, ’feather boa’: 116, ’lhasa apso’: 105, ’sea slug’: 90, ’howler monkey’: 85, ’fly’: 83, ’parking meter’: 54, ’freight car’: 50, ’ant’: 44, ’dungeness crab’: 36, ’chain link fence’: 33, ’bolete’: 14

J Per Class Accuracy for Subset Matching, openai classnames, sc

J.1 ImageNet-100

’macaw’: 0.81, ’barn’: 0.9, ’umbrella’: 0.85, ’lion’: 0.92, ’mitten’: 0.89, ’warthog’: 0.9, ’magpie’: 0.87, ’koala’: 0.88, ’banana’: 0.88, ’espresso’: 0.89, ’bagel’: 0.88, ’howler monkey’: 0.87, ’tarantula’: 0.87, ’orange’: 0.86, ’lemon’: 0.87, ’fig’: 0.87, ’broccoli’: 0.85, ’cucumber’: 0.84, ’ice cream’: 0.84, ’junco’: 0.83, ’goose’: 0.83, ’chickadee’: 0.83, ’teapot’: 0.82, ’daisy’: 0.82, ’french bulldog’: 0.82, ’vulture’: 0.82, ’stingray’: 0.81, ’guacamole’: 0.81, ’flamingo’: 0.81, ’lifeboat’: 0.81, ’suspension bridge’: 0.81, ’greenhouse’: 0.8, ’lhasa apso’: 0.81, ’ant’: 0.8, ’stupa’: 0.8, ’wallaby’: 0.8, ’bald eagle’: 0.8, ’lionfish’: 0.82, ’english springer spaniel’: 0.82, ’bison’: 0.82, ’barbell’: 0.82, ’macaque’: 0.82, ’mailbox’: 0.85, ’necklace’: 0.84, ’quail’: 0.84, ’padlock’: 0.85, ’hen’: 0.84, ’acorn’: 0.83, ’pillow’: 0.82, ’fly’: 0.83, ’vending machine’: 0.83, ’stone wall’: 0.83, ’bottle cap’: 0.83, ’popsicle’: 0.83, ’spider web’: 0.82, ’totem pole’: 0.82, ’pizza’: 0.82, ’spindle’: 0.81, ’bookstore’: 0.79, ’mongoose’: 0.79, ’border terrier’: 0.78, ’parking meter’: 0.77, ’marimba’: 0.77, ’flagpole’: 0.77, ’dungeness crab’: 0.78, ’peafowl’: 0.77, ’bubble’: 0.74, ’bell pepper’: 0.74, ’bee’: 0.7, ’corn’: 0.68, ’menu’: 0.68, ’great white shark’: 0.67, ’wine bottle’: 0.67, ’dishwasher’: 0.65, ’soccer ball’: 0.65, ’jellyfish’: 0.65, ’beer bottle’: 0.59, ’grand piano’: 0.56, ’bolete’: 0.55, ’gas pump’: 0.52, ’christmas stocking’: 0.51, ’hay’: 0.48, ’traffic light’: 0.45, ’scuba diver’: 0.45, ’cougar’: 0.45, ’feather boa’: 0.42, ’african bush elephant’: 0.4, ’siamese cat’: 0.35, ’lampshade’: 0.35, ’barbershop’: 0.35, ’baby bib’: 0.26, ’freight car’: 0.11, ’laptop computer’: 0.05, ’sea slug’: 0.04, ’shoe store’: 0.03, ’cricket insect’: 0.02, ’baluster handrail’: 0.0

J.2 OpenImages-100

’bee’: 0.03, ’pizza’: 0.08, ’goose’: 0.09, ’menu’: 0.23, ’lion’: 0.21, ’banana’: 0.14, ’umbrella’: 0.21, ’jellyfish’: 0.21, ’ice cream’: 0.24, ’orange’: 0.21, ’ant’: 0.2, ’koala’: 0.2, ’necklace’: 0.22, ’flamingo’: 0.24, ’vulture’: 0.25, ’fly’: 0.24, ’lemon’: 0.21, ’wine bottle’: 0.22, ’broccoli’: 0.26, ’bison’: 0.29, ’barn’: 0.27, ’bald eagle’: 0.3, ’hen’: 0.27, ’stupa’: 0.27, ’spider web’: 0.24, ’pillow’: 0.24, ’padlock’: 0.23, ’macaw’: 0.24, ’totem pole’: 0.23, ’traffic light’: 0.23, ’laptop computer’: 0.23, ’bubble’: 0.23, ’chickadee’: 0.23, ’cucumber’: 0.24, ’daisy’: 0.24, ’warthog’: 0.24, ’parking meter’: 0.24, ’teapot’: 0.24, ’junco’: 0.22, ’spindle’: 0.22, ’lionfish’: 0.22, ’bagel’: 0.22, ’cougar’: 0.22, ’french bulldog’: 0.21, ’mailbox’: 0.19, ’hay’: 0.19, ’stingray’: 0.19, ’magpie’: 0.19, ’wallaby’: 0.19, ’vending machine’: 0.19, ’macaque’: 0.19, ’greenhouse’: 0.19, ’espresso’: 0.18, ’quail’: 0.17, ’bottle cap’: 0.17, ’grand piano’: 0.16, ’acorn’: 0.15, ’siamese cat’: 0.14, ’guacamole’: 0.13, ’gas pump’: 0.13, ’mitten’: 0.12, ’bell pepper’: 0.12, ’fig’: 0.12, ’bookstore’: 0.11, ’barbershop’: 0.11, ’lifeboat’: 0.11, ’peafowl’: 0.11, ’great white shark’: 0.11, ’mongoose’: 0.11, ’suspension bridge’: 0.11, ’tarantula’: 0.11, ’marimba’: 0.11, ’dishwasher’: 0.11, ’stone wall’: 0.1, ’christmas stocking’: 0.09, ’bolete’: 0.09, ’lhasa apso’: 0.09, ’soccer ball’: 0.09, ’beer bottle’: 0.1, ’border terrier’: 0.09, ’howler monkey’: 0.09, ’lampshade’: 0.09, ’african bush elephant’: 0.05, ’scuba diver’: 0.06, ’mashed potatoes’: 0.05, ’english springer spaniel': 0.04, 'cricket insect': 0.04, 'feather boa': 0.04, 'dungeness crab': 0.05, 'shoe store': 0.04, 'freight car': 0.04, 'barbell': 0.04, 'baby bib': 0.03, 'sea slug': 0.03

K JANuS Spreadsheet Column Explanations

JANuS contains many different kinds of metadata, and the meaning of some of the column labels used may not be immediately apparent to the reader.

We do not provide explanations for metadata columns which are explained in one of the original dataset descriptions; for those, we recommend referring to the original authors of the datasets. (Deng et al., 2009; Fang et al., 2022; Schuhmann et al., 2021; Thomee et al., 2016; Kuznetsova et al., 2020)

BLIPCaption refers to captions generated by us using a BLIP captioning model. **BLIPTitle** captions are a combination of the BLIP caption and the title field of flickr captions. Li et al. (2022a)

FlickrCaption refers to captions sourced from flickr.

annot_caption refers to OpenImages captions that were authored by human image annotators. **prose_caption** combines BLIP and annotator captions, favoring the latter when available.

clip_idx are ImageNet labels chosen by a zero-shot CLIP ViT-L model from OpenAI.

idx_ labels refer to labels generated using various subset-matching strategies.

mc is multiclass, **sc** is single class, **strict** is strict. **Ours, default, openai** refer to the three different sets of class labels we experimented with throughout this paper.