Is Seeing Still Not Necessarily Believing?

Siyu An, Sidharth Singh, Yating Chen, JangHyeon Lee*

May 17, 2023

1 Introduction

Deep neural networks have shown revolutionary improvements in computer vision tasks. However, these models are notorious for being data-hungry, requiring vast amounts to achieve optimal performance. Without sufficient data, deep learning models are prone to overfitting, resulting in a poor representation of the underlying distribution of unseen data. This leads to a fair share of challenges, such as a lack of quality data in specific domains and costly data collection. For example, the medical domain struggles with a scarcity of data on rare cancer types, while data privacy concerns hinder full data access. Additionally, in-the-wild data often consists of artifacts, compromising the integrity of the data. On the other hand, capturing rare events demands significant patience and resources to amass large datasets.

Considerable efforts have been made to develop practical solutions to address data scarcity. Traditional augmentation techniques such as cropping and rotation have been used to reduce model bias toward object orientation. However, these methods fall short when faced with distortions, like varying weather conditions that globally affect learnable features. Consequently, advanced approaches, such as BigGANs, were introduced for data augmentation. Unfortunately, despite their ability to produce photo-realistic samples, the synthetic data failed to serve as useful extra training data. This led the authors to conclude *Seeing is Not Necessarily Believing* as their title. Our work explores the potential of replacing BigGANs with diffusion models. By mirroring prior work, our experiments answer the question: *Is Seeing Still Not Necessarily Believing?*. For convenience, our quantitative and qualitative results are on the last page.

2 Related Work

Seeing is Not Necessarily Believing [2], aimed to understand if BigGAN samples could be used for data augmentation. The authors found that the photorealistic samples were not useful as extra training data and that neither Inception Score (IS) nor Fréchet Inception Distance (FID) score was relevant to classification performance. In contrast, our work extends the findings of this study by evaluating multiple diffusion models across a broader range of benchmarks rather than focusing on a single BigGAN or ImageNet dataset. Additionally, we endeavor to understand why diffusion models are more advantageous than GANs in data augmentation.

Invariant Learning via Diffusion Dreamed Distribution Shifts [1], addressed the problem of classifiers over-relying on background information rather than foreground objects themselves. The authors introduced the D3S dataset, which was generated using Stable Diffusion and featured unusual fore-background combinations. They found that pretrained classifiers on ImageNet were not robust to these fore-background shifts. Thus, they introduced a feature extractor to learn invariant features for both fore-background elements. However, our work emphasizes the use of synthetic data as extra training data for classifiers, thus presenting a different narrative.

3 Proposed Methodology

Our pipeline in Figure 1 has two stages: augmentation and classification and an interval stage for evaluation. The first stage uses Stable Diffusion (SD) and Conditional Diffusion (CD) to generate synthetic versions of the original datasets. The SD model, pretrained on the LAION-5B dataset, takes class names from CIFAR-10 and Imagenette as input text prompts. The CD model trains a simple Denoising Diffusion Probabilistic Model (DDPM) from scratch using a domain-specific dataset (melanoma). The input consists of noise conditioned on class labels (benign or malignant) as additional input features. Then, we evaluate the synthetic data using the IS and FID score to assess the diversity of the datasets. The second stage trains a ResNet-18 model on the datasets created in the first stage. The training process consists of 1) using only original, 2) using only synthetic, and 3) combining original and synthetic data. The model is then tested on the validation subset of the original data to evaluate its usefulness in a downstream task such as classification. The hyperparameters are fixed.

4 Results & Discussion

Table 1 shows that synthetic data alone performs on par with original data suggesting that data replacement can be a feasible alternative for training models. Moreover, unlike BigGAN-generated data, most diffusion-generated data improves classifiers when mixed, answering as *not likely* to the question *Is Seeing Still Not Necessarily Believing?* Naturally, our findings prompt a question *why does diffusion-driven augmentation outperform BigGANs?* We assume that the reason lies in the absence of a robust discriminator. By training a separate classifier using only original data and testing it on synthetic data, we find that the discriminator struggles to classify diffusion-generated images. This observation suggests that diffusion-driven results better capture properties more closely aligned with the data distribution than GANs.

However, we also found samples, such as indoor churches and planes from Figures 2 and 3, that do not represent their respective classes leading to potentially adverse outcomes. Furthermore, since Stable Diffusion was pretrained on the LAION-5B dataset, there exist notable distinctions in data distribution compared to the original CIFAR-10, which may have harmed the training process. The high FID score for synthetic CIFAR-10 may indicate the cause. Finally, the generally low accuracy for Imagenette can be attributed to the resizing of images to very low resolutions (32x32).

5 Conclusion & Future Work

In this work, we identified key challenges in deep learning, especially the lack of quality data. Several attempts have been made to address this issue, such as using BigGANs for data augmentation, but the photorealistic samples have not been helpful for extra training data. In response, our study offer two contributions. First, we extended prior work by replacing BigGANs with diffusion models, and second, we introduced a challenging domain-specific dataset (melanoma) to further our investigation.

Looking forward, we plan to explore the usefulness of synthetic data for other downstream tasks, such as object detection or semantic segmentation. Next, we intend to investigate techniques for generating or filtering only *useful* data that may sacrifice photorealism for data with more useful learnable features. Lastly, we recognize the need for standardization in evaluation metrics and aim to contribute to developing a novel metric that better reflects the quality and diversity of the generated data.

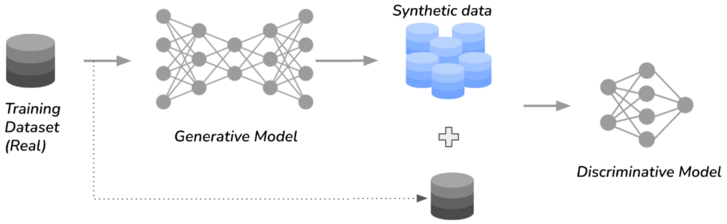

Figure 1: Pipeline

	ORG (Acc)	SYN_SD (Acc/IS/FID)	MIX_SD (Acc)	SYN_CD (ACC/IS/FID)	MIX_CD (Acc)
CIFAR-10	81.3	34.7/9.9/73.2	12.1	77.3/9.5/12.4	**83.4**
Imagenette	22.9	17.9/11.3/31.1	**24.4**	19.8/11.7/7.8	**25.9**
Melanoma	61.7	N/A	N/A	58.1/8.1/17.1	**64.9**

Table 1: Quantitative Results

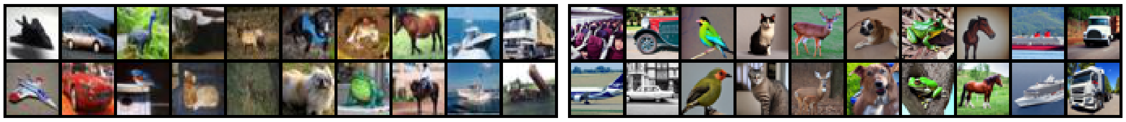

Figure 2: Real (left) vs Fake (right) for CIFAR-10

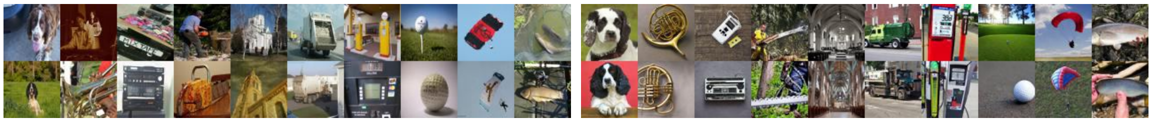

Figure 3: Real (left) vs Fake (right) for Imagenette (a subset of ImageNet)

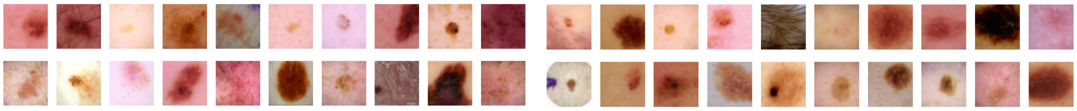

Figure 4: Real (left) vs Fake (right) for Melanoma

References

[1] Priyatham Kattakinda, Alexander Levine, and Soheil Feizi. Invariant learning via diffusion dreamed distribution shifts. *arXiv preprint arXiv:2211.10370*, 2022.

[2] Suman Ravuri and Oriol Vinyals. Seeing is not necessarily believing: Limitations of biggans for data augmentation. *International Conference on Learning Representations*, 2019.