Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP

Anonymous Author(s)
Affiliation
Address
email

Abstract

Web-crawled datasets have enabled remarkable generalization capabilities in recent image-text models such as CLIP (Contrastive Language-Image pre-training) or Flamingo, but little is known about the dataset creation processes. In this work, we introduce a testbed of six publicly available data sources—YFCC, LAION, Conceptual Captions, WIT, RedCaps, Shutterstock—to investigate how pre-training distributions induce robustness in CLIP. We find that the performance of the pre-training data varies substantially across distribution shifts, with no single data source dominating. Moreover, we systematically study the interactions between these data sources and find that mixing multiple sources does not necessarily yield better models, but rather dilutes the robustness of the best individual data source. We complement our empirical findings with theoretical insights from a simple setting, where combining the training data also results in diluted robustness. In addition, our theoretical model provides a candidate explanation for the success of the CLIP-based data filtering technique recently employed in the LAION dataset. Overall our results demonstrate that simply gathering a large amount of data from the web is not the most effective way to build a pre-training dataset for robust generalization, necessitating further study into dataset design.

1 Introduction

Large models pre-trained on web-scale datasets are becoming a cornerstone of machine learning. For instance, the past two years have witnessed the arrival of several new models such as GPT-3 [8], Chinchilla [15], and PaLM [10] for natural language processing, or CLIP [22], BASIC [20], and Flamingo [4] for computer vision. These models exhibit unprecedented generalization capabilities in zero-shot inference, in-context learning, and robustness to distribution shift. A key ingredient enabling their generalization performance are the large and diverse pre-training corpora that exceed previous datasets by multiple orders of magnitude. For instance, the training set of BASIC [20] contains 6.6 billion images, which is more than 1,000 times larger than the widely used ImageNet [25] training set from the 2012 competition containing "only" 1.2 million images.

Despite the central role datasets play for pre-trained models, little is known about them, especially for image-text models. The aforementioned CLIP [22], BASIC [20], and Flamingo [4] all rely on internal datasets, which is also the case for other models such as DALL-E [23], Florence [34], and ALIGN [17]. In addition, research publications often provide little details on the data collection processes, e.g., the data sources or data filtering mechanisms. Beyond clear issues such as reproducibility and harmful content, the opaque dataset creation practices also make it hard to identify effective methods for assembling pre-training datasets. As a result, researchers cannot build on each other's dataset

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

innovations, which obstructs the incremental research process that has successfully accumulated algorithm and architecture improvements in machine learning models. A more principled understanding of datasets will likely enable further progress in the generalization capabilities of pre-trained models.

In this paper, we take a step towards a better understanding of pre-training data distribution and investigate the impact of dataset design on the generalization capabilities of image-text models. Specifically, we focus on CLIP, the first large model of this kind that demonstrated remarkable robustness to multiple challenging distribution shifts. The main questions we ask are: (i) How much do different web data sources vary in their induced robustness? (ii) Do dataset combinations lead to better robustness? (iii) Can filtering with an existing image-text model improve data quality? We answer these questions from both an experimental and theoretical perspective.

On the experimental side, we begin by assembling a corpus of six different datasets from the web, spanning a variety of sources including Flickr, Shutterstock, Wikipedia, Common Crawl, and Reddit. We then measure the robustness of CLIP models trained on each dataset. In particular, we compare the zero-shot accuracy of these models on ImageNet and a set of canonical ImageNet distribution shifts. We find that the robustness induced by each pre-training dataset varies widely, and that sources with careful curation such as Wikipedia do not necessarily outperform those with minimal filtering.

In addition, the datasets cannot be compared along a single dimension: different datasets help with robustness to different distribution shifts. This in turn motivates studying how combining datasets affects robustness. We find that a model trained on two datasets does not inherit the robustness properties of both. Rather, while the model is exposed to both distributions, its robustness interpolates between that of the individual datasets. This indicates that dataset designers must carefully combine the pre-training data in order to preserve and enhance the robustness of the resulting models.

Building on our empirical results, we introduce a theoretical model that captures the robustness phenomena in our experiments. While the model is simple enough to allow for mathematical analysis, it can still predict the behavior of real-world, web-crawled datasets. Specifically, our theoretical model also shows that combining multiple datasets dilutes the robustness of the better data distribution. Moreover, our model provides a candidate explanation for another phenomenon in the curation of pre-training data: the LAION-400M experiment [26] demonstrated that filtering a noisy source (Common Crawl) with a CLIP model results in a dataset on which newly trained CLIP models exhibit *higher* robustness to some distribution shifts.

The outline of our paper is as follows. We first briefly review related work and the relevant background on robustness to distribution shift (Section 2), before discussing our experimental setup in Section 3. Sections 4 and 5 then measure the robustness of CLIP induced by individual data sources and their combinations. To support these empirical results, Section 6 presents our theoretical analysis. We conclude with future research directions in Section 7.

2 Background & Related Work

Robustness to distribution shift is a long-standing issue in machine learning [32, 21] and has recently received renewed attention as researchers scrutinize the generalization performance of neural networks in greater detail [29, 6, 7]. Similar to CLIP, we focus on robustness to natural distribution shifts, where the corresponding test sets contain only unmodified images and are not intentionally perturbed by adversarial examples or synthetic corruptions (such as Gaussian noise or blur patterns) [13]. Specifically, we test robustness on ImageNetV2 [24], ImageNet-R [14], ImageNet-Sketch [33], and ObjectNet [5] because prior work has established many baselines for these distribution shifts [30]. Moreover, models robust to these shifts also show improved robustness on other out-of-distribution benchmarks such as WILDS [18].

The robustness literature sometimes discusses distribution shift in terms of "in-distribution" and "out-of-distribution" data. Experimenting with different pre-training datasets complicates these notions since it is now unclear what counts as "in-distribution". Instead, we follow the more flexible definition of distribution shift employed in prior work [30] and measure robustness as accuracy difference between two related but distinct test distributions (e.g. ImageNet and ImageNet-Sketch).

The expectation is that the shift between the two test distributions should not affect an ideal robust model, for instance because the shift does not affect the accuracy of humans labelers [27]. Taori et al. [30] defined this notion of robustness under the term *effective robustness*, and Radford et al. [22] adopted this effective robustness framework in their evaluations when they introduced the CLIP model. Effective robustness is illustrated in Figure 1 and measures movement towards a perfectly robust classifier which is not affected by the shift between two test distributions.

Prior work [19] has evaluated several hundred models and demonstrated experimentally that changes to model architecture, training set size, training algorithm, and other model-related factors do *not* change effective robustness in most cases. In contrast, changes to the pre-training *distribution* can improve the effective robustness of a model. This makes effective robustness a useful metric for evaluating the influence of pre-training data on robust generalization, since it removes confounders stemming from model hyperparameters and number of data samples.

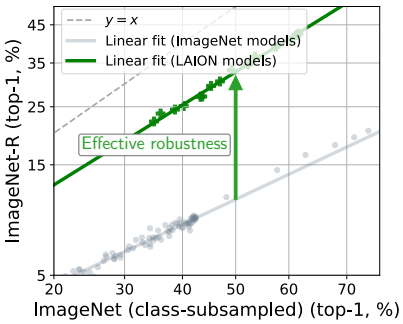

Figure 1: Models pre-trained on LAION exhibit *effective robustness* [30] compared to standard models trained on ImageNet. Effective robustness is defined as movement towards a classifier which is robust to distribution shift. A classifier which is perfectly robust to distribution shift lies on the $y = x$ line as it is not affected by the shift between distributions.

A simple baseline for dataset design is to simply train on *all* available data. However, recent work [1, 3] has introduced web-crawled datasets whose sizes exceed the compute commonly available for training. As a result, it is necessary to curate the most informative training examples from noisy sources to produce effective training datasets.

The existence of a universal linear trend for accuracies on a pair of test sets has been analytically studied in [19]. For a simple binary classification model similar to our Assumption 1, convergence to a linear trend is shown with the deviation from the line scaling as $O(1/\sqrt{d})$. However, the analysis in [19] is restricted to (i) a specific stochastic distribution shift for the out-of-distribution test data and (ii) a fixed classifier independent of the training data, and hence (iii) the dependence of the slope on the training set size, variations in the training methods, and properties of the training distribution is completely lost. We provide significantly more fine-grained analyses (Theorem 1) that captures all such tradeoffs, which allows us to draw novel insights on how data mixing (Theorem 2) and filtering (Theorem 3) affects model robustness.

3 Experiment Setup

Model. We focus on the CLIP model [22] which has demonstrated unprecedented zero-shot performance on a wide range of downstream tasks, as well as robustness to various distribution shifts [19]. Given an image-text pair, CLIP is trained to maximize the cosine similarity between the embedding of the text and that of the image, relative to the similarity of unaligned pairs. We use the CLIP implementation from *open_clip* GitHub repository [16], with ResNet-50 [12] as the image encoder architecture. We limit the total training set size to 15M samples in most of our experiments, and vary hyperparameters such as number of epochs to obtain different accuracies on each data distribution. Further training details can be found in Appendix B.

Data. To study the effects of training distributions and data availability on robust generalization at scale, we collect several datasets from publicly available sources, most of which have been studied in the context of various downstream visual and language tasks in previous work:

- YFCC: we experiment with the 15M subset of the YFCC100M dataset [31] that the original CLIP paper [22] used for dataset ablation. Images and captions are crawled from Flickr.
- LAION [26]: images and corresponding alt-texts come from web pages collected by Common Crawl [2] between 2014 and 2021. We randomly select a subset of 15M samples to experiment with, and ensure that the accompanying NSFW tags of all chosen images are marked as 'UNLIKELY'.

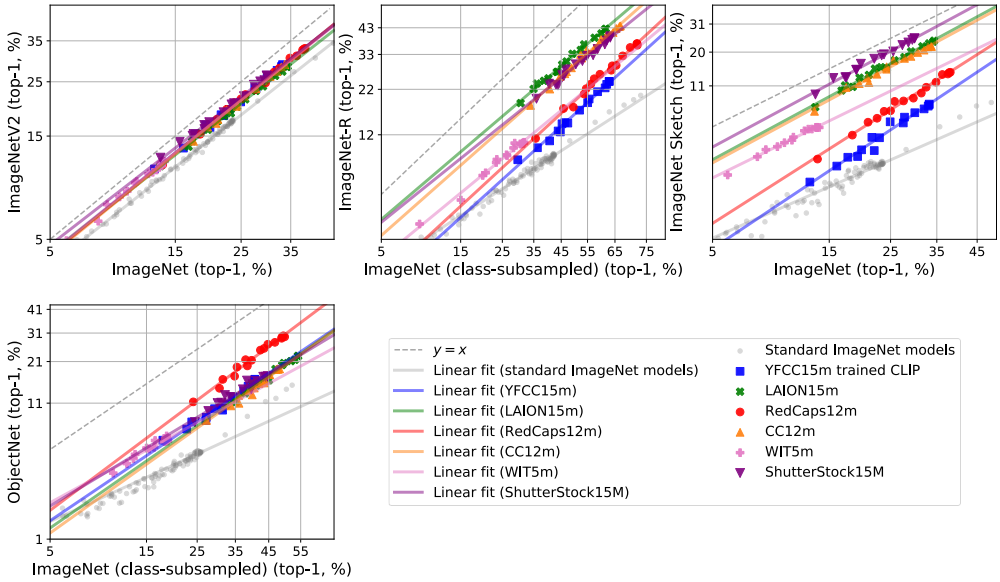

Figure 2: **Performance of the six pre-training data sources under various distribution shifts.** We find that the behavior—both in terms of accuracy and the slope of the linear trend—of the pre-training data varies substantially across distribution shifts, with no single data source dominating. Most shifts help highlight strengths and weaknesses of different data sources, except for ImageNet-V2, where the linear trends produced by individual sources are highly correlated with one another.

- Conceptual Captions [9]: we use CC-12M for our experiments, which consists of images and HTML alt-text from an unspecified set of web pages.
- RedCaps [11]: this dataset has 12M datapoints, obtained from 350 manually curated subreddit threads between 2008 and 2020. The subreddits are selected to contain high volume of image posts that are mostly photographs and not revolve around images of people.
- Shutterstock: 15M images and captions were crawled from the Shutterstock website in 2021. Refer to the website (https://www.shutterstock.com/license) for data license.
- WIT [28]: image-text pairs come from Wikipedia pages. We use reference description as the source of text data and obtain 5M datapoints in total after filtering for the English language.

Appendix A contains analysis of image and text statistics, as well as data samples of each source.

Evaluation. We pick ImageNet as the reference distribution and evaluate CLIP on four natural distribution shifts derived from ImageNet: (i) ImageNet-V2 [24]—a reproduction of ImageNet test set using the original dataset creation process, (ii) ImageNet-R [14]—renditions (e.g., sculptures, paintings) for 200 ImageNet classes, (iii) ImageNet-Sketch [33]—sketches of ImageNet class objects, and (iv) ObjectNet [5]—a test set of objects in novel backgrounds, rotations and viewpoints with 113 classes overlapping with ImageNet. Visualizations of samples randomly selected from each distribution shift can be found in Appendix A.

4 Individual pre-training Data Sources

We first investigate how well a CLIP model trained on each data source would perform under different distribution shifts of interest. As seen from Figure 2, while all sources yield the same linear trend on ImageNet-V2, some display clear advantages when evaluated on other test distributions. For example, Shutterstock offers the best out-of-distribution performance on ImageNet-Sketch, while RedCaps displays the most effective robustness on ObjectNet. On ImageNet-R, LAION, CC12M and Shutterstock seem to do much better than the rest. Overall no pre-training data distribution is consistently the most robust across all evaluation settings.

We also measure the data efficiency of each source, i.e., how much the performance would change with more samples from the same source, see Appendix C. Similar to the previous observation, the

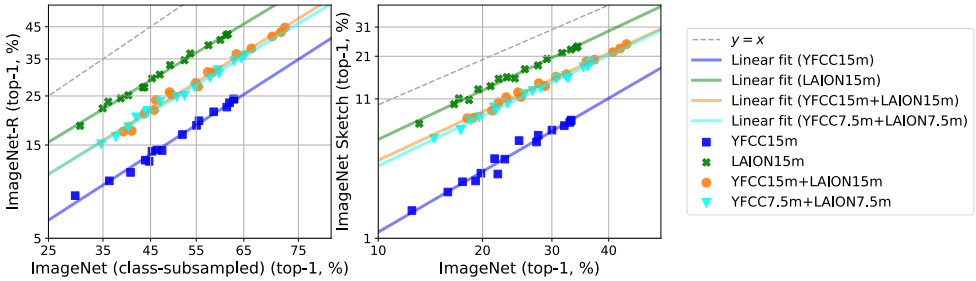

Figure 3: **Combining YFCC and LAION training data in equal ratios results in a CLIP model with intermediate robustness.** Given a fixed data budget of 15M samples, the linear trend produced by training CLIP on a YFCC-LAION data mixture, with 7.5M datapoints from each source (cyan line), lies between that of training CLIP on YFCC (blue line) or LAION (green line) entirely. Even when we increase the total training set size (30M) and use all data available from both sources (orange line), the same pattern persists.

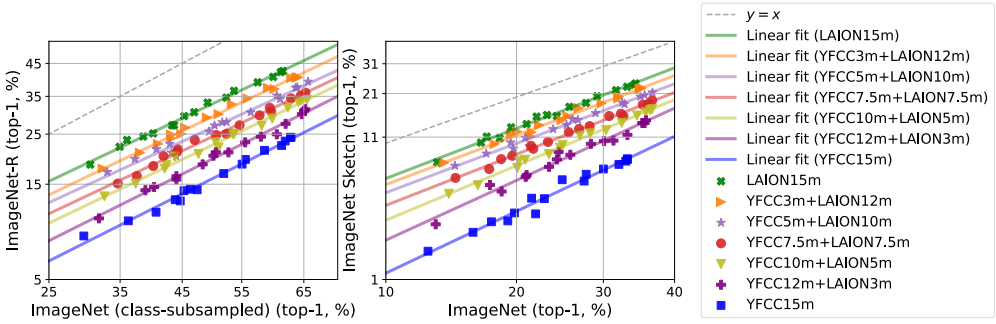

Figure 4: **Varying the sample contributions of YFCC and LAION to the input data mixture produces a smooth interpolation of the linear trend between those of training on YFCC and LAION separately.** Keeping the total number of training samples fixed at 15M, as we vary the contribution of YFCC to the dataset mixture from 15M (i.e., only training on YFCC) to 0M (i.e., only training on LAION), the resulting linear trend gradually shifts from that of YFCC-15M (blue line) to that of LAION-15M (green line).

six pre-training sources display vastly different data efficiency depending on the distribution shifts of interest. Although LAION and CC-12M behave similarly in all evaluation settings in Figure 2, this analysis reveals subtle differences between these two training distributions in the low-data regimes.

5 Combining Data Sources

In the previous section, we demonstrate the variability in behavior of different data sources based on the distribution shift at test time. A natural question then arises from this observation: does combining multiple sources help improve robustness across some, if not all, test distributions of interest? We investigate two common approaches of aggregating information from different training sets—input mixing and output ensembling. In the subsequent discussion, we focus on distribution shifts that bring out significant differences in behavior across the pre-training data sources (e.g., ImageNet-R and ImageNet-Sketch); the full results on all distribution shifts can be found in Appendices D and E.

5.1 Input Mixing

In input mixing, we combine samples from multiple data sources to build the pre-training dataset. Our findings indicate that this exposure to more training distributions doesn't help CLIP take advantage of the complimentary strengths of each source. Rather, the effective robustness of the resulting model is less than that of training on the best individual source for each distribution shift.

We start with combining data from two largest data sources in the testbed—YFCC and LAION. To remove dataset size as a confounder, we fix the total amount of training data at 15M samples, and sample 7.5M image-text pairs from each source. This data mixture yields a linear trend that lies in between the trends obtained from training on 15M YFCC and 15M LAION datapoints separately (Figure 3). Even when we remove the constraint on the training set size and use all the data available from these two sources (i.e., 30M samples), the same observation on robustness holds, and the

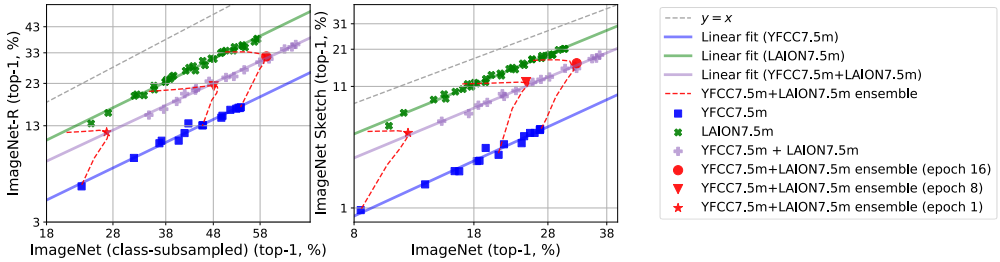

Figure 5: **Ensemble outputs of CLIP models trained on YFCC and LAION separately share the same linear trend as a single model trained on the combined data mixture (where each source contributed equally).** We ensemble the logit predictions of YFCC-trained (blue line) and LAION-trained (green line) models taken from the same epoch, with varying ensemble weights between 0 and 1 (red dashed line). When the outputs are combined with equal weights (red markers), the resulting test accuracies closely track the linear trend produced by pre-training CLIP on a data mixture with equal number of samples from each source (purple line).

resulting linear trend is highly correlated with that of training on the YFCC-7.5M + LAION-7.5M mixture.

The previous set of experiments combines YFCC and LAION data with a 50:50 ratio. In Figure 4, we find that when this ratio is varied within a fixed budget of 15M datapoints, the robustness linear trends of the corresponding mixtures form a smooth interpolation between the linear trends of training on 15M YFCC and 15M LAION samples separately. Experiments with different combinations of sources, as well as mixtures of a larger number of sources, can be found in Appendix D. There we also include results for combining data from CIFAR-10 and CINIC-10 distributions to train ResNets on image classification tasks, where we observe the same interpolation pattern. We provide a theoretical justification for this phenomenon in Section 6.2.

5.2 Output Mixing

Another common approach to take advantage of different training sets is to combine them at model output level (i.e., ensemble). Here, we train a CLIP model on each pre-training distribution of interest and combine the logit predictions of all the resulting models with equal weights. In experiments with mixing YFCC and LAION (Figure 5), the ensembles of 2 single-source-trained CLIP models taken from the same epoch of training, lie on the linear trend of a *single* CLIP model trained on the combined data. This holds across different distribution shifts that we consider. The same observation also applies when we ensemble 6 CLIP models trained separately on the 6 data sources collected, with the contribution of each model being weighted equally. Refer to Appendix E for more details.

We next show that this phenomenon extends to more complex mixtures. The fact that output mixing (i.e., ensembling the predictions of CLIPs trained on individual sources) is predictive of the linear trend produced by input mixing (i.e., training on *all* the data from these sources) presents an opportunity to estimate model robustness given new data sources, without having to train the model from scratch on the new, potentially much larger, combined dataset. For example, in Figure 6, assuming access to a CLIP model trained on the YFCC-5M + LAION-5M mixture, and another one trained on CC-5M, the ensembles of these two models across different stages of training share the same linear trend as a CLIP model trained on the YFCC-5M + LAION-5M + CC-5M mixture. In Appendix E, we show the results of ensembling 6 CLIP models trained with the 6 data sources we collected, as well as the CINIC-10 + CIFAR-10 ensemble, a uni-modal image classification setting where output mixing accuracies are also predictive of the linear trend of input mixing.

6 Analysis under Simple Binary Classification Models

Empirically we observe, e.g., in Figure 2, that for a pair of test datasets $(\mathcal{D}_1, \mathcal{D}_2)$, the corresponding test accuracies after probit transform achieved by various trained models (including different architectures and training algorithms) on the same training dataset $\mathcal{D}$ all lie on the same line. Furthermore, this line includes models trained on only a subset of $\mathcal{D}$ [19]. In other words, there is a universal line that is determined only by the training data distribution and the two test distributions, and is

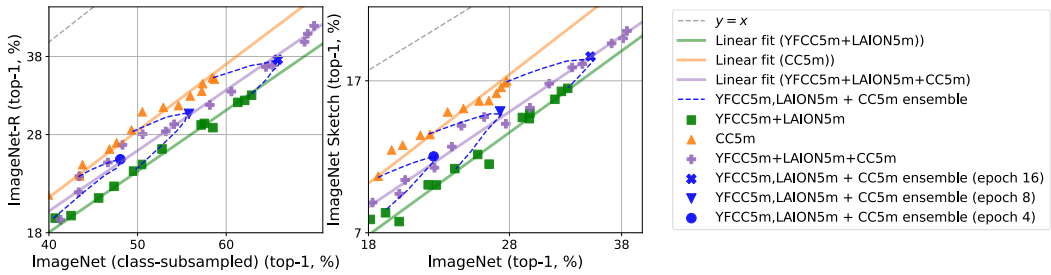

Figure 6: **Using ensemble outputs to predict the linear trend of input mixing without retraining CLIP from scratch.** A generalization of the observation made in Figure 5 is that given an existing pre-training dataset that could be a mixture (e.g., YFCC-5M + LAION-5M, green line) and a new data source (e.g., CC-5M, orange line), we could use the ensemble outputs (blue markers) of two CLIP models that have been trained separately on these two data distributions, to estimate where the linear trend for a CLIP model trained on *all* the data would lie (purple line). This removes the need to actually train CLIP from scratch on the now bigger 3-source mixture.

independent of which architecture we use, how we train the model, or how many samples we use from the training set. To explain this phenomenon, we study the following class of trained models parametrized by $(\theta \in \mathbb{R}^d, \rho \in \mathbb{R}_+)$ representing the training distribution, with n representing the training data size, and $\xi \in \mathbb{R}_+$ representing the variations due to the training algorithm.

6.1 Universality of Accuracy on the Line for Binary Classification

We analyze a simple but canonical binary classification example where each training data is parametrized by its ground-truth linear classifier $\theta \in \mathbb{R}^d$ and its Signal-to-Noise Ratio (SNR) $\rho^2 \in \mathbb{R}_+$. Concretely, a training dataset $\mathcal{D}_{n,\theta,\rho} = \{(x_i \in \mathbb{R}^d, y_i \in \pm 1)\}_{i=1}^n$ is a set of n i.i.d. paired samples from a joint distribution $P_{\theta,\rho}$ defined as follows.

Assumption 1 (Data distribution)**.** *We define a joint distribution $(x_i, y_i) \sim P_{\theta,\rho}$ as (i) $y_i = \pm 1$ uniformly at random, and (ii) $x_i = y_i\theta + (\|\theta\|/\rho)z_i$ where the noise z_i is zero-mean and has independent entries with variance one. For an observation (x_i, y_i), we refer to $\|\theta\|^2$ as the signal power and $\|\theta\|^2/\rho^2$ as the corresponding noise power with SNR ρ^2.*

Assumption 2 (Trained model distribution)**.** *We consider a (random) linear model parameter $\hat{\theta}_{n,\xi} \in \mathbb{R}^d$ that predicts a binary label $\mathrm{sign}(\langle x, \hat{\theta}_{n,\xi}\rangle)$ for a test example $x \in \mathbb{R}^d$. We assume that the random model $\hat{\theta}_{n,\xi} = \theta + (\xi\|\theta\|/(\rho\sqrt{n}))z \in \mathbb{R}^d$, trained on $\mathcal{D}_{n,\theta,\rho}$ is unbiased, $\mathbb{E}[\hat{\theta}_{n,\xi}] = \theta$, and that z has independent entries with variance one each. The randomness comes from the training data as well as any internal randomness in the training algorithm. The parameter $\xi \in \mathbb{R}_+$ captures the variations in the resulting model distribution due to changes the training algorithm.*

Concretely, one canonical example of a trained model is $\hat{\theta}_{n,\xi} = (1/n)\sum_{i=1}^n y_i x_i$. It follows that $\hat{\theta}_n = \theta + (\|\theta\|/(\rho\sqrt{n}))z$ with $\xi = 1$. Hence, ξ measures the randomness of the trained model relative to this simple training algorithm.

We analyze the resulting accuracy when evaluated on two test distributions P_{θ_1,ρ_1} and P_{θ_2,ρ_2} as defined above, with $\mathrm{Acc}_{\theta_1,\rho_1} := P_{\theta_1,\rho_1}\{\mathrm{sign}(\langle X, \hat{\theta}_{n,\xi}\rangle) = Y\}$, and $\mathrm{Acc}_{\theta_2,\rho_2}$ defined similarly. In particular, we are interested in how the accuracy pair $(\Phi^{-1}(\mathrm{Acc}_{\theta_1,\rho_1}), \Phi^{-1}(\mathrm{Acc}_{\theta_2,\rho_2}))$ behaves as we vary the sample size n and as we vary the training algorithm represented by ξ. Here, $\Phi^{-1} : [0,1] \to \mathbb{R}$ is the inverse of the CDF of a standard Gaussian distribution, defined as $\Phi(t) = \mathbb{P}(z \le t)$ where $z \sim \mathcal{N}(0,1)$, and Φ^{-1} is also called the probit function. This choice of mapping the accuracy with the probit function is critical in getting the linear relation, which we will explain in Remark 1.

Theorem 1 (Universality of accuracy on the line)**.** *Under Assumptions 1 and 2, asymptotically as the dimension d grows linearly in the sample size n such that $\lim_{d\to\infty} n/d = \alpha^2$, we have*

$$\lim_{d\to\infty} \mathrm{Acc}_{\theta_1,\rho_1} = \Phi\Big(\cos(\theta_1,\theta)\frac{\rho_1\rho\alpha}{\xi}\Big)\ , \text{ and } \lim_{d\to\infty} \mathrm{Acc}_{\theta_2,\rho_2} = \Phi\Big(\cos(\theta_2,\theta)\frac{\rho_2\rho\alpha}{\xi}\Big)\ . \tag{1}$$

Further, under Assumption 5 in the Appendix, for some universal constant $c > 0$,

$$\Phi^{-1}(\mathrm{Acc}_{\theta_2,\rho_2}) = \frac{\cos(\theta_2,\theta)\rho_2}{\cos(\theta_1,\theta)\rho_1}\Phi^{-1}(\mathrm{Acc}_{\theta_1,\rho_1}) + O\Big(\frac{e^{\frac{cn}{d}}}{\sqrt{n}}\Big)\ . \tag{2}$$

We provide a proof in Appendix F.1. This analysis implies that for any training sample size n and any (variation due to the) training algorithm ξ, the resulting accuracy pair after probit transform lies on a *universal* line determined by Eq. (2), and the slope only depends on the two test distributions $(P_{\theta_1,\rho_1}, P_{\theta_2,\rho_2})$ and the training distribution $P_{\theta,\rho}$. The similarity between the training data and each test dataset is captured by the angles: $\cos(\theta_1, \theta)$ and $\cos(\theta_2, \theta)$. More similar training data achieves a larger test accuracy. The hardness of the test distribution is captured by the SNR of each test dataset: ρ_1 and ρ_2. We emphasize that this line is universal in the sense that the slope does not depend on the sample size n and training algorithm parameter ξ. We are interested in the regime where n scales linearly in d and both are large such that the second term in the above equation is negligible.

Remark 1. Why do we get the accuracy-on-the-line phenomenon? The prediction of the trained (random) linear model on a random test data point X involves the inner product, which performs a natural spatial averaging over the d coordinates. Since the noise across the coordinates is independent and has bounded variance, the central limit theorem applies. The resulting error has a Gaussian tail. Hence, the probit mapping, $\Phi^{-1}(\mathrm{Acc}_{\theta_1\rho_1})$, is critical in translating accuracy into the relevant mean to standard deviation ratio: $\cos(\theta_1, \theta)\rho_1\rho\sqrt{n}/(\xi\sqrt{d})$. This is consequently important for getting the universal linear relation, because irrelevant parameters, n and ξ, cancel out in the slope of $\Phi^{-1}(\mathrm{Acc}_{\theta_2\rho_2})/\Phi^{-1}(\mathrm{Acc}_{\theta_1\rho_1})$. Refer to the proof of Theorem 1 (Appendix F.1) for details.

Remark 2. Intersection at the random guess: Previous work [19] that considered models with a wide range of accuracies has found that all lines cross at a point corresponding to "random guess", which in this binary example is $(\Phi^{-1}(1/2), \Phi^{-1}(1/2)) = (0, 0)$. Our analysis is consistent with this empirical observation: all universal lines intersect at $(0, 0)$ up to a small additive error scaling as $O(1/\sqrt{n})$.

6.2 Input Mixing Yields an Intermediate Slope

Figures 3 and 4 show that when a model is trained on data combined from two datasets, the resulting accuracy line is in between the lines achieved by individual datasets. We show that this trend is universally true under our current setup in Theorem 2. Note that for the linear models we consider, input mixing (combining training sets) and output mixing (combining model outputs) are equivalent.

Assumption 3. *Consider two training datasets $\mathcal{D}_{n_1,\tilde{\theta}_1,\tilde{\rho}_1}$ and $\mathcal{D}_{n_2,\tilde{\theta}_2,\tilde{\rho}_2}$ of sizes n_1 and n_2 from distributions $P_{\tilde{\theta}_1,\tilde{\rho}_1}$ and $P_{\tilde{\theta}_2,\tilde{\rho}_2}$ as defined in Assumption 1. Separately training on individual datasets gives two models $\hat{\theta}_{n_1,\xi_1}(\mathcal{D}_{n_1,\tilde{\theta}_1,\tilde{\rho}_1})$ and $\hat{\theta}_{n_2,\xi_2}(\mathcal{D}_{n_2,\tilde{\theta}_2,\tilde{\rho}_2})$ as defined in Assumption 2. A model trained on a combined (and possibly subsampled) training dataset $\mathcal{D}_{n_1',\tilde{\theta}_1,\tilde{\rho}_1} \cup \mathcal{D}_{n_2',\tilde{\theta}_2,\tilde{\rho}_2}$ is represented by $\hat{\theta}_{n_1'+n_2',\xi}(\mathcal{D}_{n_1',\tilde{\theta}_2,\tilde{\rho}_2} \cup \mathcal{D}_{n_2'\tilde{\theta}_2,\tilde{\rho}_2}) = \bar{\theta} + (\xi\|\bar{\theta}\|/(\bar{\rho}\sqrt{n_1'+n_2'}))z$ where $\bar{\theta} = (n_1'\tilde{\theta}_1 + n_2'\tilde{\theta}_2)/(n_1' + n_2')$, $\bar{\rho} = \|\bar{\theta}\|/\sqrt{(n_1'\|\tilde{\theta}_1\|^2/\rho_1^2) + (n_2'\|\tilde{\theta}_2\|^2/\rho_2^2)}$, and $z \in \mathbb{R}^d$ is a zero-mean random vector with independent entries each with variance one.*

Again, a canonical example of a trained model is $\hat{\theta}_{n_1'+n_2',\xi} = (1/(n_1'+n_2'))\big(\sum_{(x_i,y_i)\in\mathcal{D}_{n_1',\tilde{\theta}_1,\tilde{\rho}_1}} y_i x_i + \sum_{(x_i,y_i)\in\mathcal{D}_{n_2',\tilde{\theta}_2,\tilde{\rho}_2}} y_i x_i\big)$. We then have $\hat{\theta}_{n_1'+n_2',\xi} = \bar{\theta} + (\|\bar{\theta}\|/(\bar{\rho}\sqrt{n_1'+n_2'}))z$ with $\xi = 1$.
It follows from applying Theorem 1 to the model trained on the combined data (Assumption 3) that the resulting accuracy line achieves

$$\mathrm{Slope}(\hat{\theta}_{n_1'+n_2',\xi}(\mathcal{D}_{n_1',\tilde{\theta}_1,\tilde{\rho}_1} \cup \mathcal{D}_{n_2',\tilde{\theta}_2,\tilde{\rho}_2})) \quad := \quad \frac{\cos(\theta_2, \bar{\theta})\rho_2}{\cos(\theta_1, \bar{\theta})\rho_1}\,, \tag{3}$$

We show that the slope of the combined dataset lies between the two individual lines. The proof could be found in Appendix F.2. Let $\mathrm{Slope}_1 := \mathrm{Slope}(\hat{\theta}_{n_1}(\mathcal{D}_{n_1,\tilde{\theta}_1,\tilde{\rho}_1}))$ and Slope_2 be defined similarly.

Theorem 2. *Under Assumption 3,* $\mathrm{Slope}_1 \le \mathrm{Slope}\big(\hat{\theta}_{n_1'+n_2'}(\mathcal{D}_{n_1',\tilde{\theta}_1,\tilde{\rho}_1} \cup \mathcal{D}_{n_2',\tilde{\theta}_2,\tilde{\rho}_2})\big) \le \mathrm{Slope}_2$.

6.3 Filtering Data to Improve Robustness

The input mixing analysis in Section 6.2 suggests a filtering strategy to improve robustness.

Assumption 4 (Gaussian distributions for filtering)**.** *Consider a training dataset $\mathcal{D}_{n,\theta_{\mathrm{train}},\rho}$ of size n draw i.i.d. from $(X, Y) \sim P_{\theta_{\mathrm{train}},\rho}$ where $Y = \pm 1$ uniformly at random and $X|Y \sim$*

$\mathcal{N}(Y\theta_{\text{train}}, (\|\theta_{\text{train}}\|/\rho)^2\mathbf{I})$. *A model trained on an unfiltered* $\mathcal{D}_{n,\theta_{\text{train}},\rho}$ *is denoted by* $\hat{\theta}_{\text{unfiltered}} = (1/n)\sum_{(x_i,y_i)\in\mathcal{D}_{n,\theta_{\text{train}},\rho}} x_i y_i$. *Note that* $\hat{\theta}_{\text{unfiltered}} \sim \mathcal{N}(\theta_{\text{train}}, (\|\theta_{\text{train}}\|/(\rho\sqrt{n}))^2\mathbf{I})$.

This is evaluated on two test datasets: in-distribution (ID) and out-of-distribution (OOD), each with an isotropic Gaussian noise: $(X, Y) \sim P_{\theta_{\text{ID}},\rho_{\text{ID}}}$, *where* $Y = \pm 1$ *uniformly at random and* $X|Y \sim \mathcal{N}(Y\theta_{\text{ID}}, (\|\theta_{\text{ID}}\|/\rho_{\text{ID}})^2\mathbf{I})$, *and* $P_{\theta_{\text{OOD}},\rho_{\text{OOD}}}$ *is defined similarly. We let* $\text{Slope}(\theta) := \frac{\cos(\theta_{\text{OOD}},\theta)\rho_{\text{OOD}}}{\cos(\theta_{\text{ID}},\theta)\rho_{\text{ID}}}$ *and assume that all models achieve better ID accuracy than OOD accuracy, i.e.,* $\text{Slope}(\theta)$ *is smaller than one. A model is more robust if the slope is closer to one.*

Suppose we have access to a pre-trained model $\hat{\theta}_{\text{pretrained}}$ that achieves better robustness than the model $\hat{\theta}_{\text{unfiltered}}$ trained on the unfiltered data, i.e., $\text{Slope}(\hat{\theta}_{\text{unfiltered}}) < \text{Slope}(\hat{\theta}_{\text{pretrained}}) \leq 1$. We want to use the pre-trained model to filter the data $\mathcal{D}_{n,\theta_{\text{train}},\rho}$ and achieve better robustness. We consider a generic family of filtering strategies that subsamples each data point $(x_i, y_i) \in \mathcal{D}_{n,\theta_{\text{train}},\rho}$ with probability that is monotonically non-decreasing with its correlation to the pre-trained model, i.e., $\mathbb{P}((x_i, y_i)$ passes the filter$) \propto h(\langle x_i y_i, \hat{\theta}_{\text{pretrained}}\rangle)$ for some monotonic scalar function $h : \mathbb{R} \to \mathbb{R}_+$. We analyze the model $\hat{\theta}_{\text{filtered}} = (1/m)\sum_{(x_i,y_i)\in\text{filtered dataset}} x_i y_i$. The next theorem shows that any such filtering strategy improves robustness, and the full proof can be found in Appendix F.3.

Theorem 3. *Under Assumption 4, if* $\text{Slope}(\hat{\theta}_{\text{unfiltered}}) < \text{Slope}(\hat{\theta}_{\text{pretrained}}) \leq 1$ *then any model* $\hat{\theta}_{\text{filtered}}$ *that is trained on the above family of filtered datasets achieves better robustness than the model trained on unfiltered data:* $\text{Slope}(\hat{\theta}_{\text{unfiltered}}) < \text{Slope}(\mathbb{E}[\hat{\theta}_{\text{filtered}}]) \leq \text{Slope}(\hat{\theta}_{\text{pretrained}})$, *where* $\text{Slope}(\mathbb{E}[\hat{\theta}_{\text{filtered}}])$ *denotes the slope of the linear trend of a model trained on the filtered data.*

Remark 3. Relevance to empirical practice: The data filtering approach currently employed for the LAION dataset [26] motivates our theoretical setup. The authors of LAION first retrieved a large number of image-text pairs from Common Crawl [2] and then computed the cosine similarity between the image and text embeddings using one of the CLIP models originally introduced by Radford et al. [22]. Next, the authors removed all pairs with similarity less than 0.3 from the dataset. While it was initially unclear if filtering the noisy Common Crawl source with an existing CLIP model would lead to a dataset with good generalization properties, experiments with medium-scale models on LAION so far demonstrate that the resulting models are comparable to the original CLIP models [3].

Our theorem 3 does not provide a full justification for the filtering process described above, but it offers theoretical evidence that filtering a noisy data source with an existing robust model can indeed improve the generalization properties of the resulting dataset. It remains to be shown empirically that the slope of the linear trend exhibited by the original CLIP model [22] is larger than that of a model trained on the Common Crawl data pool without filtering. If this assumption holds, we can expect that training on data with better text-image alignment (according to the original CLIP model) will improve robustness compared to training on unfiltered data from Common Crawl.

7 Conclusion

Motivated by CLIP's robustness to distribution shift, we introduced a testbed of six image-text datasets to study the influence of pre-training data on the robustness of large-scale models such as CLIP. We analyzed the interactions of these datasets through both experiments and theory, finding that simply combining multiple datasets dilutes the robustness of the best-performing one. Moreover, we offered an explanation for the potential benefits of the data filtering process currently used for LAION. Our findings suggest that being more selective about what data to train on may provide more robustness than increasing the quantity of data, or mixing more data sources into the training set.

The broad societal implications of image-text models have been extensively discussed by Radford et al. [22]. As these models continue to grow in size, full-scale experiments are becoming prohibitively expensive. This necessitates small experiments that allow reliable extrapolation to larger scales. Since effective robustness is independent of model size and quantity of training data, we hope that our testbed and results can serve as a first step towards building multimodal pre-training datasets in a principled manner. There remain interesting open questions about finding better ways to filter and combine samples from multiple data distributions, or how to build the most robust dataset given a fixed set of data sources. We look forward to addressing these in our future work.

References

[1] Bigscience model training launched. https://bigscience.huggingface.co/blog/model-training-launched. Accessed: 2022-05-18.

[2] Common crawl. https://commoncrawl.org/. Accessed: 2022-05-18.

[3] Laion-5b: A new era of open large-scale multi-modal datasets. https://laion.ai/laion-5b-a-new-era-of-open-large-scale-multi-modal-datasets/. Accessed: 2022-05-18.

[4] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[5] A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. Tenenbaum, and B. Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

[6] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[7] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[9] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut. 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*, pages 3558–3568, 2021.

[10] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[11] K. Desai, G. Kaul, Z. Aysola, and J. Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations (ICLR)*, 2019. https://arxiv.org/abs/1903.12261.

[14] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, 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*, pages 8340–8349, 2021.

[15] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[16] G. Ilharco, M. Wortsman, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

[17] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[18] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[19] J. P. Miller, R. Taori, A. Raghunathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon, and L. Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021.

[20] H. Pham, Z. Dai, G. Ghiasi, H. Liu, A. W. Yu, M.-T. Luong, M. Tan, and Q. V. Le. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021.

[21] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[23] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[24] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

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

[26] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[27] V. Shankar, R. Roelofs, H. Mania, A. Fang, B. Recht, and L. Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning (ICML)*, 2020. http://proceedings.mlr.press/v119/shankar20c/shankar20c.pdf

[28] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, 2021.

[29] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[30] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

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

[32] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[33] H. Wang, S. Ge, Z. Lipton, and E. P. Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

[34] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] see Section 7
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] we study publicly available data sources
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] see Appendix F.1, F.2, F.3
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] , our models were trained using code from *open_clip* GitHub repository, which is publicly available [16]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] see Appendix B
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] see Appendix C
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] see Appendix B
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] for datasets that have been published, we refer readers to the corresponding publications. For Shutterstock, the license follows what is stated on the website (see Section 3).

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]