Data Augmentation as an Inverse Problem

Hayden Prairie Frank Collebrusco David Shilliday Ranit Gupta Ashwin Ram Andrew Wang

Abstract

With the growth in the complexity of neural networks and the amount of available compute, the need for larger datasets has become a bottleneck in the advancements of computer visions. In this paper, we explore a new approach to synthetic data generation by treating data augmentation as an inverse problem. We utilize posterior sampling with latent diffusion models to generate synthetic data through masking. Finally, we discuss potential improvements and future work that can be done. Code is available at https://github.com/Hprairie/Synthetic-ImgGen-PSLD.

1. Introduction

In recent years deep learning has made incredible strides in computer vision tasks such as image classification, object detection, and semantic segmentation. Most advancements within this field can be attributed to the improvement of novel architectures, an increase in compute, and access to large datasets. However, the ability to scale a model without scaling the size of its corresponding dataset has been shown to have diminishing returns (Kaplan et al., 2020). The ability to generate large labeled datasets by hand is infeasible and thus the need for unsupervised techniques to create augmented or synthetic data becomes necessary. Naive approaches to data augmentation such as random cropping, flipping, rotation, etc. are effective in some cases, but unfeasible in other cases such as medical imaging, where naive approaches create unrealistic image domains. More advanced techniques such as generative adversarial networks (Goodfellow et al., 2014) are more effective in creating realistic synthetic data, however, these often require large amounts of data to fine-tune models.

In the last year, the surge in the power of large language models such as GPT-4 has also shown their potential to create synthetic data (Møller et al., 2023). The ability of LLMS to generate realistic text responses to prompts could be viewed as an inverse problem, in which the LLM is attempting to find the most likely text given a prompt. A similar approach can be taken with images, where diffusion models can be used to solve inverse problems and generate synthetic data.

In this paper we explore the use of posterior sampling with latent diffusion models (PSLD) (Rout et al., 2023) to create synthetic data. PSLD allows us to leverage the power of latent diffusion models to solve linear inverse problems in images. Using masking and reconstruction with the PSLD algorithm, variations of the original image can be created, due mainly to the inherent lossy relations of masking pixels. We found these synthetic images to be effective in improving the performance of object detection models.

The rest of the paper is organized as follows. First, we cover the related work in comparison with previous naive and advanced data augmentation techniques. Then we cover our approach to utilize diffusion models for synthetic data augmentation. Finally, we cover the results of our initial experiments and potential improvements along with future work that can be done.

2. Related Work

Data augmentation and synthetic data generation aim to artificially increase the size of the training set to improve the robustness of computer vision models and prevent overfitting to a training set. Similar to dropout in neural networks, data augmentation can be useful in leading a model to a more generalizable solution (Zhong et al., 2017).

2.1. Niave Data Augmentation

Most naive data augmentation techniques apply simple transformations to the input image while maintaining the label, such as flipping, cropping, rotating, translation, color jittering, and adding noise. Other more complex techniques such as random erasing (Zhong et al., 2017) aim to occlude parts of the image allowing the model to learn to focus on other features. These techniques are often effective in improving the performance of image classification models, however, they are often negligible in object detection settings (Yang et al., 2022). Furthermore, these transformations are often ignorant to certain domains such as medical imaging, where simple transformations do not apply to potential real-world domains.

For example in brain CT scans, linear transformations of the image would create unrealistic domains that are not

representative of test/validation data. The use of linear transformations to improve the robustness of a model in these domains often doesn't work as inference time samples will be centered and cropped such as in CT scans. Thus the need for variation in the training set under specific domain constraints is necessary.

2.2. Advanced Data Augmentation

In more recent years the use of generative adversarial networks (GANS) (Goodfellow et al., 2014) has been effective in creating realistic images. Several papers have attempted to leverage their power to create synthetic (Besnier et al., 2019), (Zhang et al., 2021), (Jahanian et al., 2021). While the techniques have been seen to be effective, their ability to generate realistic images pales in comparison to latent diffusion models (Rombach et al., 2021). Especially when considering the power of latent diffusion when trained on datasets such as LAOIN-5B (Schuhmann et al., 2022), the power of GANs pales in comparison.

2.3. Diffusion Models

As the power of latent diffusion models grows, the desire to apply their power to synthetic data generation has become increasingly more popular. Some work attempts to graft target images partway through the diffusion process creating more variations of the original images at the cost of faithfulness to the target class (Meng et al., 2021). This was then tested in zero-shot and few-shot settings, where it was shown to be effective in improving the performance of image classification (He et al., 2022). In both of these situations, the diffusion model is guided by the text prompt, creating another issue as the model is unable to generalize to new images where vocabulary is used outside of its training set. Other work has attempted to solve this by inserting embeddings into the textual encoder and then using textual inversion to fine-tune the model, allowing it to generalize new vocabulary (Trabucco et al., 2023). However, the significant issue with these approaches is their higher likelihood to generate unfaithful images and their need to fine-tune the diffusion model, which often requires a large amount of compute.

There is some work on the use of inpainting in semantic segmentation tasks, where morphological erosion was applied to the mask which allowed the model to better generalize the inpainting process and inpaint more faithful images (Pobitzer, 2023). However, there are still some issues with this approach as well, as the model still uses prompt guidance and is unable to generalize to domains not in its training vocabulary.

3. Our Approach

A desirable way to create synthetic data would create more variation than naive data augmentation techniques, but without the need for large data and compute to train GANs. Thus we propose a method of treating data augmentation as an inverse problem, in which synthetic samples can be generated by utilizing PSLD and allowing the latent diffusion models to fill in missing pixels. For example, consider the matrix A which transforms the image x to a 'corrupted' image $\hat{x}$ and can then be passed to the PSLD algorithm which will attempt to reconstruct the original image x. The resulting image is a synthetic sample $\tilde{x}$ which can then be used as a training sample. In this paper we only attempt to reconstruct an image through masking, however, other transformations such as gaussian blurring, motion blurring, and lossy compression could potentially be used as well.

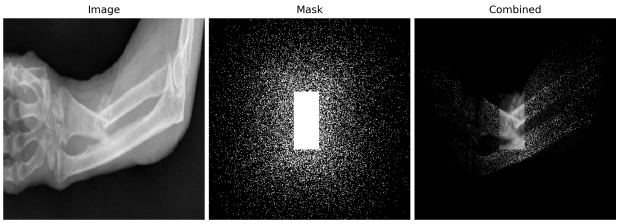

Figure 1. Example of sample passed into PSLD with corresponding image and mask.

While several masking techniques could potentially be useful, i.e. random, gaussian, in bounding box, out of bounding box, etc. we only tested on gaussian out of the bounding box masking, which is displayed in Figure 1. A mask can be generated by first completely masking the inside of a given sample's bounding box and then creating a matrix D which is the same shape of our image but contains the distance from any given pixel to the nearest masked pixel. We can then sample a gaussian distribution $N \sim (0, \sigma)$ at each pixel, and then mask any pixel where $d \in D$ is greater than its given sample from N. This results in a mask where pixels closer to the bounding box are more likely to be masked than pixels further away from the bounding box.

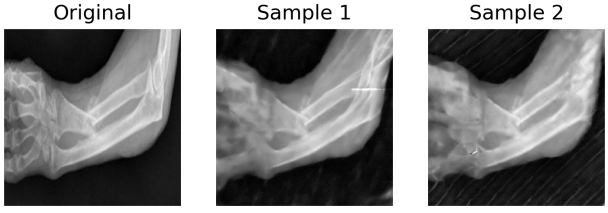

Figure 2. Example of sample passed into PSLD with corresponding results.

The image with its corresponding mask can then be *solved* as an inverse problem using the PSLD algorithm, creating new synthetic samples due to the lossy nature of estimating

Table 1. Results of experiments on bone fracture dataset.

Augmentation	Train Box ↓	Train Objective ↓	Val mAP@0.5 ↑	Val F1 ↑	Test mAP@0.5 ↑	Test F1 ↑
Baseline	0.02726	0.004067	0.6969	0.73	0.699	0.73
Naive	0.03984	0.005202	0.7298	0.74	0.746	0.73
PSLD	**0.0254**	**0.003194**	**.7349**	**0.77**	**0.782**	**0.82**

pixel values. Examples of resulting images can be seen in Figure 2.

As described this approach allows the creation of synthetic samples that are not prompt guided or limited to a given vocabulary, don't require model finetuning, and are not limited to a given domain. However, the downside to this approach is the diminished realism in the generated images, as a preference for class faithfulness comes at the cost of realism.

4. Experiment

We ran several experiments on the effectiveness of synthetic data generation using PSLD. We used a bone fragment dataset which consists of 1000 images (750 train, 50 val, 200 test) of bone fractures from all over the body (Fracture, 2023). The reason for using this dataset is its increased probability of lying outside the training set. We used stable diffusion v1.4 (Rombach et al., 2022), a latent diffusion model trained on LAION-Aesthetics V2 (Schuhmann et al., 2022). We created roughly 1000 synthetic samples using PSLD and then compared it with other naive data augmentation techniques where we created a similar number of samples.

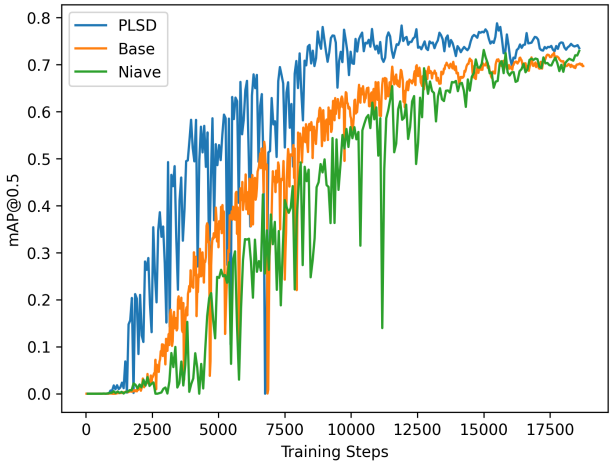

Figure 3. Training curves of YOLOv7 model concerning each data type.

The results of the experiment can be seen in Table 1 where we compare the performance of a YOLOv7 model (Wang et al., 2023). We found that the use of synthetic data was incredibly effective in improving the performance of our baseline YOLOv7 model. Compared with other naive data augmentation techniques, we found that the use of PSLD was also more effective in improving the performance of our model. The training curves of the model can be seen in Figure 3, which is adjusted for gradient steps instead of epochs.

One thing interesting to note is that samples are often not too visually different, however, there are often large differences at the pixel level. This can be seen in Figure 4, where the difference between the two images at a pixel level is shown scaled to an extremely large exponent of 400. Nevertheless, this still enables to model to learn more robust features and generalize to the validation and test set.

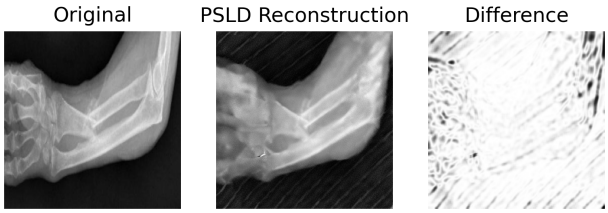

Figure 4. Difference between the original image and its synthetic sample. Pixel difference x is scaled to x^{400}.

5. Discussion and Future Work

Treating data augmentation as an inverse problem allows for better variation of synthetic samples compared to naive data augmentation techniques, and can be done without the need to retrain on a specific dataset. However as seen, the realism of the generated images is not as comparable to GANs or other diffusion based techniques but is inherently more faithful to the original class. Currently, the PSLD algorithm is unable to effectively use the prompt as a guide to reconstruction, however, if the effect of the prompt is considered within the gradient updates of PSLD, the potential for its implementation could be explored. Furthermore finetuning the diffusion model on the training set using methods such as ambient diffusion (Daras et al., 2023) could also be explored to improve the ability of the diffusion model to generalize to a specific domain, i.e. medical imaging. Additionally, an understanding of the effects that ambient diffusion has on the size of data needed to finetune diffusion models would be interesting to explore. This could be taken

a step further by using approaches described by (Trabucco et al., 2023), where embeddings could be learned through textual inversion and allow better class-specific guidance through the inpainting process.

Finally, the use of other inverse problems such as motion blurring, gaussian blurring, and lossy compression could also be explored as ways to induce other types of variation within the diffusion model. The use of inpainting is intuitive, however, there is no reason to believe that other inverse problems could be used to create unique synthetic sample reconstructions compared to other domains.

Importantly the need to test the effectiveness of these methods for OOD datasets from diffusion models training sets is necessary to understand the generalizablity of these methods. While the use of these diffusion models is promising, the need to ensure their ability to generalize to new domains is important in understanding their ability to create synthetic data. Further understanding of whether diffusion models are creative or have simply memorized their training set is essential to understanding their abilities, as synthetic data generators.

Acknowledgements

We would like to thank Litu Rout for spending time to help explain his work on solving inverse problems with PSLD. We would also like the thank the Machine Learning Lab for providing credits to run our experiments.

References

Besnier, V., Jain, H., Bursuc, A., Cord, M., and Pérez, P. This dataset does not exist: training models from generated images, 2019.

Daras, G., Shah, K., Dagan, Y., Gollakota, A., Dimakis, A. G., and Klivans, A. Ambient diffusion: Learning clean distributions from corrupted data, 2023.

Fracture. Bone fracture detection dataset. https://universe.roboflow.com/fracture-uofxm/bone-fracture-detection-ivsy6, 2023.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014.

He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., and Qi, X. Is synthetic data from generative models ready for image recognition?, 2022.

Jahanian, A., Puig, X., Tian, Y., and Isola, P. Generative models as a data source for multiview representation learning, 2021.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models, 2020.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2021.

Møller, A. G., Dalsgaard, J. A., Pera, A., and Aiello, L. M. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks, 2023.

Pobitzer, M., 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Rout, L., Raoof, N., Daras, G., Caramanis, C., Dimakis, A., and Shakkottai, S. Solving linear inverse problems provably via posterior sampling with latent diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=XKBFdYwfRo.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

Trabucco, B., Doherty, K., Gurinas, M., and Salakhutdinov, R. Effective data augmentation with diffusion models, 2023.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., and Shen, F. Image data augmentation for deep learning: A survey, 2022.

Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.-F., Barriuso, A., Torralba, A., and Fidler, S. Datasetgan: Efficient labeled data factory with minimal human effort, 2021.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. Random erasing data augmentation, 2017.