Exploring Text-Guided Synthetic Distribution Shifts for Robust Image Classification

Ryan Ramos[1 [0009−0004−9914−8386]], Raphael Alampay[1 [0000−0001−7498−8830]], and Patricia Abu[1 [0000−0002−8848−6644]]

ALIVE, Ateneo de Manila University, Quezon City, Philippines
ryan.ramos@obf.ateneo.edu

Abstract. The empirical risk minimization approach of contemporary machine learning leads to potential failures under distribution shifts. While out-of-distribution data can be used to probe for robustness issues, collecting this at scale in the wild can be difficult given its nature. We propose a novel method to generate this data using pretrained foundation models. We train a language model to generate class-conditioned image captions that minimize their cosine similarity with that of corresponding class images from the original distribution. We then use these captions to synthesize new images with off-the-shelf text-to-image generative models. We show our method's ability to generate samples from shifted distributions, and the quality of the data for both robustness testing and as additional training data to improve generalization.

Keywords: robustness to distribution shift · synthetic data · foundation models.

1 Introduction

Contemporary machine learning models are heavily reliant on data, to the point that at smaller scales of training data, models fail to adapt to novel datapoints outside their original training distribution [6]. A classic illustration of this is image classification: if images of cows in a model's training distribution frequently involved grassy fields, the said model is likely to fail in identifying cows when they are present in beaches [3]. This phenomenon is a consequence of combining the empirical risk minimization paradigm and a non-large-scale data regime. With a limited training distribution, there is less information available to models for them to disentangle spurious correlations from target labels, and thus these shortcuts become leveraged as predictive features despite their lack of robustness under distribution shifts [6, 20].

To address this issue in practice, researchers evaluate their models on test sets in order to examine how well a model adapts to data it has explicitly not been optimized for during training. However, these test sets tend to be independent and identically distributed, and potentially share the same underlying distribution as the training data [20], thus eliminating the effectiveness of using a test set to probe the ability of a model to adapt to new domains. In line with

this, studies have taken steps towards building specifically-engineered datasets and benchmarks to probe models' abilities to generalize to out-of-distribution data [13, 8].

While many traditional datasets are carefully sourced and manually annotated, recent advancements in generative modelling and foundation models have proven able to generate clean, usable data at scale without being restricted to natural data [24, 11, 1]. More relevant to this study, they have also been used to successfully simulate distribution shifts in data [22, 12].

A persistent question in the field of synthetic distribution shifts is how far of a distribution shift can we successfully simulate, and what methods make synthetic distribution shifts possible. We continue this line of work by proposing a novel method for this task. We train a language model to generate captions that maximize their CLIP [16] distance from reference classes, and use these captions to steer generative image models towards outputs that align with predefined classes but do not represent the original distribution. Our method generates data that is on average 15% lower in CLIP similarity to images from the same class when compared to in-distribution data. As a robustness benchmark, our method reduces classification accuracy by approximately 40%. Our generated data also leads to performance gains when leveraged as additional training data.

2 Related Work

Given a distribution of training data, an out-of-distribution dataset would introduce some distribution shift relative to the original distribution. Specifically, if the training distribution is said to be the result of some causal model or environment E, the new dataset should be sampled from E', the product of intervening on or changing the joint distribution of E. When evaluated on a new joint distribution, model A is said to be more robust than model B if it sees a smaller drop in performance resulting from leveraging reliable features that are invariant across domains; less robust models would be exposed for relying on spurious correlations that are not invariant across domains.

Probing for robustness is traditionally performed with curated datasets, many of which require manual annotation or image post-processing. With ImageNet assumed as a train set for the task of image classification, distribution shifts include varying viewpoints [2], rendering images in different styles [8, 24], corrupting or perturbing the images [9], etc. Another popular approach would be to specifically engineer an adversarial benchmark, where data is selected particularly based on a reference model's inability to correctly classify the data with high confidence [10].

As curated datasets, these methods are essentially limited by what data is naturally available. Generative models have provided ways to artificially induce distribution shifts in data, such as texture changes [15] and optimally adversarial augmentations [7]. Particularly relevant to this study are methods such as [22] and D3S [12], which leverage recent advancements in diffusion models [4, 18] to generate out-of-distribution data for image classification. While these works

either utilize simple yet effective prompting techniques with frozen models, or leverage gradient information in the diffusion process, our method investigates a new space for optimizing generative model outputs for out-of-distribution data. We specifically investigate optimizing not the diffusion process, but a text generation process prior to text-guided diffusion instead. We discuss this further in Section 3.

3 Distribution-shifted Data Generation

Given a distribution of n images $I_C := \{I_0, I_1, \ldots, I_n\}$ sampled from some class C, our proposed method generates images from that class that fall under C but lie away from the original distribution. We achieve this by leveraging a pretrained text-to-image model G. Unlike previous works [22, 12], we leave the mechanics of G untouched but optimize the text generation process that would steer G.

In order for the generated images to lie away from the original distribution, the corresponding texts that will guide the image generation must be sensitive to the original data and not represent its distribution. To implement this, we consider the goal of maximizing the distance of our generated captions, and consequently generated images, from the original distribution in a shared image-text embedding space modelled by some multimodal embedder $\phi(\cdot)$. Thus, we finetune a generative language model L to generate a caption P_C conditioned on a given class C that minimizes the following objective function R:

$$R(P_C) = s\left(\phi(P_C), \frac{\sum_{i=0}^{n} \phi(I_i)}{n}\right) \tag{1}$$

where $s(\cdot, \cdot)$ measures the similarity between two points in an embedding space, with higher values indicating stronger similarity. This value compares a generated caption with the mean of ϕ-encoded images of I_C, thus minimizing this value would maximize the distance of the caption from the original data distribution. Because a language model can create arbitrary noise to maximize the distance of its outputs from a reference distribution in the embedding space, we further modify R by returning the lower bound of s for any any non-sensical (incomplete sentence and failure to mention class C in prompt) output. Generated captions can then be used as prompts for G without further finetuning or modifications.

4 Experiments

4.1 Implementation

For our experiments, we focus on the distribution of images found in ImageNet. Specifically, we focus on the 86 class subset intersection between ImageNet-A [10] and ImageNet-R [8] for comparability purposes. To calculate mean embeddings for each class, we sample 128 images per class.

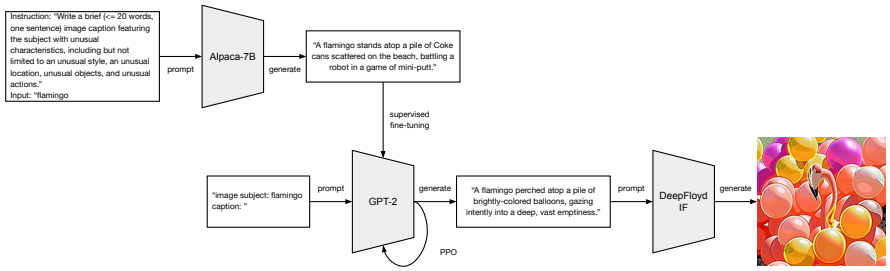


Fig. 1: Overview of method.

"A bright yellow school bus cradles up a cobroad road filled with colorful balloons, orange skies, and a peaceful blue sky, amidst a smooth green hue of vibrant stars."

"The school bus swerved down the track, its strange-looking driver screeching loudly as it sped through the streets."

"A school bus standing in a meadow with bright lights, its seats only dimly lit with dimming lights."

"An old school bus trundles down the cobblestone street, its bright colors shimmering in the sunlight."

Fig. 2: Sample generated texts with corresponding generated images for the ImageNet class "school bus".

An overview of the generation pipeline is shown in Fig. 1. We leverage DeepFloyd IF, an open-access implementation of Imagen [19], for G; CLIP [16] for ϕ; and base our L on GPT-2 [17]. Our similarity scorer s is cosine similarity. We also use CLIP cosine similarity as a filter to remove low-quality data from our generations, using a threshold of 0.2.

In training L, we first use an Alpaca-7B [23] to generate 16 sample captions per ImageNet class to create 16000 examples of class-conditioned prompts. We then use this data to perform supervised finetuning on a GPT-2 followed by reinforcement learning with PPO [21].

For all reference methods, in order to maintain comparability as regards image fidelity, we reimplement each method using the same generative image model backbone, model precision, and diffusion sampling steps.

Sample generations of our method are presented in Fig. 2.

4.2 Measuring Distribution Shift

Table 1 describes our generated data compared to both the original data and similar methods. Mahalanobis distance was calculated in CLIP space, reduced to

Table 1: Distance metrics describing relation of datapoints to class mean.

Method	CLIP similarity ↓	Mahalanobis distance ↑
Base distribution	83.50	7.90
Low-density [22]	70.26	24.70
D3S [12]	73.74	24.49
Ours	73.93	23.30

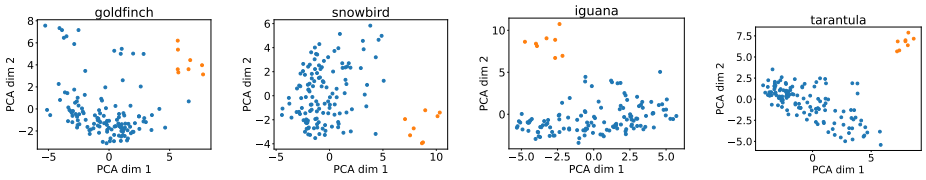


Fig. 3: PCA-reduced plots of datapoints in CLIP space. Blue points represent the original distribution while orange points represent our generated data.

64 dimensions via PCA. While our method does not outperform other methods in CLIP similarity or Mahalanobis distance, we maintain similar performance as regards maintaining a gap from the base distribution. Examples of this gap can be visualized in Fig. 3, which shows examples in a PCA-reduced space. Furthermore, Fig. 4 shows the percent reduction in cosine similarity of datapoints to the class mean in CLIP space across all examined classes, with our method only creating datapoints closer to the mean in one class while reducing similarity by approximately 9.5% on average.

4.3 Probing Adversarial Ability

We also measure the strength of our distribution shift via zero-shot evaluation on our data, where pretrained classifiers are expected to drop in accuracy under stronger distribution shifts, assuming the model relies on spurious correlations rather than causal features. The results of these experiments are seen in Table 2. Though we do not outperform both methods, our method is able to reduce classification accuracy by around 40% on each model, indicating the potential for this method to act as a challenging benchmark for image models.

4.4 Use as Training Data

Aside from acting as a robustness benchmark, another potential use of out-of-distribution data is to act as training data in order to widen the training distribution and induce better generalization ability during inference. We fine-tune a ConvNeXt-Atto [14] on data generated with our method as well as two other synthetic distribution-shift methods, with 8 generated samples per class, then evaluate performance gains in the out-of-distribution setting, in this case ImageNet-A and ImageNet-R. Our results are seen in Fig. 6.

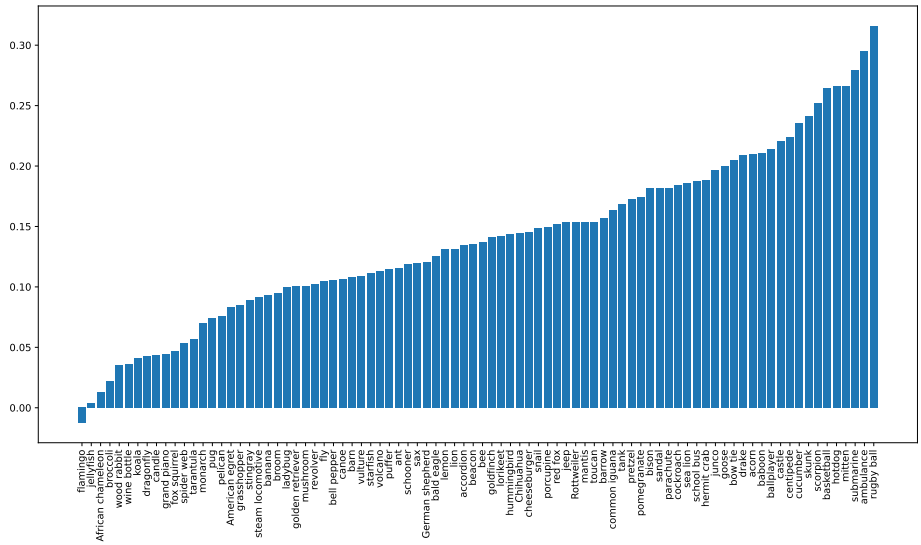


Fig. 4: Percent reduction in average cosine similarity of generated datapoints in a class compared to original datapoints from the same class, in CLIP space.

Table 2: Accuracy reduction of ImageNet classifiers on distribution shift validation datasets.

Method	ConvNeXt-Atto [14]	ViT-Tiny [5]
Low-density [22]	-62.04	-58.73
D3S [12]	-34.71	-43.04
Ours	-39.51	-41.87

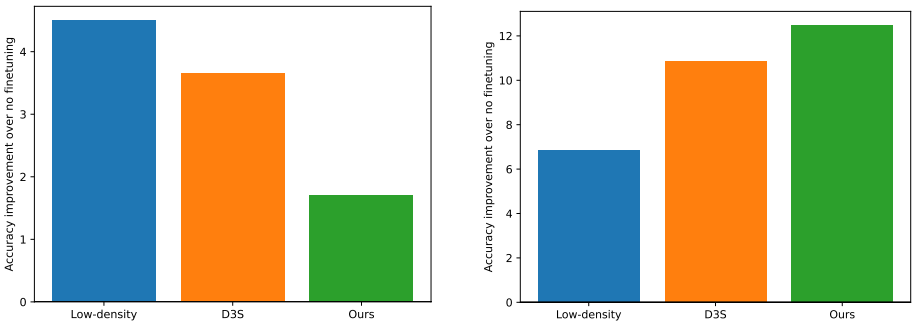
We show the most improvement on ImageNet-R, while introducing the least improvement on ImageNet-A. We believe this to be attributed to the data generation process of each method. While our method produces a mix of realistic and artistically rendered images, the other methods mainly produce photo-like images. At the current data scale, the other methods have a greater impact for ImageNet-A, which mainly includes photos, while our data has the most impact for ImageNet-R, which does not exclusively focus on photos.

5 Conclusion

We introduce and explore the potential of a new method for generating synthetic out-of-distribution data by focusing on the text generation process used to steer a generative text-to-image model. We describe our method's ability to successfully generate samples that are distanced away from the original distribution, and quantitatively report its ability to act as a test set for robustness probing and as a

(a) ImageNet (b) Low-density [22] (c) D3S [12] (d) Ours

Fig. 5: Images of the ImageNet class "German shepherd" drawn from (5a) the ImageNet training set, (5b) sampling from low-density regions, (5c) generating images with different backgrounds, and (5d) our proposed method of generating captions to guide out-of-distribution image generation.



(a) ImageNet-A [10] (b) ImageNet-R [8]

Fig. 6: Accuracy improvements on adversarial datasets by fine-tuning on distribution shift datasets.

train set to increase generalization ability. Future studies can explore this method at larger scales with stronger foundation models or increased data generation.

References

1. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. arXiv preprint arXiv:2304.08466 (2023)
2. Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., Katz, B.: Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. Advances in neural information processing systems **32** (2019)
3. Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European conference on computer vision (ECCV). pp. 456–473 (2018)
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)

5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
7. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
8. 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. pp. 8340–8349 (2021)
9. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)
10. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15262–15271 (2021)
11. Honovich, O., Scialom, T., Levy, O., Schick, T.: Unnatural instructions: Tuning language models with (almost) no human labor. arXiv preprint arXiv:2212.09689 (2022)
12. Kattakinda, P., Levine, A., Feizi, S.: Invariant learning via diffusion dreamed distribution shifts. arXiv preprint arXiv:2211.10370 (2022)
13. Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Beery, S., et al.: Wilds: A benchmark of in-the-wild distribution shifts 2021. arXiv preprint arXiv:2012.07421 (2020)
14. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
15. Pintor, M., Angioni, D., Sotgiu, A., Demetrio, L., Demontis, A., Biggio, B., Roli, F.: Imagenet-patch: A dataset for benchmarking machine learning robustness against adversarial patches. Pattern Recognition **134**, 109064 (2023)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
17. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
19. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
20. Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y.: Toward causal representation learning. Proceedings of the IEEE **109**(5), 612–634 (2021)
21. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)

22. Sehwag, V., Hazirbas, C., Gordo, A., Ozgenel, F., Canton, C.: Generating high fidelity data from low-density regions using diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11492–11501 (2022)
23. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca (2023)
24. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in Neural Information Processing Systems **32** (2019)