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 to in order 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 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.

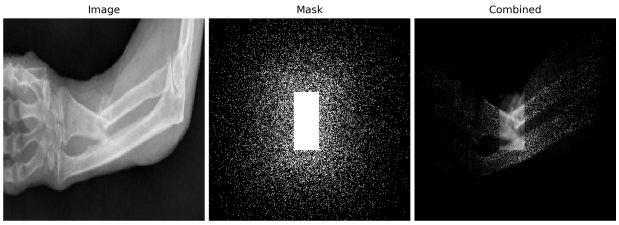

Figure 1. Posterior sampling with latent diffusion (PSLD) is a method of solving linear inverse problems in images by utilizing the power of latent diffusion models such as Stable Diffusion.

In this paper we explore the use of posterior sampling with latent diffusion models (PSLD) to create synthetic data. PSLD is a method of solving linear inverse problems in images by utilizing the power of latent diffusion models such as Stable Diffusion. 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 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. Data Augmentation Techniques

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 jit-

tering, 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.

2.2. Advanced Data Augmentation

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

3. Our Approach

4. Results

5. Discussion and Future Work

Acknowledgements

We would like to thank Litu Rout for spending time to help explain his work on solving inverse problems with PSLD.

References

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.

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