Data Augmentation as an Inverse Problem

Hayden Prairie

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 an new approach to synthetic data generation by treating data augmentation as an inverse problem. We ultize posterior sampling with latent diffusion models in order 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 feild 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 correspondeing dataset has been show to have diminishing returns (Kaplan et al., 2020). The ability to generated large labeled datasets by hand is infeasable and thus the need for unsupervised techniques to create augmented or synthetic data become's necissary. Niave approaches to data augmentation such as random cropping, flipping, rotation, etc. have been shown to be effective in some cases, but unfeasable in other cases such as medical imaging, where niave approaches create unrealistic image domains. More advanced techniques such as generative adversarial networks (GANs) have been shown to be effictive in creating realistic synthetic data, however these often require large amounts of data in order 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. The ability for LLMS to generate realistic text responses to prompts can be viewed as an inverse problem, where the model is attempting to find the most likely text given a prompt. A similar approach can be applied to images, where diffusion models can be used to solve inverse problems and in turn generate synthetic data.

In this paper we explore the use of posterior sampling with latent diffusion models (PSLD) to create synthetic data. PSLD is a algorithm which allows the use of latent diffusion models to solve linear inverse problems in images. By generating a mask and then using PSLD to reconstruct the image, variations of the original image can be created, due mainly the inherent lossy relations of masking pixels. We found these synthetic images to be effective in improving the performance of object detections models.

The rest of the paper is organized as follows. First we cover the related work along with the benefits and downsides of niave and advanced data augmentation techniques. Then we cover our approach of treating augmentation as an inverse problem and utilize diffusion models for synthetic data generation. Finally we cover the results of our initial expirements and potential improvements along with future work that can be done.

2. Related Work

Data Augementation and synthetic data generation aim to artifically increase the size of training set in order to improve 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 niave data augmentation techniques apply simple tranformations 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 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 computer vision models to an extent, however they are often ignorant to certain domains such as medical imaging, where simple transoformations are not applicable 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 in order to improve the robustness of a model in these domains often doesn't work as infrence 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 are necissary.

2.2. Advanced Data Augmentation

In more recents years the use of GANs have been show to be effective in creating realistic synthitic data.

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 insert 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 unfailful images and their need to fintune 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 faithfull images (Pobitzer, 2023). However there are still some issues with this approach as well, as the model is still uses prompt guidance and is unable to generalize to new domains not in its training vocabulary.

3. Our Approach

A desirable way to create synthetic data would create more variation than niave 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 '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 be then be used as a training sample. In this paper we only attempt to reconstruct an image through masking, however other transformations such as gaussion blurring, motion blurring, and lossy compression could potentially be used as well.

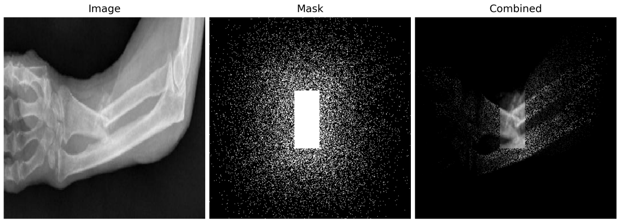

Figure 1. Example of sample passed into stable diffusion with corresponding image and mask.

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 samples 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 it's given sample from N. This results in a mask where pixels closer to the bounding box are more likely be masked than pixels further away from the bounding box.

The image with it's 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 pixel values. Examples of resulting images can be seen in Figure 2.

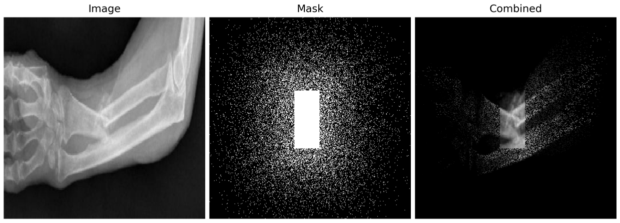

Figure 2. Example of sample passed into stable diffusion with corresponding image and mask.

As described this approach allows the creation of synthetic samples 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 that realism in the generated images, as a preference to class faithfulness comes at the cost of realism.

Table 1. Your caption

Augmentation	Train Box	Train Objectiveness	Validation mAP@0.5	Validation F1	Test mAP@0.5	Test F1
Baseline	0.02726	0.004067	0.6969	0.73	0.699	0.73
PSLD	**0.0254**	**0.003194**	**.7349**	**0.77**	**0.782**	**0.82**

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. The reasoning for using this dataset is that it is more likely to be outside the training dataset. We used stable diffusion v1.4, which is a latent diffusion model trained on TEMP NEED TO FIX. 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.

The results of the experiment can be seen in Table 1 where we compare the performance of a YOLOv7 model. 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.

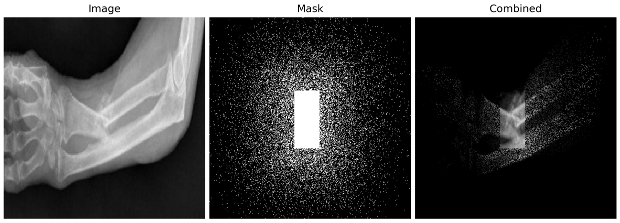

Figure 3. Example of sample passed into stable diffusion with corresponding image and mask.

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 are not as comparable to GANs or other diffusion based techniques, but are 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 affect of the prompt is consider within the gradient updates of PSLD, the potential for it's 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 in order to improve the ability of the diffusion model to generalize to a specific domain, i.e. medical imaging. Furthermore, an understanding into 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 farther 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 in order 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 wether diffusion models are creative or have simple 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.

References

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

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.

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.

Pobitzer, M., 2023.

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

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