Data Determines Distributional Robustness
in Contrastive Language-Image Pre-training (CLIP)

Alex Fang[1]   **Gabriel Ilharco**[1]   **Mitchell Wortsman**[1]   **Yuhao Wan**[1]
Vaishaal Shankar[2]   **Achal Dave**[2]   **Ludwig Schmidt**[1,3]

Abstract

Contrastively trained language-image models such as CLIP, ALIGN, and BASIC have demonstrated unprecedented robustness to multiple challenging natural distribution shifts. Since these language-image models differ from previous training approaches in several ways, an important question is what causes the large robustness gains. We answer this question via a systematic experimental investigation. Concretely, we study five different possible causes for the robustness gains: (i) the training set size, (ii) the training distribution, (iii) language supervision at training time, (iv) language supervision at test time, and (v) the contrastive loss function. Our experiments show that the more diverse training distribution is the main cause for the robustness gains, with the other factors contributing little to no robustness. Beyond our experimental results, we also introduce ImageNet-Captions, a version of ImageNet with original text annotations from Flickr, to enable further controlled experiments of language-image training.

1. Introduction

Large pre-trained language-image models such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and BASIC (Pham et al., 2021) have recently demonstrated unprecedented robustness on a variety of natural distribution shifts. In contrast to prior models that are trained on images with class annotations, CLIP and relatives[1] are directly trained on images and their corresponding unstructured text from the web. The resulting models achieve large robustness even on

---

[1]University of Washington [2]Amazon [3]Allen Institute for Artificial Intelligence. Correspondence to: Alex Fang <apf1@cs.washington.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

[1]Following Radford et al. (2021), we use *CLIP* as a name for the general training technique, not only their specific models.

challenging distribution shifts such as ImageNetV2 (Recht et al., 2019) and ObjectNet (Barbu et al., 2019). No prior algorithmic techniques had enhanced robustness on these datasets even after multiple years of intensive research in reliable machine learning (Djolonga et al., 2021; Taori et al., 2020). As CLIP also improves robustness on a wide range of other distribution shifts, an important question emerges: *What causes CLIP's unprecedented robustness?*

The fact that language-image models were the first to achieve large robustness gains suggests that multimodal learning on language and image data may be key to more robust image representations. However, pinpointing the exact cause of CLIP's robustness is complicated by the fact that CLIP relied on several changes to the common supervised training paradigm for image classification models. For instance, the CLIP models with highest accuracy follow the vision transformer (ViT) architecture (Dosovitskiy et al., 2020). Radford et al. (2021) already investigated model architecture and size, showing that these factors do not affect the robustness of their CLIP models. Nevertheless, there is still a long list of possible causes for CLIP's robustness:

- The large training set size (400 million images)
- The training distribution
- Language supervision at training time
- Language supervision at test time via prompts
- The contrastive loss function

Understanding the mechanism underlying CLIP's robustness is important as it may guide the way towards more reliable machine learning more broadly.

In this paper, we answer the question of CLIP's robustness via a series of controlled experiments that test the five possible causes listed above. Our main result is that CLIP's robustness is determined almost exclusively by the training distribution. Language supervision at training time does *not* make the resulting models more robust than standard supervised learning when the images in the training set are the same. Hence language supervision only has an *indirect* effect on robustness. In particular, language supervision simplifies training on a diverse distribution of images by removing the need for consistent annotation with class labels.

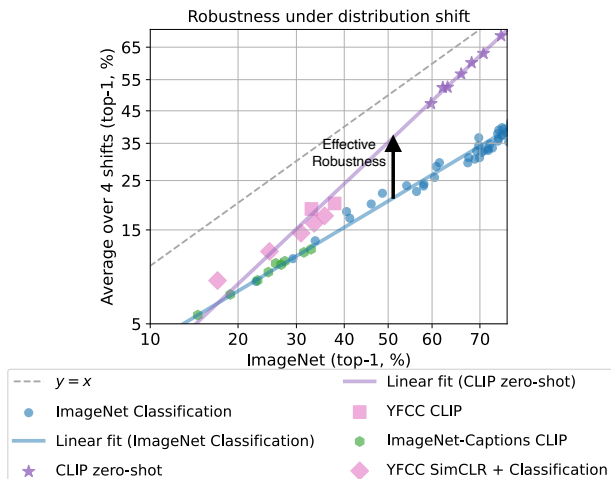The more diverse training distribution — not the language supervision — then leads to more robust representations.

Figure 1. We compare models trained using different methods and on different datasets, measuring their robustness on a range of natural distribution shifts (ImageNetV2, ImageNet-R, ImageNet-Sketch, and ObjectNet). The CLIP models stand out with their consistent performance in the presence of distribution shift. We find that large gains in effective robustness (improvement over ImageNet models) only come from varying the training distribution. Language supervision alone does not cause robustness.

Our investigation of CLIP's robustness rests on two further contributions. First, we introduce ImageNet-Captions, a new dataset for training on paired language-image data. ImageNet-Captions augments 463,622 of the 1.2 million images in the ImageNet 2012 training set (Russakovsky et al., 2015) with the original text data sourced from the corresponding Flickr images. ImageNet-Captions enables controlled experiments comparing standard (class-based) ImageNet training with language-image training on the same set of images. Such experiments precisely pinpoint the effect of utilizing language when training computer vision models.

Second, we provide a new baseline for language-image training that minimizes the interaction between the vision and language components yet achieves accuracy similar to CLIP training. Specifically, we introduce the following training procedure and illustrate its behavior on the YFCC-15M dataset (Thomee et al., 2016; Radford et al., 2021):

1. Use SimCLR (Chen et al., 2020a) to pre-train a representation on only the *images* in YFCC-15M.

2. Fine-tune the resulting representation by matching examples in YFCC-15M to ImageNet classes with simple *text matches* in the corresponding captions.

In particular, our approach relies on *no* language model,

demonstrating that it is possible to match the performance of CLIP training with much simpler language processing. Besides serving as a useful baseline to understand CLIP training, our simplified approach may open the way for further algorithmic innovations in language-image training.

The remainder of our paper proceeds as follows. The next two sections introduce relevant background and our new dataset ImageNet-Captions as experimental framework. Sections 4 and 5 then describe our main experiments testing the impact of language supervision and the training distribution on the robustness of the resulting models. Sections 6 and 7 present the evidence against test time prompts and contrastive training losses as causes for CLIP's robustness. We summarize our findings in Section 8.

2. Background

Pinpointing the cause of CLIP's robustness requires a precise experimental setup for comparing robustness across a range of models. To this end, we follow the *effective robustness* framework first proposed by Taori et al. (2020) and later utilized by Radford et al. (2021) to demonstrate the robustness gains of their CLIP models. We first review this measurement framework and then survey further related work.

2.1. Experimental setup for measuring robustness

An important goal of reliable machine learning is to design models that consistently perform well across a diverse range of test distributions. For instance, a model that achieves 75% accuracy on ImageNet should ideally also achieve 75% accuracy on the closely related ImageNetV2 distribution shift because humans can do so (Shankar et al., 2020). But instead of consistent performance, most current ImageNet models see a 12 percentage point (pp) accuracy drop on this distribution shift (Recht et al., 2019). In contrast, the CLIP models of Radford et al. (2021) are more robust and only have a 6 pp accuracy drop on ImageNetV2. Compared to earlier models, CLIP also exhibits substantially smaller accuracy drops on many other distribution shifts (Radford et al., 2021; Jia et al., 2021; Pham et al., 2021).

More formally, our experiments measure the accuracy of a model f on two test distributions D_1 and D_2, which we abbreviate as $\text{acc}_{D_1}(f)$ and $\text{acc}_{D_2}(f)$. Usually D_1 is the ImageNet (ILSVRC-2012) test set and D_2 is one of multiple out-of-distribution test sets. An ideal model would achieve close to 100% accuracy on both distributions. Since such machine models currently do not exist, we instead have to compare the robustness of models with varying accuracies across the two distributions. In these comparisons, an important confounder is that simply increasing accuracy on distribution D_1 often already results in accuracy gains on D_2

(Taori et al., 2020; Miller et al., 2021). For instance, Figure 1 shows a range of ImageNet models (blue points) in a scatter plot with ImageNet accuracy on the x-axis ($\mathrm{acc}_{D_1}(f)$) and accuracy under distribution shift on the y-axis ($\mathrm{acc}_{D_2}(f)$). The models achieve higher accuracy under distribution shift just by virtue of having higher ImageNet accuracy.

In order to address the confounder of ImageNet accuracy when evaluating robustness, Taori et al. (2020) quantified robustness as *accuracy beyond the baseline* given by ImageNet models. The authors called this quantity *effective robustness*. In Figure 1, effective robustness corresponds to the vertical lift of a model above the blue baseline given by ImageNet-trained models. Radford et al. (2021) then demonstrated that their CLIP models achieve high effective robustness (the purple line). Mathematically, we first fit a baseline function $\beta : \mathbb{R} \to \mathbb{R}$ that maps from the accuracy $\mathrm{acc}_{D_1}(f)$ of baseline models f to the corresponding $\mathrm{acc}_{D_2}(f)$. For a new model f', the effective robustness is then given by $\rho(f') = \mathrm{acc}_{D_2}(f') - \beta(\mathrm{acc}_{D_1}(f'))$. This is the main quantity we visualize in this paper to understand the robustness of CLIP models.

Similar to Taori et al. (2020) and Radford et al. (2021), we focus on *natural* distribution shifts, which arise from natural variations such as lighting, geographic location, crowdsourcing process, etc. Natural distribution shifts stand in contrast to *synthetic* distribution shifts, where an existing test set is intentionally computationally modified to reduce model accuracy (e.g., by adding Gaussian noise, blur, or adversarial perturbations). Since natural distribution shifts resemble real data, we choose the following popular distribution shifts:[2]

1. ImageNet-V2 (Recht et al., 2019): a reproduction of the ImageNet validation set with distribution shift due to changes in the crowdsourcing process.

2. ImageNet-Sketch (Wang et al., 2019): black and white sketches of ImageNet images.

3. ImageNet-R (Hendrycks et al., 2021): renditions (e.g., art, patterns, etc.) of 200 ImageNet classes.

4. ObjectNet (Barbu et al., 2019): real-world objects from ImageNet with crowd-sourced random backgrounds, rotations, and viewpoints.

5. ImageNet-A (Hendrycks et al., 2019): naturally occurring examples filtered so they are misclassified by a ResNet-50 model.

---

[2]See Appendix A for examples of the shifts. In some figures we omit ImageNet-A due to the piecewise linear response created by the adversarial filtering process. We refer the reader to Taori et al. (2020) for details.

An important property of effective robustness on these distribution shifts is that only varying the *size* of the training set (holding its *distribution* constant) does *not* influence effective robustness. In particular, Taori et al. (2020) and Miller et al. (2021) showed that randomly sub-sampling the training set changes the accuracy, but not the effective robustness of the resulting models. This rules out the training set size as a cause for CLIP's high effective robustness (the training set size is still important for the ImageNet *accuracy* of the CLIP models).

2.2. Additional related work

Language image pre-training has been an active area of research for multiple years, including initial contributions such as VirTex (Desai & Johnson, 2021), ICMLM (Sariyildiz et al., 2020), and ConVIRT (Zhang et al., 2020). Radford et al. (2021) and Jia et al. (2021) continued this line of work and trained on significantly larger datasets to achieve competitive performance on a variety of tasks, as well as obtain models with unprecedented robustness.

Related recent work also studies exactly *where* the generalization capabilities of CLIP come from. Devillers et al. (2021) investigate whether models that use multimodal information (such as text & images) have superior generalization capabilities – as measured by few-shot and linear probe performance – to models that use only one type of information (images or text). Their analysis found that for both few-shot and linear probe settings there was no consistent advantage of multimodal models over models using only a single modality. In contrast, our work studies the *robustness* of CLIP and how language specifically affects its capability to generalize out of distribution. An important difference between our experiments and those of Devillers et al. (2021) is that we control for in-distribution accuracy in our comparison between the models to separate accuracy and robustness. Furthermore, Andreassen et al. (2021) study the effect of fine-tuning on robustness. They find that effective robustness decreases almost monotonically during the fine-tuning process, pointing to the zero-shot capability of CLIP as a source of its robustness.

Since the original CLIP paper (Radford et al., 2021), there have been a series of follow up works, including ALIGN (Jia et al., 2021), BASIC (Pham et al., 2021) and LiT (Zhai et al., 2021), each of which has made contributions to improving either the robustness or base accuracy of large pre-trained image-text models. Most related to our experiments in Section 5 is LiT, which uses a pre-trained image model and fine-tunes only the text head of the language-image model to achieve high accuracy on downstream tasks. However, we note that this work differs from our contribution in that LiT still fine-tunes a language model on a dataset of 4 billion image-caption pairs to achieve its zero-shot capability,
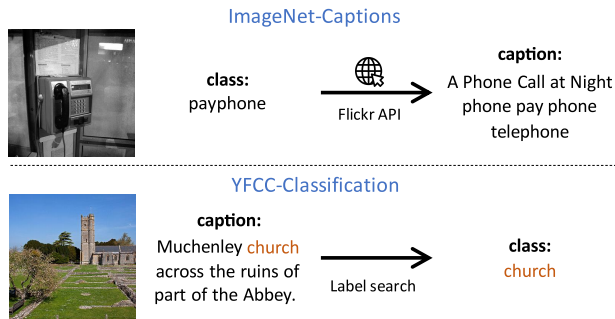
Figure 2. Overview of the two main training sets in our experiments. (Top) We introduce the ImageNet-Captions dataset, where we *augment* a subset of the ImageNet 2012 training set images with the corresponding original captions collected from Flickr. (Bottom) We convert the YFCC image-caption dataset into YFCC-Classification by searching for class labels in the YFCC captions and then *removing* the text annotations. These two datasets allows us to evaluate the impact of language-image training on robustness because we can compare language-image training with standard classification training *on the same set of images*.

while we simply convert the captions to class labels using substring matching and train a regular image classifier.

Image captioning datasets.   Existing literature offers a variety of public datasets with image-text pairs. Examples range from medium to large scale, including MS-COCO (Chen et al., 2015), SBU (Ordonez et al., 2011), Conceptual Captions 3M (Sharma et al., 2018) and 12M (Changpinyo et al., 2021), RedCaps (Desai et al., 2021), WIT (Srinivasan et al., 2021), YFCC 100M (Thomee et al., 2016) and LAION 400M (Schuhmann et al., 2021). Compared to these datasets, ImageNet-Captions contains high quality classification labels along with text associated with each image. Moreover, ImageNet-Captions is designed such that the distribution of images strongly resembles that of ImageNet, which is widely used for training and evaluating models, enabling controlled experiments such as comparisons between multiclass supervised training and image-text training.

3. ImageNet-Captions

We now describe ImageNet-Captions[3], our new dataset for experiments with image-text supervision. Four desiderata guided the creation of ImageNet-Captions:

1. To isolate the effect of natural language supervision on effective robustness, we require a dataset that contains *both* natural language supervision and traditional classification labels. This setup allows us to train clas-

---

[3]The dataset is available at https://github.com/mlfoundations/imagenet-captions.

sifiers separately with contrastive image-text losses and with standard classification losses on the *same images* and compare the resulting models. Differences in the models are then solely due to different loss functions, not architectural differences or different training distributions.

2. The text annotations in the dataset should come from the original image source, as opposed to synthetically generated captions from curated templates or an image captioning model. This helps ensure that the dataset is representative of image-text data "in the wild" and minimizes artifacts from templates or machine models.

3. The dataset should be related to commonly studied benchmarks, such as ImageNet, in order to have good baselines and comparable training methods.

4. The dataset should be large enough to support training on contemporary neural networks.

Before our paper, no dataset satisfied these constraints.

We constructed ImageNet-Captions to satisfy all four desiderata. ImageNet-Captions is a subset of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 training set, paired with the original image title, description, and tags from Flickr (recall that a large part of ImageNet was sourced from the Flickr image hosting website). Figure 3 shows three sample image-text pairs from our dataset.

3.1. Constructing ImageNet-Captions

Since ImageNet is a widely used image classification benchmark, our goal was to augment the 2012 ImageNet training set with original text data. A priori, this is a difficult task since the standard 2012 ImageNet release does not contain any metadata for the images. As a starting point, we leveraged three facts about ImageNet:

- A large fraction of ImageNet is sourced from Flickr.

- The ImageNet fall 2011 release contained URLs for each image in the full ImageNet dataset.

- For a given photo identifier, the Flickr API provides the associated text data.

Our dataset construction began with filtering the 14,197,122 image URLs in the ImageNet fall 2011 release to only include images from Flickr. In addition, we restricted the images to just the 1,000 classes included in the 2012 ImageNet competition (every entry in the fall 2011 release contains both a URL and a class label). After this filtering, we were left with 642,147 images belonging to 999 classes (all classes in ILSVRC-2012 except "teddy bear").

Title: `Reflected Duck`
Description:
Tags: `lake`, `water`, `bird` [6 tags omitted]

Title: `SILENT ROCKER`
Description: `MOSE'S MOTHER HAS LEFT THE BuILDING` [10 words omitted]
Tags: `rockingchair`, `rock`, `chair` [2 tags omitted]

Title: `A Phone Call at Night`
Description: `I might have a thing with telephones` [174 words omitted]
Tags: `phone`, `telephone`, `blackandwhite` [7 tags omitted]

Figure 3. Three sample images from ImageNet-Captions. Their respective ImageNet labels are: drake, rocking chair, payphone.

Table 1. Images from ImageNet-Captions contain three types of metadata: titles, description, and tags. For each type of metadata, this table shows the number of images that have corresponding metadata that contains the class label of the image. For most images, the class label is in at least one text field, indicating that ImageNet-Captions is suitable for language-image training.

Caption Type	# Images	% of Total
Title Only	239,495	51.6
Description Only	134,387	28.9
Tags Only	342,340	73.8
Title, Tag and Description	435,239	93.8

Next, we ran the image deduplication routine of Jain et al. (2019) to remove images that were not in the ILSVRC-2012 training set. In addition, we removed text containing profanity. This left us with a dataset of 463,622 images that are in the ILSVRC-2012 training set, along with the newly obtained corresponding text data. In particular, for each image we extracted a title (the text at the top of the Flickr image), description (the text at the bottom of the Flickr image), and user-provided tags. Since these images are a subset of ILSVRC-2012, we also have a corresponding class label that can be used for standard ImageNet training.

3.2. Properties of ImageNet-Captions

The resulting dataset contains captions from a mix of 127 different languages with the bulk (90%) coming from English. We further inspected the quality of image-text pairs by checking for the presence of the desired class label in the associated text. Table 1 summarizes the analysis. We find that for 94% of the images, the name of the ImageNet class is present in the corresponding text. This indicates that most of the captions contain relevant information about the class and are suitable for training image-text models. For additional statistics, see Appendix M.

4. ImageNet-Captions experiments

In this section, we use the ImageNet-Captions dataset to investigate the effect of language on robustness. ImageNet-Captions provides a simple comparison with vision-only methods because ImageNet is considered the premier benchmark for image classification. We train the ResNet-50 based CLIP model on ImageNet-Captions with a contrastive loss, as well as the vision encoder of that CLIP model with an additional linear layer on the equivalent image classification dataset. Training details are in Appendix B.

4.1. Caption construction

When constructing ImageNet-Captions, we had to choose which parts of the metadata to include in the caption. To do so, we ran experiments on variants that included just the title, the title followed by the description, and the title followed by the tags followed by the description. Furthermore, Radford et al. (2021) use a filter to keep only images with captions in English. We create additional variants of the dataset by applying a similar filter. As shown in Table 2, captions that include more information appear to perform better. Furthermore, it seems that filtering for cleaner captions does not make up for the loss of image-caption pairs. In caption construction ablations, images with empty captions were dropped, causing variation in dataset size across the experiments in Table 2.

4.2. Robustness

To determine the robustness of models trained on ImageNet-Captions, we evaluate on ImageNet and compare with natural distribution shifts in ImageNetV2, ImageNet-R, ImageNet Sketch, ObjectNet, and ImageNet-A. In Figure 4, we

Table 2. Evaluating different caption variants across ImageNet (IN) natural distribution shifts. Results are reported in top-1 accuracy (%). The best performing caption uses title, tags, and description. Although the language filter makes captions cleaner, the decrease in overall dataset size decreases performance.

Title	Desc	Tags	Filter	Size	Relative Size (%)	IN	IN-V2	IN-R	IN Sketch	ObjectNet	IN-A
✓			✓	197K	42.6	15.7	12.2	6.6	1.1	5.5	2.4
✓				459K	99.0	26.2	20.7	9.5	2.6	8.4	2.7
✓	✓		✓	312K	67.4	21.9	16.5	8.0	1.7	6.1	2.2
✓	✓			461K	99.4	27.8	21.6	9.6	3.0	8.0	2.7
✓	✓	✓	✓	367K	79.3	26.5	20.3	8.9	2.3	7.9	2.5
✓	✓	✓		464K	100.0	31.5	24.0	10.9	2.7	9.1	3.0

see that ImageNet-Captions CLIP models roughly follow the same linear trends as ImageNet-Captions classification models across the various distribution shifts. This shows that CLIP models are not more robust than classification models trained on the same dataset, despite the difference of language supervision. This is a better comparison than that with ImageNet classification models because there is no longer the potential confounding factor of the datasets having different image distributions. Nevertheless, these models do not achieve the robustness seen in CLIP models from Radford et al. (2021). Additional experiment details can be found in Appendix C and D.

4.3. Pre-training on language

While the above experiments show that language supervision from ImageNet-Captions does not contribute to a model's robustness, it does not rule out robustness coming from the language supervision of OpenAI's proprietary dataset used to train CLIP. Therefore we ran additional experiments where we loaded the pre-trained OpenAI CLIP model onto the language encoder, while randomly initializing the vision encoder. We trained ImageNet-Captions on this setup, with an additional variant where we also freeze the language encoder's weights. As seen in Figure 5, while both the unfrozen and frozen variants of the pre-trained language encoder increased the accuracy of the model when compared to the completely randomly initialized model, neither variant provided additional effective robustness. Detailed experiment results can be found in Appendix E.

4.4. Effect of using templates

Given that images in ImageNet-Captions have a corresponding ImageNet class, we can try to leverage this information to investigate the effect of captions and class information on both accuracy and robustness. Radford et al. (2021) introduces prompt templates in formats similar to "A photo of a {label}." Creating templates for ImageNet-Captions is different than doing so for other image-text datasets because each image already has an assigned label; for other datasets,

creating a template requires looking through the caption for classes, which are not guaranteed to be in the caption.

We found that attaching templates at the beginning of captions (followed by Title+Tags+Description) achieves 34.7% ImageNet top-1 accuracy, which is 3.2% more than without the templates. However, using the templates by themselves as the captions achieves 50.5% ImageNet top-1 accuracy, suggesting that the additional information in the captions hurts ImageNet performance. The model trained on the equivalent classification task achieves 48.7%, which suggests that with additional parameter tuning, classification may be similar to CLIP training on templates. Detailed experiment results can be found in Appendix F.

While using templates instead of captions can increase ImageNet performance, it does not improve robustness. Figure 8 in Appendix F shows that using templates on top of the captions follows linear trends similar to ImageNet-Captions. In fact, training a model on all of ImageNet using templates behaves like an equivalent classification model.

4.5. Improving ImageNet performance using captions

While language supervision does not improve robustness, it is still possible that the additional information may improve ImageNet accuracy. We investigate this by running experiments on ImageNet, augmented with ImageNet-Captions. It is well known that ResNet-50 achieves 77.15% top-1 ImageNet accuracy (He et al., 2016). As a similar baseline, we achieve 76.62% top-1 ImageNet accuracy by training the CLIP model with templates as the caption.

We have tried improving this baseline by initializing the language head with the OpenAI pre-trained model, using the combined ImageNet (templates) and ImageNet-Captions for training, using ImageNet-Captions as text augmentation when available, and contrasting the image encoding with both the template and the ImageNet-Captions caption encodings when available. However, all of these fall within ±1% of the baseline. Note that concatenating the captions to the templates changes the image distributions, while some of
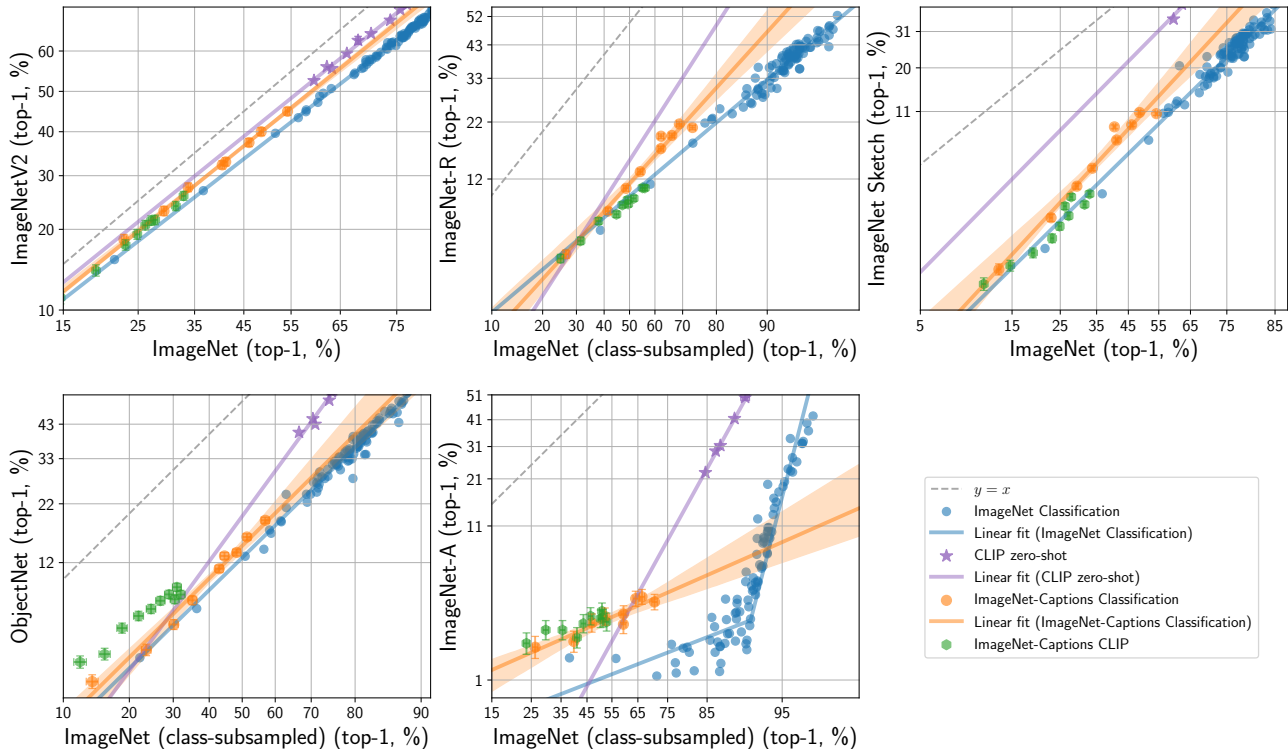
Figure 4. On most natural distribution shifts, models trained with language information from ImageNet-Captions follow the same trend as models trained without it. Neither comes close to achieving the robustness of OpenAI's CLIP models.

the other approaches do not. On the other hand, restricting the images used to those within ImageNet-Captions hints that language may help improve ImageNet performance.

Appendix G presents detailed results. The experiments we have run are non-exhaustive, and we leave it to future work to find whether language information can improve ImageNet performance, and more broadly, vision task performance.

5. YFCC experiments

Our experiments in the previous section show that language supervision alone does not improve robustness. To further understand the source of CLIP's robustness, we now investigate whether it is possible to train a representation with minimal or even no language supervision that still yields the same robustness as CLIP. These results will provide further evidence that CLIP's robustness stems from the more diverse data distribution, not the presence of language supervision.

Our experiments in this section start with a language-image training set on which CLIP exhibits improved robustness: the Yahoo Flickr Creative Commons dataset (YFCC) (Thomee et al., 2016). To test whether the image data in YFCC alone can improve robustness, we con-

trastively pre-train a "standard" image representation on YFCC that does not involve the language part of the dataset. Building on this image-only representation, we then train a zero-shot classifier with only minimal text processing (substring matches). The resulting classifier achieves effective robustness close to CLIP. This demonstrates that the training distribution, not language supervision at training time, is the main reason behind CLIP's robustness.

Dataset. In this section, we use the YFCC-15M (Radford et al., 2021) dataset, a subset of YFCC-100M (Thomee et al., 2016) filtered to only images with English titles or descriptions. The dataset contains 14,829,396 images with natural language captions associated with each image.

To train image classifiers on YFCC-15M, we convert YFCC-15M into a classification dataset with class labels for each image, which we denote YFCC-15M-Cls. We assign ImageNet labels to each image using a simple strategy: if the title or description contains the name of an ImageNet synset or synonym (Miller, 1995), we assign the corresponding synset label to the image. If an image contains no or multiple ImageNet synsets, we discard that image. This results in 1,694,125 images (11.4% of the full dataset) covering 953 ILSVRC classes. The least common class has 1 image, while the most common has 280,351 images.
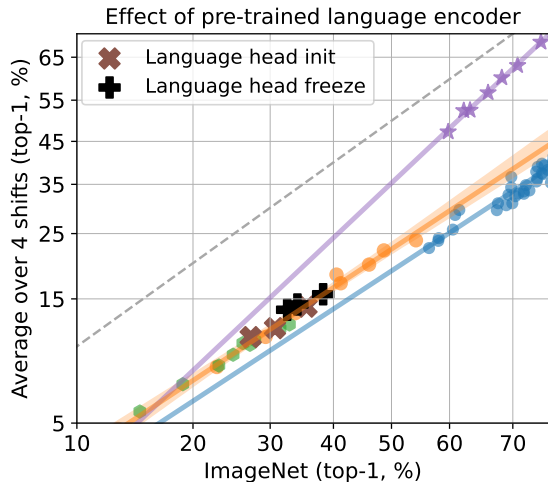
Figure 5. Using the weights from OpenAI's pre-trained CLIP model does not improve robustness, despite the large size of the full CLIP training set (400 million images). This is further evidence that language supervision does not increase robustness. See Figure 4 for remaining legend elements.

Table 3. Comparing CLIP training with (language model free) classification models on YFCC-15M. All experiments use a ViT-B/16 backbone. The CLIP results are from Mu et al. (2021). Image-only contrastive learning followed by a simple text matching stage for classification nearly matches the performance of CLIP with a full language model.

Training style	ImageNet	Avg OOD
CLIP	37.9	19.9
SimCLR → Classification	35.7	18.8

Classification training. We use a ViT-Base (ViT-B/16) model fine-tuned using the softmax cross-entropy loss on YFCC-15M-Cls. Since this data is only a fraction of YFCC-15M, we initialize the classification model with a SimCLR model pre-trained on YFCC-15M from Mu et al. (2021). Appendix J includes implementation details and ablations.

Results. We present our results in Table 3 and Figure 1. A CLIP model trained on all images and captions from YFCC-15M yields an ImageNet top-1 accuracy of 37.9%. Our baseline classification model[4], which trains SimCLR on YFCC-15M, but fine-tunes on only a small fraction (about 11%) of the supervision in YFCC-15M, results in an accuracy of 35.7%, which we found surprisingly close to CLIP. Further, as shown in Figure 1 ("YFCC SimCLR + Classification"), our baseline model's effective robustness is similar to that of CLIP.

---

[4]We call this baseline "NoCLIP", for "**No**w we use Sim-CLR+Classification instead of **C**ontrastive **L**anguage-**I**mage **P**re-training

Appendix K provides results for training on YFCC-15M-Cls from scratch. Since the training set is now about nine times smaller than YFCC-15M, the resulting models achieve much lower accuracy and are hard to compare to CLIP.

Overall, we find that despite largely eschewing language, and training on a fraction of the supervision, our baseline model results in high effective robustness, similar to CLIP. These results indicate that image-only pre-training followed by classification fine-tuning can match the robustness of CLIP, and that language pre-training is *not* necessary for effective robustness. Models trained on YFCC consistently achieve higher effective robustness than models trained on ImageNet, which shows that different training distributions have different levels of effective robustness.

6. Effect of test time prompts

As another hypothesis, we study whether natural language prompts affect CLIP's robustness. Recall that prompts consist of a template (e.g., *"a photo of ___"*) and the name of a class in the dataset. Radford et al. (2021) showed how to use multiple templates by averaging their text representations. Similarly, it is also possible to use multiple class names for each class if synonyms exist (e.g. *microwave* and *microwave oven*). To investigate the influence of specific prompts in the robustness of CLIP, we conduct a series of experiments using a trained CLIP model and multiple prompting strategies. Specifically, we vary:

- The templates used, using one of the following three options:

    i) Templates from Radford et al. (2021);

    ii) No templates (i.e., only the class names);

    iii) Random words appended before and after the class name.[5]

- The names of the classes, using one of the following three sources:

    i) Class names from Radford et al. (2021);

    ii) Class names from WordNet synset (Miller, 1995);

    iii) A combination of the previous two sources.

- The number of templates used, chosen from $\{1, 2, 4, 8, 16, 32, 80\}$.

- The maximum synonyms per class, one of $\{1, 2, 4\}$.

Figure 6 (left) shows the results from over a hundred experiments. We find that specific choices of prompts can have

---

[5]Templates are composed by one to ten random words along with the class name, in an arbitrary position. Random words are drawn using https://pypi.org/project/Random-Word/.
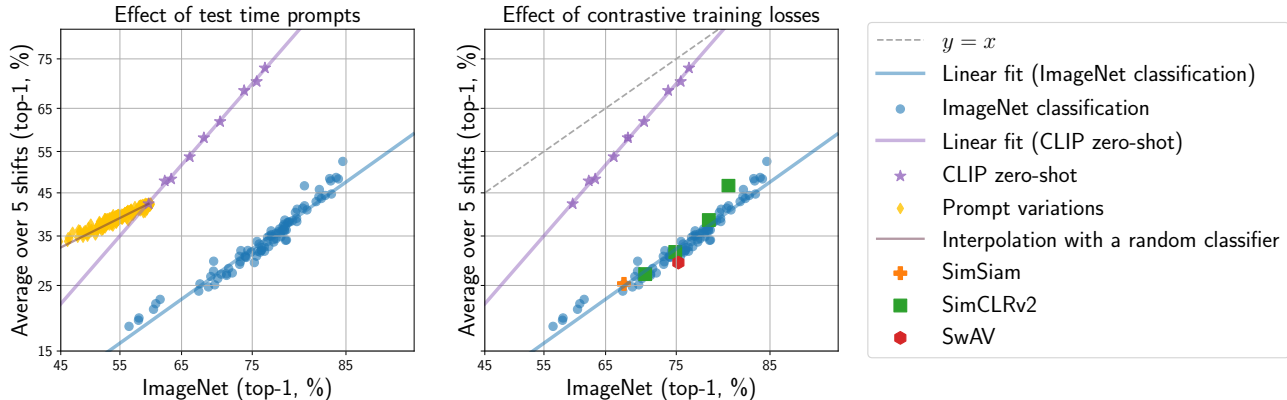
Figure 6. Effect of prompting strategies and contrastive objectives on robustness. (Left) On most natural distribution shifts, effect of prompting on effective robustness is similar to that of random interpolation. (Right) Models pre-trained with various contrastive objectives on ImageNet do not achieve the same effective robustness as CLIP models.

a substantial impact on performance. While some prompt variations did increase effective robustness, this increase is entirely due to the substantially reduced accuracy. In particular, one can achieve the same change in effective robustness by simply interpolating with a random classifier (which sees no performance change under distribution shift). We illustrate this behavior with the brown line in Figure 6 (left). Overall, these results show that prompts are not the source of robustness of CLIP models.

7. Effect of contrastive training losses

Finally, we explore contrastive pre-training as a potential source of CLIP's robustness. Contrastive pre-training is a popular method for self-supervised representation learning that encourages similar pairs to be close and dissimilar pairs to be far apart in a learned representation space. In Radford et al. (2021), the similar pairs are images and their corresponding captions. In SimCLR (Chen et al., 2020a), similar pairs are the same images with different data augmentation.

As the contrastive loss is core to CLIP's approach, we explore whether contrastive approaches independently promote effective robustness. Figure 6 (right) shows results for various popular contrastive methods, including SimCLRv2 (Chen et al., 2020b), SimSiam (Chen & He, 2020) and SwAV (Caron et al., 2020), pre-trained on ImageNet. We evaluate on ImageNet and the five distribution shifts. While the methods differ significantly from each other (e.g., different augmentation strategies, memory banks, and feature clustering techniques), they consistently exhibit little to no effective robustness. See Appendix H for further details.

8. Conclusion

The previous sections have systematically ruled out the training set size, language supervision, and the contrastive loss function as explanations for the large robustness gains achieved by the CLIP models of Radford et al. (2021). In addition, Section 5 has demonstrated that changing the training distribution from ImageNet(-Captions) to YFCC substantially affects the robustness of the resulting models. We arrive at a clear conclusion: CLIP's robustness is dominated by the choice of the training distribution, with other factors playing a small or non-existent role. While language supervision is still helpful for easily assembling training sets, it is not the primary driver for robustness.

Our paper connects the fields of robustness, learning from language & vision, and data-centric machine learning. Moreover, our results add to a growing body of evidence that the training distribution plays a central role for mitigating real-world distribution shifts. Hence, we believe that the sometimes overlooked area of dataset design offers a promising avenue for increasing the robustness of machine learning models, and we hope that the community invests more efforts into this direction.

9. Acknowledgements

References

Andreassen, A., Bahri, Y., Neyshabur, B., and Roelofs, R. The evolution of out-of-distribution robustness throughout fine-tuning. 2021. https://arxiv.org/abs/2106.15831.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://proceedings.neurips.cc/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. 2020. https://arxiv.org/abs/2006.09882.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. https://arxiv.org/abs/2102.08981.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020a. http://proceedings.mlr.press/v119/chen20j.html.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. Big self-supervised models are strong semi-supervised learners. 2020b. https://arxiv.org/abs/2006.10029.

Chen, X. and He, K. Exploring simple siamese representation learning. corr abs/2011.10566 (2020). 2020. https://arxiv.org/abs/2011.10566.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. 2015. https://arxiv.org/abs/1504.00325.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *Conference on Computer Vision and Pattern Recognition Workshops*, 2020. https://arxiv.org/abs/1909.13719.

Desai, K. and Johnson, J. Virtex: Learning visual representations from textual annotations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 11162–11173. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Desai_VirTex_Learning_Visual_Representations_From_Textual_Annotations_CVPR_2021_paper.html.

Desai, K., Kaul, G., Aysola, Z. T., and Johnson, J. Redcaps: Web-curated image-text data created by the people, for the people. 2021. https://arxiv.org/abs/2111.11431.

Devillers, B., Choksi, B., Bielawski, R., and VanRullen, R. Does language help generalization in vision models? 2021. https://arxiv.org/abs/2104.08313.

Djolonga, J., Yung, J., Tschannen, M., Romijnders, R., Beyer, L., Kolesnikov, A., Puigcerver, J., Minderer, M., D'Amour, A., Moldovan, D., Gelly, S., Houlsby, N., Zhai, X., and Lucic, M. On robustness and transferability of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. https://arxiv.org/abs/2007.08558.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. https://arxiv.org/abs/2010.11929.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. https://doi.org/10.1109/CVPR.2016.90.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. 2021. https://arxiv.org/abs/2111.06377.

Heckel, R. and Yilmaz, F. F. Early stopping in deep networks: Double descent and how to eliminate it. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tlV90jvZbw.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples.(2019). 2019. https://arxiv.org/abs/1907.07174.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. https://arxiv.org/abs/2006.16241.

Jain, T., Lennan, C., John, Z., and Tran, D. Imagededup. https://github.com/idealo/imagededup, 2019.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. https://arxiv.org/abs/2102.05918.

Miller, G. A. Wordnet: A lexical database for english. *Commun. ACM*, Nov 1995. https://doi.org/10.1145/219717.219748.

Miller, J. P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*. PMLR, 2021. https://arxiv.org/abs/2107.04649.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. 2021. https://arxiv.org/abs/2112.12750.

Ordonez, V., Kulkarni, G., and Berg, T. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. https://proceedings.neurips.cc/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

Pham, H., Dai, Z., Ghiasi, G., Liu, H., Yu, A. W., Luong, M., Tan, M., and Le, Q. V. Combined scaling for zero-shot transfer learning. *CoRR*, abs/2111.10050, 2021. https://arxiv.org/abs/2111.10050.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021. http://proceedings.mlr.press/v139/radford21a.html.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*. PMLR, 2019. https://arxiv.org/abs/1902.10811.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

Sariyildiz, M. B., Perez, J., and Larlus, D. Learning visual representations with caption annotations. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J. (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VIII*, volume 12353 of *Lecture Notes in Computer Science*, pp. 153–170. Springer, 2020. doi: 10.1007/978-3-030-58598-3_10. URL https://doi.org/10.1007/978-3-030-58598-3_10.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. 2021. https://arxiv.org/abs/2111.02114.

Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., and Schmidt, L. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*. PMLR, 2020. https://proceedings.mlr.press/v119/shankar20c.html.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

Srinivasan, K., Raman, K., Chen, J., Bendersky, M., and Najork, M. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. 2021. https://arxiv.org/abs/2103.01913.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. 2020. https://arxiv.org/abs/2007.00644.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Wang, H., Ge, S., Xing, E. P., and Lipton, Z. C. Learning robust global representations by penalizing local predictive power. 2019. https://arxiv.org/abs/1905.13549.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. Lit: Zero-shot transfer with locked-image text tuning. *CoRR*, 2021. https://arxiv.org/abs/2111.07991.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. *CoRR*, abs/2010.00747, 2020. URL https://arxiv.org/abs/2010.00747.

A. Distribution shift examples

Figure 7. Samples of the class candle from the various distribution shifts that we evaluate on in our experiments.

B. ImageNet-Captions experiments training details

CLIP experiments are trained with cross-entropy losses using AdamW optimizer with initial learning rate of 0.001 and a cosine-annealing learning rate schedule with 500 warmup steps. Hyperparameters for AdamW are set at $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon =$1e-8. The batch size is set to 1024. CLIP models trained on ImageNet-Captions are trained for 32 epochs, while ClIP models trained on all of ImageNet are trained for 90 epochs.

ImageNet-Captions classification models are trained with cross-entropy loss for 90 epochs using SGD with Nesterov momentum, setting weight decay to 0.0001, momentum to 0.9, and batch size to 256. The initial learning rate is 0.1, and is decayed by 0.1 at epochs 30, 50, and 70.

The default augmentation is random resized crop to size 224 with scale set to $(0, 9, 1.0)$, and then normalization. Additional augmentation indicates using random resized crop to size 224, random horizontal flips, and then normalization. Normalization is done with mean set to $(0.48145466, 0.4578275, 0.40821073)$, and standard deviation set to $(0.26862954, 0.26130258, 0.27577711)$.

C. ImageNet-Captions CLIP subsampled experiments

All models used here are the ResNet-50 based CLIP model used in Radford et al. (2021). Experiments are trained on a class-balanced subset of ImageNet-Captions (IN-Captions).

Experiment	IN (top-1, %)	IN-V2 (top-1, %)	IN-R (top-1, %)	IN Sketch (top-1, %)	ObjectNet (top-1, %)	IN-A (top-1, %)
IN-Captions 30%	10.8	8.6	4.7	0.8	3.8	1.8
IN-Captions 40%	14.7	11.2	5.8	1.0	4.2	2.2
IN-Captions 50%	18.9	14.2	7.4	1.2	5.7	2.2
IN-Captions 60%	23.0	17.6	8.0	1.6	6.6	2.0
IN-Captions 70%	24.9	19.2	8.9	1.9	7.1	2.5
IN-Captions 80%	27.1	21.4	9.0	2.2	7.8	2.8
IN-Captions 100%	31.5	24.0	10.9	2.7	9.1	3.0

D. ImageNet-Captions classification experiments

Unless otherwise specified, all models used here are the ResNet-50 based visual encoder used in Radford et al. (2021), with an additional linear layer at the end. Experiments are trained on a class-balanced subset of ImageNet-Captions (IN-Captions).

Experiment	IN (top-1, %)	IN-V2 (top-1, %)	IN-R (top-1, %)	IN Sketch (top-1, %)	ObjectNet (top-1, %)	IN-A (top-1, %)
IN-Captions 10%	12.9	10.0	5.0	1.0	3.0	1.7
IN-Captions 20%	22.8	18.4	8.3	2.2	4.4	1.9
IN-Captions 30%	29.3	23.1	10.8	3.6	6.0	2.5
IN-Captions 40%	33.8	27.6	13.0	4.7	7.9	2.8
IN-Captions 60%	41.2	33.0	16.7	7.2	11.2	2.9
IN-Captions 80%	46.1	37.4	19.2	9.1	13.4	3.7
IN-Captions 100%	48.7	40.0	21.6	10.8	15.8	3.8
IN-Captions 100%, Aug	54.3	45.0	20.8	10.7	18.7	3.5
ResNet-18, IN-Captions 100%	40.5	32.3	19.1	8.8	12.9	2.4

E. ImageNet-Captions language encoder experiments

Title	Desc	Tags	Filter	IN (top-1, %)	IN-V2 (top-1, %)	IN-R (top-1, %)	IN Sketch (top-1, %)	ObjectNet (top-1, %)	IN-A (top-1, %)
Language Initialized									
✓			✓	19.9	15.3	8.4	1.9	6.5	2.2
✓				27.2	21.7	10.8	2.8	8.1	3.0
✓	✓		✓	26.5	20.7	10.4	2.8	8.3	2.9
✓	✓			30.7	23.5	11.2	3.3	8.8	3.0
✓	✓	✓	✓	31.2	24.0	12.0	3.2	10.0	2.8
✓	✓	✓		35.6	28.0	13.3	4.0	10.9	3.1
Language Initialized and Frozen									
✓			✓	23.4	18.5	11.0	3.2	8.6	2.9
✓				32.6	26.0	14.1	4.3	10.4	3.0
✓	✓		✓	29.3	23.5	12.3	3.6	9.2	2.9
✓	✓			34.1	27.0	14.7	4.3	11.3	2.9
✓	✓	✓	✓	35.4	27.5	14.7	4.8	11.0	3.3
✓	✓	✓		38.3	30.2	15.5	5.0	11.6	3.1

F. ImageNet-Captions template experiments

| Experiment	IN (top-1, %)	IN-V2 (top-1, %)	IN-R (top-1, %)	IN Sketch (top-1, %)	ObjectNet (top-1, %)	IN-A (top-1, %)
Title+Tags+Description (Base)	31.5	24.0	10.9	2.7	9.1	3.0
Templates+Base	34.7	27.1	10.5	3.0	9.8	2.9
Templates+Base, Aug	42.1	32.9	11.2	3.6	10.8	2.5
Base as Templates	50.5	39.6	17.4	7.5	13.9	3.4
Base as Templates, Aug	59.0	47.6	18.3	8.4	16.1	3.2
Base as Classification	48.7	40.2	21.5	10.8	15.7	3.7
Base as Classification, Aug	54.2	45.0	20.7	10.6	18.7	3.6
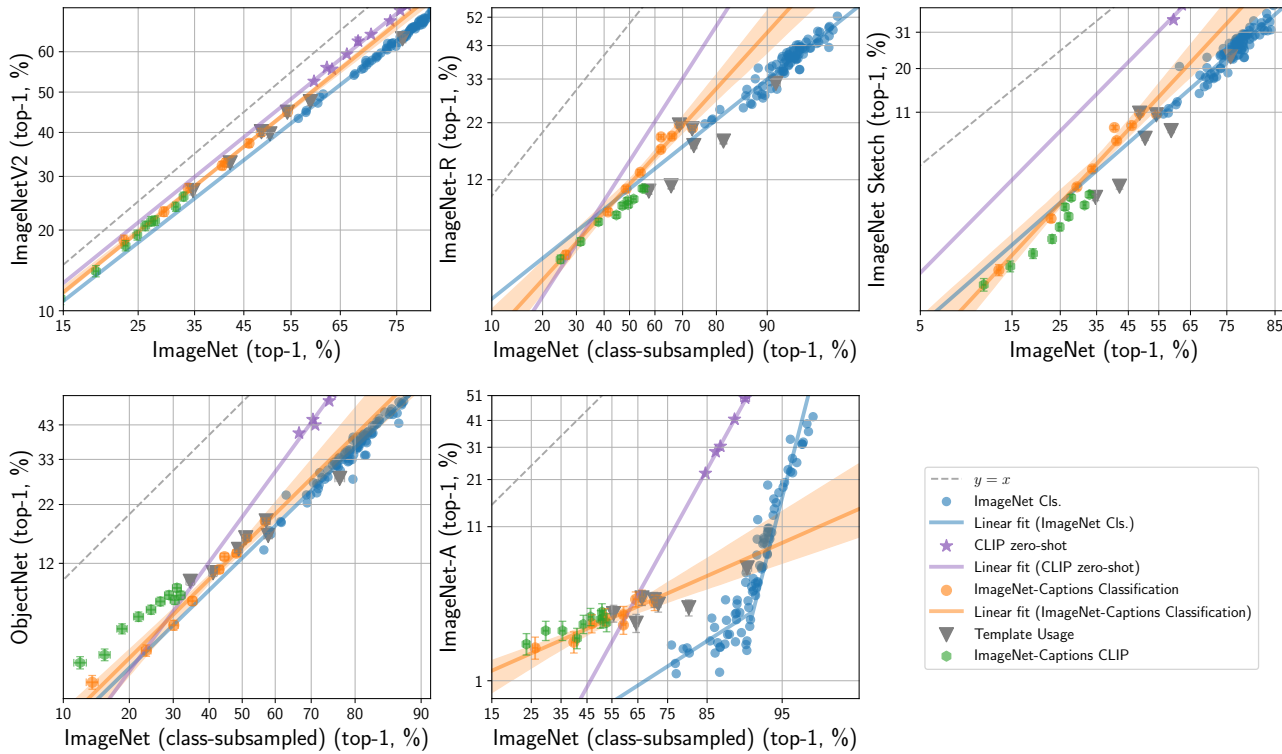
Figure 8. For the experiments that use templates in Appendix F, the templates do not increase a model's robustness.

G. ImageNet templates and captions training variation experiments

Experiments in this section use all of ImageNet with the class labels replaced with templates. Experiments that mention captions use captions for the subset of ImageNet in ImageNet-Captions.

Experiment	ImageNet (top-1, %)
ImageNet using Templates (Base)	76.6
Base, Language Initialized	76.6
Base concatenated with Captions	76.8
Base concatenated with Captions, Language Initialized	76.9
Base, Captions as Text Augmentation	76.4
Base, Contrast with Template and Captions	76.0

H. Self-supervised training variation experiments

Figure 9 supplements our analysis of self-supervised methods with the recent MAE models of (He et al., 2021). In contrast to SimCLRV2 or MoCo, MAE is not a contrastive training approach. While MAE provides more effective robustness than other approaches, it is still much less robust than CLIP. However, the source of this robustness is an open question.
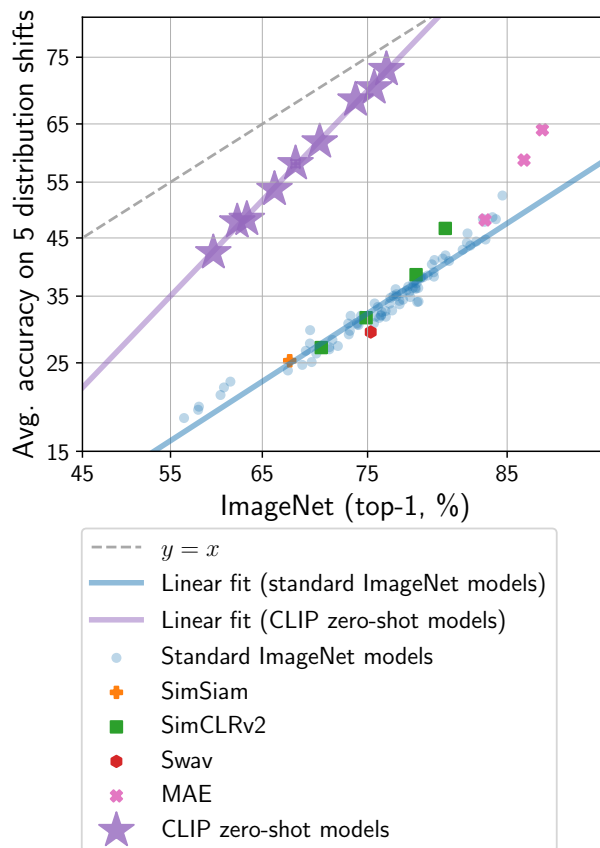

Figure 9. On most natural distribution shifts, models pre-trained on ImageNet with various contrastive objectives do not achieve effective robustness.

I. Data cleaning

The dataset is cleaned by removing samples thought to be offensive or profane. More specifically, the captions are inspected via three independent mechanisms: i) matches using a list of bad words and expressions; ii) a profanity detector model trained on human-labeled samples; iii) and human annotations, when applicable.

For the first filtering step, we use `better-profanity` library.[6] The list of words and expressions is initialized from the 835 default expressions from the library.[7] The authors manually reviewed these 835 expressions, finding 18 that could potentially be associated with ImageNet classes in non-profane captions. Captions containing any of these 18 expressions were marked for subsequent human validation, while captions that contained any of the remaining 817 expressions were automatically excluded. This step is responsible for the largest portion of filtered samples, around 14 thousand samples (approximately 3% of the data). We now list the 18 expressions (**warning, the following words might be offensive**): *breasts, cock, cocks, coon, cowgirl, dyke, nappy, nipple, nipples, organ, paddy, pot, sandbar, screw, screwed, screwing, sniper, titi*.

Data is additionally filtered using the `profanity-check` library.[8] The library detects profane or offensive language using a linear SVM model trained on 200 thousand human-labeled samples. We use a threshold of 0.95, which filters 482 samples.

Finally, remaining captions that are found to contain any of the 18 expressions listed above are manually reviewed. A total of 114 samples were found to be offensive or profane, and were removed from the dataset.

Combined, the three filtering steps filter 14,322 samples, approximately 3% of the data.

J. NoCLIP ablations

We train NoCLIP using RandAug (Cubuk et al., 2020) augmentation with N=3, magnitude=9, and magnitude std=0.5. We use a cosine annealing learning rate initialized at `1e-3` with no warmup, and a batch size of 64, with a class-balanced sampler, training for 1 epoch. We use early stopping because training for more epochs hurts performance. Heckel & Yilmaz (2021) showed that early stopping can be helpful when there is label noise. We present ablations with different hyperparameters in Table 4.

Table 4. NoCLIP ablations. "Label match" indicates whether we search for the ImageNet synset or the synset and synonyms in the YFCC captions. "Init" indicates whether we train from scratch, or using SimCLR pre-training. "Augmentation" indicates different augmentation strategies; for "RandAug" we use N=3, magnitude=9, and magnitude std=0.5. "Sampler" is either class-balanced ('Class-bal') or 'Random' (i.i.d., no balancing).

Label match	Init	Augmentation	Sampler	Epochs	ImageNet (top-1, %)
Synset	Scratch	Crop, Flip, Jitter	Class-bal	1	3.6
Synset	Scratch	Crop, Flip, Jitter	Class-bal	20	5.7
Synset	SimCLR	Crop, Flip, Jitter	Random	1	15.0
Synset	SimCLR	Crop, Flip, Jitter	Class-bal.	1	32.1
Synset	SimCLR	Crop, Flip, Jitter	Class-bal.	2	30.4
Synset	SimCLR	Crop, Flip, Jitter	Class-bal.	5	27.1
Synset + Synonyms	SimCLR	Crop, Flip, Jitter	Class-bal.	1	34.5
Synset + Synonyms	SimCLR	RandAug	Class-bal.	1	35.7

K. YFCC-15M-Cls additional experiments

In addition to the experiments that trained on top of a model pre-trained on YFCC-15M, we also train a model on YFCC-15M-Cls from scratch. Note that this model is at a much lower accuracy regime than the rest of the models we look at.
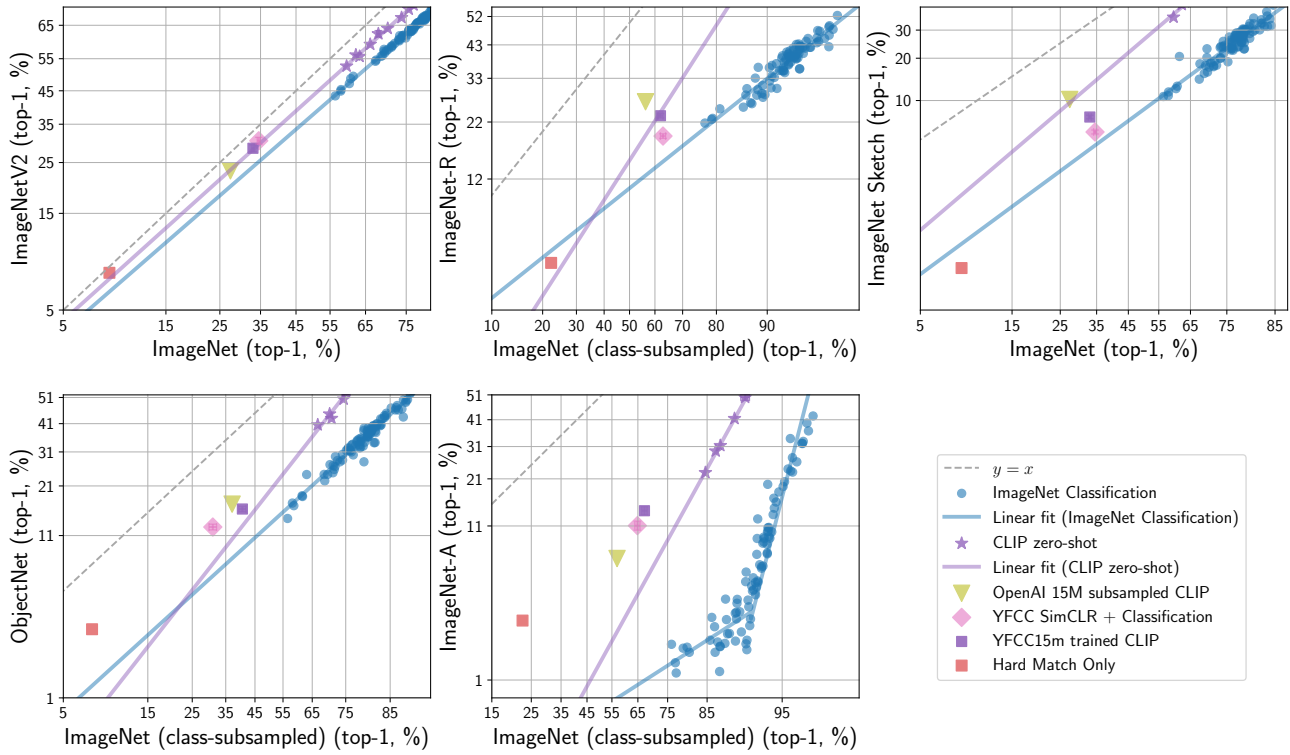
Experiment	IN (top-1, %)	IN-V2 (top-1, %)	IN-R (top-1, %)	IN Sketch (top-1, %)	ObjectNet (top-1, %)	IN-A (top-1, %)
YFCC-15M-Cls Only	8.3	7.7	4.5	0.4	2.8	2.6

---

[6] https://pypi.org/project/better-profanity/
[7] https://github.com/snguyenthanh/better_profanity/blob/master/better_profanity/profanity_wordlist.txt
[8] https://pypi.org/project/profanity-check/

L. YFCC-15M-Cls classes

There were 47 ImageNet classes that did not show up in the YFCC captions. They are the following:

tiger shark, boa constrictor, partridge, bee eater, crane bird, sea lion, toy terrier, Black and Tan Coonhound, English foxhound, Otterhound, Curly-coated Retriever, Brittany dog, Kuvasz, Groenendael dog, Greater Swiss Mountain Dog, Entlebucher Sennenhund, brussels griffon, tiger cat, tiger beetle, guinea pig, bath towel, bell tower, cassette player, cliff dwelling, construction crane, espresso machine, fountain pen, French horn, harp, one-piece bathing suit, measuring cup, missile, oxygen mask, plate rack, radio telescope, rain barrel, balaclava ski mask, slide rule, steel drum, totem pole, waffle iron, whiskey jug, window screen, Windsor tie, acorn squash, bell pepper, gyromitra

M. ImageNet-Captions additional statistics

Table 5. The most frequently occurring languages in ImageNet-Captions, according to PYCLD2 top-1. English also appears as a top-3 language in 91% of the captions.

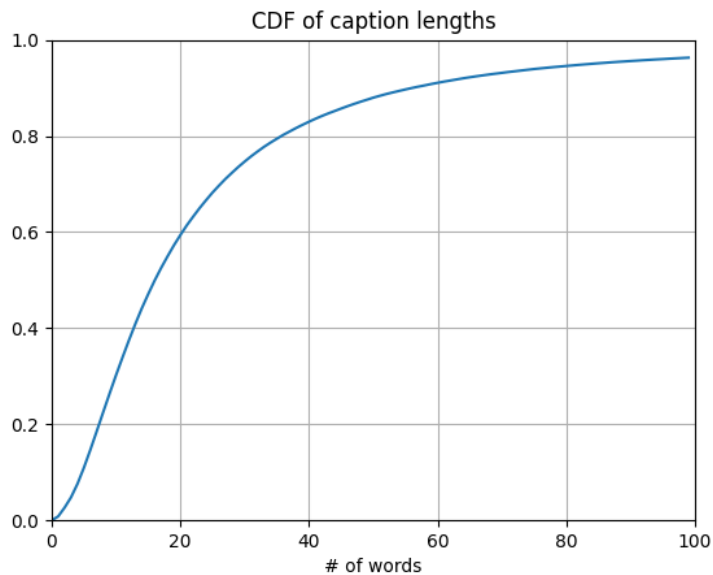| Language	# Captions	% of Total
English	416,601	89.9
Chinese	5,357	1.2
Spanish	3,893	0.8
Danish	2,993	0.6
Italian	2,598	0.6
German	2,263	0.5
Portuguese	2,104	0.5
Dutch	1,924	0.4
French	1,433	0.3
Scottish	1,404	0.3

Figure 10. The cumulative distribution function of the caption lengths (in number of words) for the title-tag-description variant of ImageNet-Captions. We limit the maximum of the x-axis to 100 words, as only 3.6% of captions are between 100 and 4,924 words. The median caption length is 17 words.