CAPTION SUPERVISION ENABLES ROBUST LEARNERS: A CONTROLLED STUDY OF DISTRIBUTIONALLY ROBUST MODEL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models such as CLIP have proven unusually robust to natural distribution shifts, such as changes in lighting and object orientation. CLIP learns on unstructured data using a technique called caption supervision; interpreting human-authored text whose contents are associated with a particular image as a ground-truth label. In a carefully controlled comparison study, we show that CNNs trained on a standard cross-entropy loss also benefit from caption supervision, in some cases producing more robust models than vision language objectives on the same data. This indicates that distributional robustness is a product of a complex interaction between the loss function, the underlying dataset, and the selection of shifts for evaluation. We also release CaptionNet, a 100-class subset of ImageNet with over 50,000 new human-labeled samples; to the best of our knowledge, this is the first version of ImageNet which duplicates the original distribution's class balance and supervision properties, while also being fully captioned with its original web-scraped labels.

1 INTRODUCTION

Motivation. Real world uses of deep learning require consistent model behavior under distribution shift. *Effective robustness*, as defined by Taori et al. (2020), is a measure of the effects of data distribution shift on image classifiers. Since 2019, numerous works have studied effective robustness on ImageNet classification, quantifying the impact of various interventions (Hendrycks & Dietterich, 2019; Taori et al., 2020; Miller et al., 2021; Fang et al., 2022; Radford et al., 2021).

The majority of "standard" deep network models have been found to perform significantly worse under naturally occurring distribution shifts, such as changes in lighting or stylized renderings of object classes (Hendrycks & Dietterich, 2019; Miller et al., 2021), leading researchers to reconsider how close computer vision models had truly come to human-level accuracy, and underscoring the need for more robust models and diverse evaluation datasets. The now-popular ViT-L CLIP model by Radford et al. (2021) was the first to show natural robustness comparable to humans across a wide range of ImageNet shifts, at the cost of some base-task accuracy. Subsequent work from Jia et al. (2021); Pham et al. (2021) showed that human-level robustness is possible even as base accuracy approaches SOTA, as long as sufficient data is available for training. Given that CLIP differs from standard deep networks on several design choices (contrastive training, much larger datasets, stylized training procedures, the role of natural language captions) it is of great interest to isolate the effect of various factors. The following important question arises: *what factors contribute to distributional robustness of models?*

Our current understanding of this question is incomplete. On the one hand, Fang et al. (2022) posit that the intrinsic diversity of training image data is the main cause for the robustness gains of CLIP-like models, with other factors such as *language supervision contributing little to no robustness*. On the other hand, Santurkar et al. (2022) seemingly contradict this finding; given a sufficiently large pretraining dataset and *descriptive, low variability captions*, contrastive CLIP-like models, which they refer to as "caption-supervised" models, outperform models trained with contrastive SIMCLR-loss in the transfer learning paradigm.

So, does caption supervision enable added robustness? Interestingly enough, this question is not so simple to answer from the above existing results, primarily due to one major challenge: one cannot directly compare CE-loss and CLIP-loss models on the same dataset because of vast discrepancies in *base accuracy* (Taori et al., 2020). CLIP-loss models are often trained on massive uncurated datasets

Table 1: *Related works cataloging emergent distributional robustness in computer vision models. We catalog the contributions of each paper with with respect to certain key factors. CLIP vs CE-loss indicates whether the paper conducted controlled comparisons on the effects of VL-loss versus cross-entropy loss. Captioning strategy indicates whether the study evaluated the effects of captioning strategy on model performance. Our paper is the first to catalog the performance of CE-loss and CLIP-loss models trained and evaluated on multiple datasets at both low and high accuracies.*

Citation	CLIP-vs-CE-loss	Non-ImageNet Training	Transfer Learning	Captioning Strategy	High Acc. Models?
Recht et al. (2019)	No	No	No	No	Yes
Taori et al. (2020)	No	Yes	No	No	Yes
Radford et al. (2021)	Yes	No	Yes	No	Yes
Miller et al. (2021)	No	Yes	No	No	Yes
Fang et al. (2022)	Yes	Yes	Yes	No	No
Santurkar et al. (2022)	No	Yes	Yes	Yes	Yes
Nguyen et al. (2022)	No	Yes	No	No	No
This paper	**Yes**	**Yes**	**No**	**Yes**	**Yes**

from the web, but perform poorly when restricted to training on standard benchmark datasets; collecting massive datasets for cross-entropy models is an expensive undertaking. Therefore, it is difficult to conduct comparisons isolating the effects of dataset size, dataset composition, loss function, data filtration method, and data supervision method (Santurkar et al., 2022).

Our contributions. This paper addresses the aforementioned challenges in several important ways:

1. Following the lead of Hendrycks & Dietterich (2019), many recent works (cf. Table 1) extrapolate trends from low accuracy models to high accuracy models. Our results that when the loss function or training dataset is changed, *the validity of these extrapolations may not apply*. Given that several controlled studies are performed in the low accuracy regime, their results must be interpreted with appropriate caveats in mind.
2. To enable proper evaluation, we introduce a new dataset, CaptionNet, a 100-class image classification dataset composed of images, human-authored captions and human-annotated labels. We source these from four existing benchmark datasets, and augment with over 50,000 newly supervised creative commons samples sourced from Flickr.
3. We train cross-entropy (CE) and CLIP-loss models models on CaptionNet, and find we are able to compare robustness with existing models while achieving *high base accuracy*. In this way, we are able to isolate the effect of labelling strategy, as well as loss functions and architectures.
4. From our observations, we are able to conclude that the interaction between loss function and data filtration strategy contribute much more to robustness than was previously believed.
5. We also provide new insight on the impact of caption styles in vision-language models, showing that improved caption supervision can have a positive impact on robustness.

Related work. Nguyen et al. (2022) were an important precursor to the work we present here; their extensive experiments on different caption-supervised datasets in the low accuracy regime made it evident that controlling for the pretraining dataset was absolutely essential for understanding distributional robustness. We extend this result and show that the loss function and labeling strategy can also play a pivotal role in distributional robustness, even when the size of the dataset and the source of image data are the same. Unlike Nguyen et al. (2022), our results are also presented in a high-accuracy regime.

In addition to their work showing the importance of data to distributional robustness, Fang et al. (2022) introduced ImageNet-Captions, which added Flickr-captions to nearly 450,000 ImageNet-1000 images. We made use of this dataset when building CaptionNet, but added over 50,000 new human-supervised samples in order to rebalance the classes, as it has been shown that CE-loss models often struggle with class imbalances (Phan & Yamamoto, 2020)

Santurkar et al. (2022) focused on two contrastive learning models in a fixed-feature setting, where model weights are frozen and a linear probe is trained using task data. Our work compares the zero-shot performance of a generalist VL model to a cross-entropy model that is, in some sense, fully trained, but using an inherently noisy labeling technique. The choice of a different learning paradigm

and a different set of controls for the model allows for a novel set of insights. We also provide new insights into the effects of captioning and filtering strategies on model robustness.

2 SETUP AND INITIAL OBSERVATIONS

Labeling schemes and captioning. In order to train cross-entropy loss models on unlabeled images, we make use of a labeling heuristic we refer to as subset matching, also used in Fang et al. (2022). This deliberately simple strategy, described in detail in Section E, avoids introducing a confound of adding either a language model or an additional vision model into the comparison process.

Whenever ground-truth labels exist, we refer to these as annotated labels. When ground-truth labels do not exist, we refer to the labels as synthetic or subset-matched. We also explore multiple captioning techniques in this paper. Flickr captions refer to captions scraped from a combination of Flickr titles, tags and descriptions. Alt-text captions are captions taken from alt-text image descriptions on websites. BLIP captions refer to captions generated using the BLIP model from Li et al. (2022). Finally, annotated captions refer to human authored captions which describe an image.

Training Datasets and Distribution Shifts A distribution shift is test-time data which differs in some patterned way from the data on which the model was trained. In this paper, we focus exclusively on natural distribution shifts, since this is the area in which CLIP outperformed prior methods. A natural distribution shift is a shift in the data distribution generated by some naturally occurring articulation parameter; sunlight, changes in perspective, abstract representations of objects. Details of training datasets and distribution shifts are deferred to the Appendix in Section A and Section B.

Measures of effective robustness and their limitations We use average robustness as our main metric for evaluating robustness. We also make use of a measure called effective robustness, introduced by Taori et al. (2020) to further help position our results among existing conclusions. Further we leverage Effective Robustness Ratio (ERR) (Feuer et al., 2022), computed as average shift accuracy over base task accuracy, to compare models with similar base accuracy.

Effective robustness is described in Taori et al. (2020) as a measure of how robust a model is on a given shift, beyond what is expected from having higher accuracy on the original test set. It is generally represented as the deviation in performance with respect to a set of baseline models on ImageNet and its subsamples, or the performance of human evaluators on the shift, which is typically plotted as $y = x$ for natural distribution shifts; typically this follows a *linear* trend.

Recent works such as Fang et al. (2022) rely on effective robustness to explain performance differences in VL models. However, using this measure requires caution. 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. On the other hand, Taori et al. (2020) and Nguyen et al. (2022) showed that models trained on more (or different data) could significantly change the 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 demonstrate, in some granular detail, certain nonlinearities that appear in many of these shifts as models begin training on larger datasets and at higher accuracies. See Section C.

In Fig. 1, we demonstrate a novel phenomenon in models trained on captioned data. Even when we hold the training dataset constant, the effective robustness of a model may differ depending on the choice of loss function and labeling strategy:

1. On LAION-15m and CC-12m, a simple subset matching labeling strategy produces CE-loss models which are comparable in robustness to VL models.
2. On YFCC, the same strategy produces CE-loss models which are less robust than VL models.

These results hold at both low and high accuracies and are unaffected by the size of the dataset.

In addition, our experiments in Table 2 show that even when we hold the loss function constant, differences in labeling strategy, such as choosing a different set of target terms for subset matching, can significantly affect effective robustness.

Our findings highlight a fundamental challenge in evaluating results on distributional robustness. Researchers cannot rely on observations made on logit-transformed linear fits (as in Recht et al. (2019)) to hold both in the low- and high-accuracy regimes, unless model architecture, loss function, dataset and evaluation metrics are fixed. Ratio-based measures such as the one described in Feuer et al. (2022) also cannot be used to effectively compare models whose base accuracy differs widely.

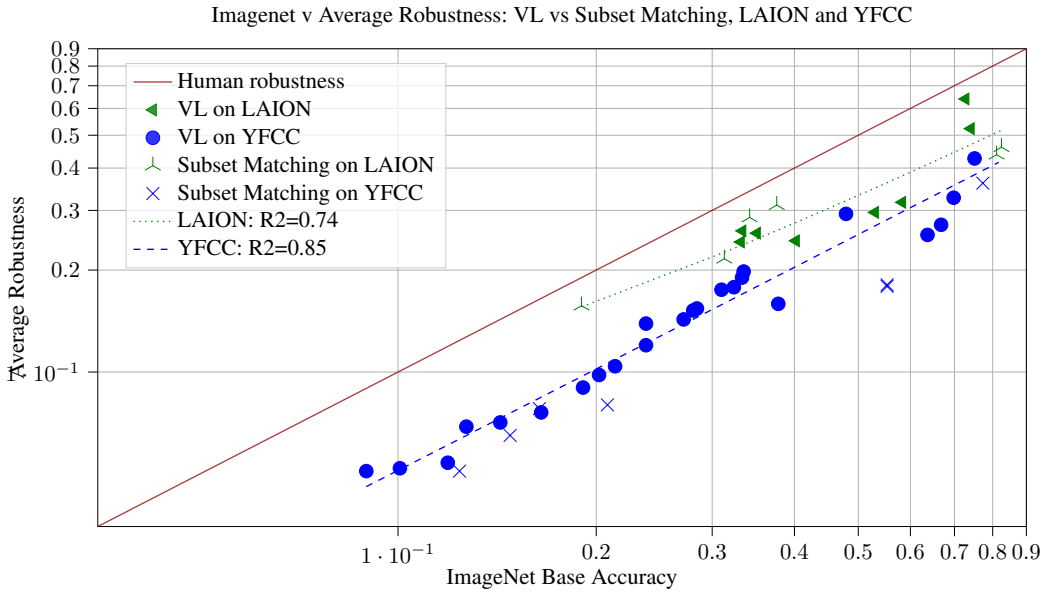

Figure 1: **Choice of loss function impacts effective robustness.** *Cross-entropy loss models fit the CLIP-loss robustness line for LAION, but not for YFCC. A ResNet-50 trained with synthetic subset-matched labels and cross-entropy loss is comparable in robustness to a ResNet-50 trained using CLIP loss. On YFCC, cross-entropy loss models underperform compared to their VL counterparts.*

Table 2: **Subset matching strategy can impact effective robustness.** *Here, we compare models with our labels to models using the default ImageNet class labels. We find that under a strict subset matching strategy, our labels perform better than default labels and a random subsample of YFCC.*

No. of samples	Dataset	Technique	Labels Used	Strategy	Val. Acc. (Imagenet)	Avg. Robust Acc.	Eff. Rob. Acc. Ratio
2m	yfcc	VL	None	Subsample	0.119	0.054	0.454
2.3m	yfcc	VL	default	Strict	0.143	0.071	0.497
2.3m	yfcc	Subset matching	default	Strict	0.124	0.051	0.411
2.2m	yfcc	VL	ours	Strict	0.165	0.076	0.461
2.2m	yfcc	Subset matching	ours	Strict	0.148	0.065	0.439

Ideally, we should be comparing models trained with the VL and CE losses in a setting where both strategies were capable of achieving high base accuracy. Unfortunately, ImageNet-1000 does not provide such a setting; even VL models trained on the largest publicly available unsupervised dataset, LAION, have yet to match the performance of CLIP, and training even one such model is enormously demanding.

3 CaptionNet

To resolve this challenge, we introduce CaptionNet, a collection of four new training datasets designed to be evaluated on a subset of ImageNet-1000. Each dataset in CaptionNet is either a subset or a superset of an existing dataset, and each is fully captioned and fully labeled, either using annotated or synthetic labels.

1. **ImageNet-100:** A superset of ImageNet with over 50,000 newly annotated samples, including flickr-captions and blip-captions.
2. **OpenImages-100:** A subset of OpenImages with new mappings from OpenImages to ImageNet classes, restored original flickr-captions, and new BLIP-captions.
3. **LAION-100:** A subset of LAION-15m with samples selected via subset matching.
4. **YFCC-100:** A subset of YFCC-15m with samples selected via subset matching.

To the best of our knowledge, CaptionNet contains the only version of ImageNet which duplicates the original distribution's class balance and supervision properties, while also being fully captioned with its original web-scraped labels. We find that both VL and CE models trained on relatively small

Figure 2: **ImageNet-100 samples from CaptionNet.**

Figure 3: **OpenImages-100 samples from CaptionNet.**

amounts of data can achieve high base accuracy on CaptionNet, making it possible for the first time to compare model robustness while controlling for base accuracy.

More information on the process used to create CaptionNet is available in Section F. Additional details on the size and composition of each CaptionNet subset can be found in Section **??**.

Experimental Setup. In order to minimize differences in model architecture, we train two families of models: A ResNet-50 for cross-entropy models, and a CLIP-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; as noted in (Radford et al., 2021; Santurkar et al., 2022), this does not affect CLIP performance. In order to control for dataset size, we train models on various subsets of CaptionNet and measure base accuracy and robustness.

CE-loss models are typically trained with full (32 bit floating point) precision with batch size of 128 and gradient clipping. VL models are typically trained with mixed precision (32-bit for gradients, and 16-bit for weights), for a batch size of 256, and do not use gradient clipping. Models are typically distributed across a single node with 4 NVIDIA GPUs; our very largest models were trained on 16 NVIDIA GPUs. We use the AMP library to implement the training process.

We train our models for 32 or 64 epochs unless otherwise specified. For CaptionNet models, we train for 256 epochs on all models, except for the models trained on YFCC-2.2Mn and YFCC-3.9Mn, which train for 64 epochs. We report the best ID/OOD accuracy score for each model, with best being determined by the model's peak performance on ImageNet-100 validation. Following Santurkar et al. (2022), we use SimCLR transformations rather than CLIP transformations for all model training on CaptionNet. In the case of ImageNet-R and ImageNet-A, which are themselves subsets of ImageNet, we evaluate on only the classes which are shared between ImageNet-100 and ImageNet-A/R, approximately 35 classes for each dataset.

3.1 Results on CaptionNet

The results discussed in this section can be found in Table 3.

Baseline performance. In order to ensure that the baseline performance of VL and CE models was essentially comparable on ImageNet-100 and ImageNet-1000 despite the newly added images, we

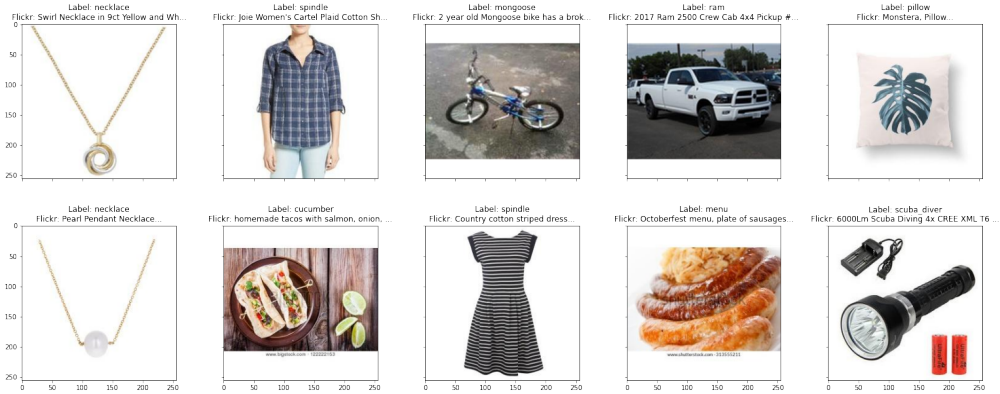

Figure 4: **LAION-100 samples from CaptionNet.**

train a VL (using "A photo of a $CLASSNAME" captions) and CE model from scratch on in100 and compare the evaluation results to those of models trained on all of ImageNet; we find that evaluation accuracy of the smaller models is very similar to their larger counterparts; in100 is a good proxy for in1000.

OpenImages-100 performance. OpenImages-100 and ImageNet-100 are both drawn from Flickr creative commons; both datasets are approximately 125,000 samples. Both have ground truth labels. There are two major differences between these datasets; OpenImages was not filtered prior to labeling and has imbalanced classes. The baseline performance of oi100 CE is considerably worse than in100 CE, almost certainly because of the class imbalance. The baseline performance of oi100 VL was, we found, caption-dependent – at least some of the time, captioning strategy makes a difference. 4.3

LAION-100 and YFCC-100 performance. In order to understand the ability of models to learn on data with noisy labels, we train laion100 and yfcc100 models. CE fails to learn anything, probably because the subset matched labels are too noisy. VL, however, learns nearly as much as it does from in100, and more than oi100, probably because laion100 and yfcc100 are larger. These dramatic differences in performance underscore the importance of the choice of loss function and labeling strategy to robustness.

Combining CaptionNet Ideally, we would like to understand how to design a dataset that offers the best of both worlds. Is this possible? We attempt to answer this question by experimenting on various CaptionNet combinations and comparing them to our baseline in100 models. We provide further analysis in Section 4.2

- **in100+oi100**: in100+oi100 is about 2x the size of in100 alone, and is fully supervised. We find that in VL, this adds a small amount of robustness; in CE, we gain considerably accuracy and robustness. From this, we conclude that CE-loss benefits much more from adding supervised data than VL-loss.

- **in100+yfcc100, in100+laion100, in100+yfcc100+laion100**: When we scale up on noisily supervised data, we increase the dataset between 4x and 10x its original size. In this paradigm, we find that VL-loss models improve fairly smoothly with data scale, whereas CE-loss models improve only on the cleaner laion-100 data.

- **in100+oi100+laion100**: What happens when we combine all of our supervised data and the cleanest source of unsupervised data? VL performance actually degrades slightly; this result is quite surprising, indicating again that scale matters more than anything else in VL. CE-loss, on the other hand, improves in robustness without losing base accuracy, and comes close to rivaling CLIP, trained on 400m samples.

Combining CaptionNet and YFCC To test the effects of scaling up on previously unseen classes, our final set of experiments on CaptionNet involved training on the union of in100 and a subset of YFCC15m selected by subset-matching on all 1000 ImageNet classes. We train on the original captions for VL models, and synthetically generated labels for CE models. In VL training, we see steady improvements to robustness and accuracy, whereas with integer-label models, we see dramatic declines in both. We analyze these results in Section 4.2

Table 3: **A direct comparison of VL and CE-loss models on CaptionNet datasets highlights the importance of the loss function.** *The best-performing VL model is marked in* ***boldface****, and the best performing CE model is in italics. Datasets with high-noise labels are harder to learn for a cross-entropy model trained from scratch, but when blended with sufficient supervised data, the resulting model is more accurate and more robust than VL for the same dataset size. VL robustness is proportionately higher in low and medium regimes, but base task accuracy improves only after large amounts of additional data are added.*

Dataset	Size	Val. Acc. (VL)	Avg. Rob. (VL)	ERR (VL)	Val. Acc. (CE)	Avg. Rob. (CE)	ERR (CE)
in100	124k	0.574	0.217	0.378	0.801	0.333	0.416
oi100	135k	0.283	0.131	0.464	0.571	0.274	0.48
yfcc100	375k	0.378	0.159	0.421	0	0	0
laion100	454k	0.402	0.244	0.607	0	0	0
in100+oi100	259k	0.575	0.251	0.437	0.834	0.395	0.474
in100+laion100	578k	0.583	0.317	0.544	0.811	0.438	0.54
in100+yfcc100	499k	0.637	0.254	0.399	0.773	0.361	0.467
in100+oi100+laion100	*713k*	0.53	0.296	0.558	*0.833*	*0.546*	*0.655*
in100+yfcc100+laion100	953k	0.62	0.322	0.519	0.817	0.467	0.572
in100+yfcc2.2m	2.4M	0.668	0.272	0.407	0.553	0.181	0.327
in100+yfcc3.9m	4.1M	0.698	0.327	0.468	0.553	0.179	0.324
yfcc15m	14.8M	0.751	0.427	0.569	N/A	N/A	N/A
laion15m	13.7M	0.741	0.523	0.706	N/A	N/A	N/A
CLIP-WiT400m	**400M**	**0.9**	**0.707**	0.786	N/A	N/A	N/A
ImageNet	1.2M	N/A	N/A	N/A	0.736	0.238	0.323

4 Understanding performance differences on CaptionNet

Our CaptionNet experiments in 3 control for base accuracy, model architecture, dataset composition and dataset size. Consequently, the most likely explanations for the differences we observe in CE and VL models are the loss function and the labeling strategy. We investigate the effects of subset matching as a labeling strategy in Section E. Here, we focus on isolating and analyzing the effects of captioning and the impact of the choice of loss function.

4.1 Label noise

Radford et al. (2021) empirically show that VL models are much more tolerant of large amounts of noise in the data distribution than cross-entropy models. We confirm these results on CaptionNet; in Table 3, we find that even when label noise exceeds 90 percent, as it does on OpenImages-100 with flickr-captions, the model is still able to achieve nearly 20 percent accuracy on ImageNet.

CaptionNet CE models can be highly sensitive to noise in dataset labels; performance degrades significantly when we train using subset matching techniques and 10 percent of the labels are noisy, even when we control for the difference in dataset size. However, not all noise is created equal; in Table 4, we see that two models with similar error rates can have very different performance, depending on what those errors are. We discuss this surprising result in greater detail in Section E.4.

Table 4: **Not all errors are created equal; CLIP labels outperform subset matched labels on ImageNet-100.** *Labels generated by a high-accuracy CLIP model perform better than subset matched labels, even though the true accuracy of each labeling method is very similar, indicating that label accuracy and match count cannot fully explain differences in model performance on a dataset. Reducing dataset utilization has a strong effect on base accuracy, similar to the effect of subsampling on ImageNet-1000.*

name	true_acc	ds_util	in-val	in-a	in-r	in-s	in-v2	avg-rob	eff-rob-rat
in100-annot-labels	1	1	0.801	0.094	0.23	0.297	0.71	0.333	0.416
in100-clip-labels	0.9	1	0.778	0.099	0.225	0.286	0.66	0.318	0.409
in100-submat-oai-ttd-sc-size	0.89	1	0.714	0.074	0.184	0.223	0.606	0.272	0.381
in100-submat-oai-ttd-sc	0.89	0.72	0.651	0.065	0.179	0.194	0.537	0.244	0.375
in100-submat-oai-tags-sc	0.87	0.58	0.608	0.065	0.147	0.168	0.506	0.222	0.365
in100-submat-oai-title-sc	0.94	0.49	0.579	0.062	0.154	0.168	0.47	0.214	0.37

4.2 Data: Quantity and Quality

Cross-entropy models are capable of learning very accurate representations with relatively little data; the best-performing subset-matched model in 3 required fewer than 1 million samples, whereas the

Table 5: **Eliminating all tokens not used in evaluation reduces evaluation accuracy.** *VL models perform worse on validation and under shift when all tokens which are not in the prompt or classnames are mapped to 0, showing that VL models learn about evaluation classes from non-evaluation classes.*

Model	in-val	Avg. Rob.	Eff. Rob. Ratio
in100-flickr-caption	0.574	0.217	0.378
in100-tok-strip	*0.476*	*0.177*	*0.372*

best performing VL model required 400 million. However, cross-entropy models are much more sensitive to noisy data than VL; in our experiments with YFCC2.2Mn and YFCC3.9Mn in 3, we see that CE performance declines as more noisy data is added to the model, whereas VL model performance steadily improves. This may be because VL offers a very large space of potential classes for matching. Consider examples like "lioness" for the class "lion", or "lamp shade" for "lampshade" – subset matching approaches miss these positive matches (unless we use heuristic methods to correct them), but a VL model would use a very similar set of tokens to encode both strings. Since we know from G that CLIP-style VL models are not sensitive to token position in the string, this could result in a correct classification where a subset-matching model would fail.

On a related note, VL models fail to benefit fully from highly accurate labels. Specifically ,we find that VL models trained with CLIP loss on ground truth labels such as "A photo of a $CLASSNAME" perform worse than either flickr-captions or blip-captions on ImageNet-100, and that VL models with short, accurate captions struggle to learn. D

VL models learn from large numbers of noisy classes In our experiments with YFCC-2.2Mn and YFCC-3.9Mn (Table 3), we find that cross-entropy model performance degrades on known classes when new noisy classes are introduced during training, even when the architecture is the same.

VL-loss models improve when new tokens are seen, even if those tokens are not used during evaluation. In 5, we experiment with training a model on ImageNet-100 in which we convert all tokens which are not in the prompt or classnames to 0. In this situation, we find that both accuracy and robustness decline substantially. We see the same effect on YFCC – when we add data from non-evaluation classes, VL performance improves. While this learning improves both zero-shot performance and performance under shift, it appears to disproportionately benefit performance under shift.

4.3 Captioning strategy can affect robustness

Table 6: **Captioning strategy can affect model 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	in-val	Avg. Rob.	Eff. Rob. Ratio
in100-blip-caption	VL	0.574	0.216	0.376
in100-flickr-caption	VL	0.574	0.217	0.378
oi100-blip+annot-caption	VL	0.283	0.131	0.464
oi100-flickr+annot-caption	VL	0.225	0.11	0.489
oi100-flickr-caption	VL	0.197	0.095	0.482

As we noted in Section 1, recent works differ in their opinion on the importance of captions. We offer new insight into these apparent differences through experiments on CaptionNet. The results can be found in 6.

1. On ImageNet-100, we found no change in performance whatsoever when using BLIP+Title captioning
2. On OpenImages-100, model performance improved around 10 percent when using annotated+BLIP+Title captioning

The effect of synthetic captions is strongest when images are labeled but unfiltered, suggesting that the machine captions act as a kind of pseudo machine-label for the VL model.

What makes a caption more robust? We cannot yet answer this question; however, in 4.3, we observe that by many common measures, robust and non-robust captions are very similar; to the

extent they are dissimilar, the measures would seem to favor the non-robust YFCC captions rather than the robust LAION captions. Measures such as length, language of origin (as estimated by a common Python language detection package) or token diversity are unlikely, then, to explain the difference in model performance.

Table 7: **Alt-text and flickr captions do not differ substantially by many measures.** *Average caption length and variance of caption length, number of unique tokens, frequency of subset matching, and frequency of encountering non-english classes are either similar or would suggest YFCC to be more robust, which it is not.*

dataset	ds-size	non-english	avg-len	stddev	unique-tok-noclean	supervision	filtering
LAION-15m	13775512	2300223	59.79	70.43	6020847	unsupervised	filtered
YFCC-15m	14825134	1367899	114.98	77.36	6747440	unsupervised	unfiltered
OpenImages-100	136175	53346	164.21	328.06	469154	supervised	unfiltered
ImageNet-100	124351	23563	195.34	274.17	355452	supervised	filtered
YFCC-100	374975	16941	181.14	249.24	480458	unsupervised	unfiltered
LAION-100	453860	69189	65.32	64.05	402230	unsupervised	filtered

5 Discussion and Recommendations

Our results help illuminate the importance of captions to robust representation learning. We conclude by highlighting some aspects that may guide future work on caption-supervised models.

1. **Losses matter.** The choice of VL or CE-loss can have a substantial effect on model robustness and base accuracy, even when dataset sizes are similar.
2. **Label noise impacts VL-model robustness.** VL-loss models tend to be more robust at low accuracy, and can learn a basic representation of a very large number of classes even when labels are less than 10 percent accurate; however, they do not benefit as much from high-accuracy labels.
3. **Length matters.** VL models learn better with long, noisy, captions than with short, accurate captions.
4. **Quality trumps descriptivity.** Descriptive caption supervision can result in better performance when caption quality is low, but has no effect on accuracy or robustness when caption quality is high.
5. **CE-models can also be robust.** CE-loss models tend to gain base accuracy quickly and learn robustness only much later; they can learn good representations with much less data, but require much stricter supervision and struggle when the number of classes grows large.
6. **CE-models are very sensitive to label noise.** CE-loss models are sensitive not just to quantity, but also quality of noise; two sets of 90 percent accurate labels can produce models which differ by over ten points in base accuracy.

Broader Impact, Provenance, and Reproducibility of Large-scale Web Datasets. As more works begin to explore training regimes which make use of unstructured data, we strongly urge that groups assembling a dataset retain all of the metadata associated with the samples.
While it is reasonable to offer filtered and preprocessed datasets for the sake of convenience, our results demonstrate that any filtration and preprocessing pipeline will significantly affect model robustness and performance, including eliminating data streams which may provide valuable information on potential correlations. Bearing this in mind, we should 'first do no harm' to the data, allowing for the possibility that future advancements in the field may unlock potential applications of datasets we had not previously anticipated.

REFERENCES

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

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.

Alexander W. Fang, Gabriel Ilharco, Mitchell Wortsman, Yu Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *ICML*, 2022.

Benjamin Feuer, Ameya Joshi, and Chinmay Hegde. A meta-analysis of distributionally-robust models, 2022. URL https://arxiv.org/abs/2206.07565.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ArXiv*, abs/1903.12261, 2019.

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.

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

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.

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*, 128(7):1956–1981, mar 2020. doi: 10.1007/s11263-020-01316-z. URL https://doi.org/10.1007%2Fs11263-020-01316-z.

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

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. *ArXiv*, abs/2107.04649, 2021.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip, 2022.

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.

Trong Huy Phan and Kazuma Yamamoto. Resolving class imbalance in object detection with weighted cross entropy losses, 2020. URL https://arxiv.org/abs/2006.01413.

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

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.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *ArXiv*, abs/1902.10811, 2019.

Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a caption worth a thousand images? a controlled study for representation learning, 2022. URL https://arxiv.org/abs/2207.07635.

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, 2021a.

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. *ArXiv*, abs/2111.02114, 2021b.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *ArXiv*, abs/2007.00644, 2020.

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.

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

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021. https://arxiv.org/abs/2109.01903.

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.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. *ArXiv*, abs/2111.07991, 2021.