7C-06

Self-Supervised Pre-training of Vision Transformers Using Stable Diffusion-Generated Images

Luiz H. Mormille[1,a)]   Masayasu Atsumi[1]

Abstract: Traditional dataset building involves time-consuming tasks such as web scraping, cleaning, and labeling. Our proposed method utilizes a stable diffusion technique to efficiently generate synthetic images from text prompts, eliminating the need for manual data collection while mitigating biases and mislabeling. We conduct experiments with a vision transformer, comparing models trained on real datasets and datasets enhanced with synthetic images. The results showcase the efficacy of stable diffusion-synthesized images in enhancing model generalization and accuracy, highlighting the potential of this approach in the realm of computer vision.

1. Introduction

Gathering data from the real world is a complex, costly, and time-consuming endeavor. Traditional machine learning datasets are frequently noisy or lack sufficient curation and size [1]. As a result, obtaining high-quality data remains a critical yet challenging aspect of developing effective predictive systems.

In recent years, large-scale Stable Diffusion models, trained on extensive amounts of noisy data, have showcased robust generative capabilities [2]. Hence, given the vastness of the cross-domain knowledge instilled in them, Diffusion models emerge as a powerful assistant or even alternative to real data, generating high-quality training samples for discriminative models.

In the past, synthetic images contained a limited diversity, hindering these samples possibilities of training high accuracy classifiers, specially when compared to those trained on real-image datasets [3–5]. However, recent state-of-the-art works on diffusion-based text-to-image models showcase impressive capabilities in synthesizing visually faithful images.

Therefore, the quality of synthetic images may no longer be an issue when using them to pre-train discriminative models, and the advantages of employing them go beyond overcoming the cost constraints linked to collecting and annotating real images.

One strategy for using these models is to augment the original dataset using prompt-based generated images. This approach often employs domain specific vocabulary, which are further enhanced using natural language techniques, in order to produce prompts that are then used for generation. This approach can indeed yield high-quality and diverse images [3, 6].

Nevertheless, in spite their high quality and diversity, generation based on prompts frequently produces images that are unrelated to the target domain, leading to the creation of subpar datasets [3, 4, 7]. Furthermore, this method often overlooks the distribution of the original train set, generating images from a distribution distant from the original data, thus leading to substantial differences between real and generated datasets [1].

In response to the aforementioned challenges, our study introduces a framework that revolves around the self-supervised pre-training of a small-scale vision transformer — ARViT [8] — leveraging the expansive ImageNet dataset enriched with synthetic data generated through a stable diffusion model [9]. Images are synthesised based on text prompts taken from a proposed *prompt sampler* and the Self-supervised pre-training specifically employs the rotation estimation task, which in practice eliminates data mislabeling. To gauge the impact of the incorporated synthetic data on the pre-training process, subsequent fine-tuning of the models is conducted on four small-scale benchmark downstream classification tasks.

The findings elucidate that augmenting the training dataset, even if by incorporating synthetic data sampled from a disparate distribution, increased the accuracy of models on all four downstream classification tasks.

2. Method

The central proposition of this work is to leverage a Stable Diffusion to produce synthetic images from text prompts, combine the generated images with real images, and use them to pre-train a small vision transformer. The stable diffusion model adopted was the *Stable Diffusion 2* by Stability-AI[*1]. The schematics of the pre-training process is shown in Figure 1. To generate the text prompts to be used by the stable diffusion model, a text prompt sampler was devised. It contains a total of 1000 words for categories, 8 for size variations, 12 for colors, 10 for angles and views and 10 for illuminations, culminating in a total of 9.6 million possible prompts which are randomly sampled. Examples of prompts and the respective images generated from them are shown in Fig-

[1] Soka University
[a)] luiz@soka.ac.jp

[*1] available from ⟨https://github.com/Stability-AI/stablediffusion⟩

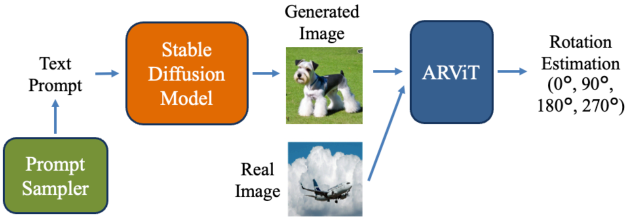


Fig. 1 Overview of the pre-training routine. Text prompts are sampled from a space with more than 9.6M prompts. The prompts are then fed to the stable diffusion model, which generates images based on them. Images are then used, alongside real images, to pre-train a vision transformer (ARViT) on a self-supervised rotation estimation task.

ure 2. Finally, the images generated from the text prompts by the stable diffusion model are then utilized, alongside real images sampled from ImageNet, to pre-train a vision transformer.

3. Experiments

A total of 5 distinct models were pre-trained. They varied based on the amount of synthetic data used to augment the original dataset. A first baseline model was pre-trained using exactly 1M real images sampled from ImageNet, and we denote it ARViT-Base. Another model was pre-trained using 1M real images, supplemented by 250K synthetic images, and it was denoted ARViT-250K. The remaining 3 models were pre-trained on 1M real images plus 500K, 750K and 1M synthetic images, and denoted ARViT-500K, ARViT-750K and ARViT-1M, respectively.

Models were trained over 90 epochs using the adam optimizer and a base learning rate of 0.0001, with batch size of 80, on a NVIDIA GEFORCE RTX 4090 GPU, with a capacity of 24 Gb. The pre-trained models were then evaluated on 4 benchmark small-scale downstream classification tasks. They are: CIFAR10, CIFAR100, IMAGENETTE and IMAGEWOOF.

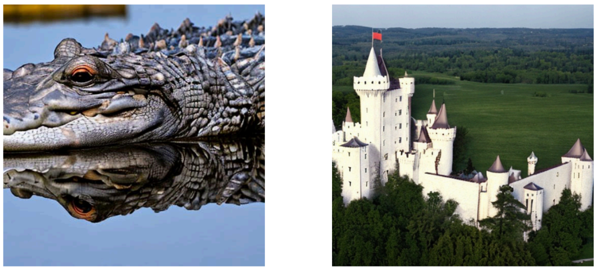
4. Results

Experiment results indicate that, on self-supervised pre-training routine, augmenting a real dataset with imaged synthesised by a stable diffusion model do increase the models ability to generalize when fine-tuned on unseen data.

On all cases, the more synthesised data added to the training set, the higher the absolute accuracy attained on downstream tasks. The top-1 accuracy of all 5 models are depicted on Table 1. When fine-tuned on CIFAR10, ARViT-1M surpassed the baseline model's accuracy in roughly 2%. Similar results were observed on CIFAR100 and IMAGEWOOF. And on IMAGENETTE, gains closing on 3% were attained. This observation highlights the potential and consistency of this method.

Table 1 Top-1 accuracy of each model on 4 different downstream classification tasks: CIFAR10, CIFAR100, IMAGENETTE and IMAGEWOOF.

Model	CIFAR10	CIFAR100	IMAGENETTE	IMAGEWOOF
ARViT-Base	83.13	76.27	91.18	73.01
ARViT-250K	83.91	77.04	91.80	73.46
ARViT-500K	84.32	77.12	92.12	74.24
ARViT-750K	85.05	77.96	93.75	74.81
ARViT-1M	85.20	78.14	94.01	74.84



American Alligator, Small, Gray, Left-Side View, Twilight

Castle, Big, White, From Above, Daytime

Fig. 2 Example of two text prompts e the images generated from it. The randomly selected category is shown in red, followed by also randomly sampled words for size, color, angle/view and illumination. The reflection of the alligator's body on the water highlights the quality of the images produced from this prompt.

5. Conclusion

In conclusion, our investigation focused on the challenges of acquiring high-quality data for machine learning, emphasizing the potential of Stable Diffusion models to generate synthetic data for pre-training discriminative models.

Our study introduced a simple yet consistent framework, employing self-supervised pre-training of the vision transformer ARViT with real data augmented with synthetic images generated through a Stable Diffusion model. Results showed that, by augmenting the training dataset, even with synthetic data from a disparate distribution, improved model accuracy across four downstream classification tasks was observed. While this breakthrough promises cost-effective and efficient model training, it underscores the importance of nuanced strategies to address potential distributional discrepancies between real and synthetic data. Future research will further investigate augmentation of real datasets as well as the substitution of real datasets with synthetic data.

References

[1] Y. Zhou, H. Sahak, and J. Ba, "Using synthetic data for data augmentation to improve classification accuracy," *https://openreview.net/forum?id=42xAKgIb2P*, 2023.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "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.

[3] R. He, S. Sun, X. Yu, C. Xue, W. Zhang, P. Torr, S. Bai, and X. Qi, "Is synthetic data from generative models ready for image recognition?," *arXiv preprint arXiv:2210.07574*, 2022.

[4] H. Bansal and A. Grover, "Leaving reality to imagination: Robust classification via generated datasets," *arXiv preprint arXiv:2302.02503*, 2023.

[5] B. Zhao and H. Bilen, "Synthesizing informative training samples with gan," *arXiv preprint arXiv:2204.07513*, 2022.

[6] J. Yuan, F. Pinto, A. Davies, A. Gupta, and P. Torr, "Not just pretty pictures: Text-to-image generators enable interpretable interventions for robust representations," *arXiv preprint arXiv:2212.11237*, 2022.

[7] A. Borji, "How good are deep models in understanding the generated images?," *arXiv preprint arXiv:2208.10760*, 2022.

[8] L. H. Mormille, C. Broni-Bediako, and M. Atsumi, "Regularizing self-attention on vision transformers with 2d spatial distance loss," *Artificial Life and Robotics*, vol. 27, no. 3, pp. 586–593, 2022.

[9] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," *arXiv preprint arXiv:2202.00512*, 2022.