Semantic-Guided Generative Image Augmentation Method with Diffusion Models for Image Classification

AI6103 - Deep Learning Project

Ho Chek Hui[†]
G2303025C
S230075@e.ntu.edu.sg

Adnan Azmat[†]
G2303265K
ADNAN002@e.ntu.edu.sg

Aradhya Dhruv[†]
G2303518F
AR0001UV@e.ntu.edu.sg

Maheep Chaudhary[†]
G2303665G
maheep001@e.ntu.edu.sg

Abstract

Our project aims to implement and evaluate the Semantic-Guided Generative Image Augmentation Method with Diffusion Models (SGID) for image classification, as proposed by Bohan Li, Xiao Xu and Xinghao Wang [3]. **SGID introduces a novel approach** to image augmentation, leveraging semantic guidance through image labels and captions to generate diverse and semantically consistent augmented images.

Motivated by the limitations of existing augmentation methods, particularly in balancing image diversity and semantic consistency, SGID represents an attempt to address this challenge. Our project focused on replicating and assessing SGID's effectiveness against baseline methods on various datasets and backbone architectures.

Anticipated contributions include a detailed performance analysis of SGID, supporting the authors' claim of its superiority. The iterative nature of our implementation has prioritized key components, ensuring a balance between project scope and available resources. The project's overarching goal is to contribute insights into the field of generative image augmentation for image classification.

Introduction

The demand for extensive labeled data in deep learning has attracted significant attention in recent years, as the quantity of training samples plays a pivotal role in unlocking the potential of deep neural networks . The manual collection and annotation of large-scale datasets are, however, resource-intensive efforts, prompting the exploration of Data Augmentation (DA) methods. These methods, integral to various applications like image classification, aim to diversify training datasets efficiently.

[1]† equal contribution

In the context of image classification, DA methods traditionally fall into two categories: perturbation-based and generation-based. Perturbation-based methods introduce predefined modifications to original images, such as erasing or mixup, ensuring local differences while limiting diversity. Conversely, generation-based methods, employing generative models like diffusion models, synthesize diverse augmented images directly from label-related captions and/or original images. However, these methods often compromise semantic consistency.

This paper addresses the critical challenge of maintaining a fine balance between image diversity and semantic consistency in generative image DA methods. Existing approaches struggle to systematically explore this balance, leading to potential shortcomings in performance on image classification tasks. In response, the research proposes SGID, a Semantic-guided Generative Image Augmentation method with Diffusion models, designed to ensure both semantic consistency and image diversity. It leverages generative baselines, SGID+DiverseCaption and SGID+InstructPix2Pix, to guide augmentation using the essential semantics of original images.

Methodology

Our methodology strictly follows the implementation details of the Semantic-Guided Generative Image Augmentation with Diffusion Models (SGID) research paper. We use the BLIP model [4] for caption generation, employing both nucleus sampling and beam search for diverse captions. The CLIP model [5] calculates the similarity between the original image and generated captions. We focus our analysis on two datasets, OxfordPets and Flowers102. We adopt a prompt weighting strategy, assigning a weight of 1.50 to labels and 0.90 to captions, striking a balance between semantic consistency and image diversity. Additionally, we incorporate a guidance mapping function

$$f(s^*) = -4 \cdot (s^*)^2 + 2 \cdot s^* + 1$$

This function is applied to the similarity score (s^*) between the original image and the chosen caption. It is used to control the extent of semantic guidance on image generation, where higher similarity scores result in higher guidance scales.

To assess the performance of SGID, we evaluated it on classification tasks against augmentation methods such as CutMix [7] and RandAugment [1]. Through these comprehensive analyses, our methodology seeks to understand and leverage the strengths of SGID in achieving a nuanced balance between image diversity and semantic consistency for improved image classification.

Generation of Augmented Images

The generation of images includes several steps and an example of how Figure 2 is generated is listed below :

1. **BLIP Caption Generation:**
   - BLIP (Bootstrapped Language-Image Pre-training) model - specifically Salesforce/blip-image-captioning-base from HuggingFace - is used to produce captions for each image.
   - Two caption-generating algorithms are used: nucleus sampling and beam search, and a total of 10 captions for each algorithm were generated. Nucleus sampling produces more diverse and startling captions, whereas beam search produces safer and more common captions. One of the captions generated for Figure 2 using beam search is "a fluffy white dog laying in the grass with a stuffed animal" while another caption generated using nucleus sampling is "white furry puppy dog".
2. **Calculation of Similarity Using CLIP:**
   - In this step, the CLIP model (openai/clip-vit-base-patch32) is applied to determine the degree of similarity between the original image and each generated caption. The caption with the highest similarity score will be used in the prompt to generate the image.
   - The similarity score indicates how well the caption content fits the content of the source image. Similarity score for the caption above generated by beam search is 0.396 while the score for the caption above using nucleus sampling is 0.301.
3. **Prompt Construction:**
   - Both the label of the image and the caption generated are used to form the prompt for image-to-image generation. The prompt used for Figure 2 is "a fluffy white dog laying in the grass with a stuffed animal. A picture of Great Pyrenees".
   - The Stable Diffusion model [6] uses the prompt from the above step to guide the augmentation process.
4. **Prompt Weighting:**
   - In this step, assign labels a weight of 1.50 and the caption a weight of 0.90. This is done to counteract the effects of semantic guidance on image diversity and consistency.
5. **Guided Mapping Function:**
   - A guidance mapping function defined as in the original paper is used with s* being the similarity score for the caption.

$$f(s^*) = -4 \cdot (s^*)^2 + 2 \cdot s^* + 1$$

   - The guidance scale determines how closely the image production process adheres to the semantic guidance offered by the textual prompt.
6. **Stable Diffusion Enhancement:**
   - For image augmentation, use the pre-trained "stable-diffusion-v1-5" model [6] and create augmented photos via image-to-image using the textual prompt, guidance, and noise rate.
   - Next, Gaussian noise of 0.3 is used to create minor modifications to the original image while adhering to the semantic requirements.

Figure 2: Comparison of original and SGID generated image

Experimentation and Results

Pre-processing of images

Using the methodology described, realistic-looking images could be generated that still retain the style, background, and camera angle while introducing some variation from the original image as shown in Figure 3. However, we observed a significant number of images are poorer in quality and the style of the image completely changed to an abstract art form. Through error analysis, we discovered that this is prevalent in images of smaller resolution which may not work well on the stable diffusion model which was trained on 512 x 512 images. By upscaling images of smaller resolution, the quality of the generated images improved. In addition, 3 images are generated per original image and human-in-the-loop approach is used to select the best image out of the 3. An example is seen in Figure 4 where an image is generated before and after using the steps described above, with a improvement in cosine similarity between the original and generated image by around 19x.

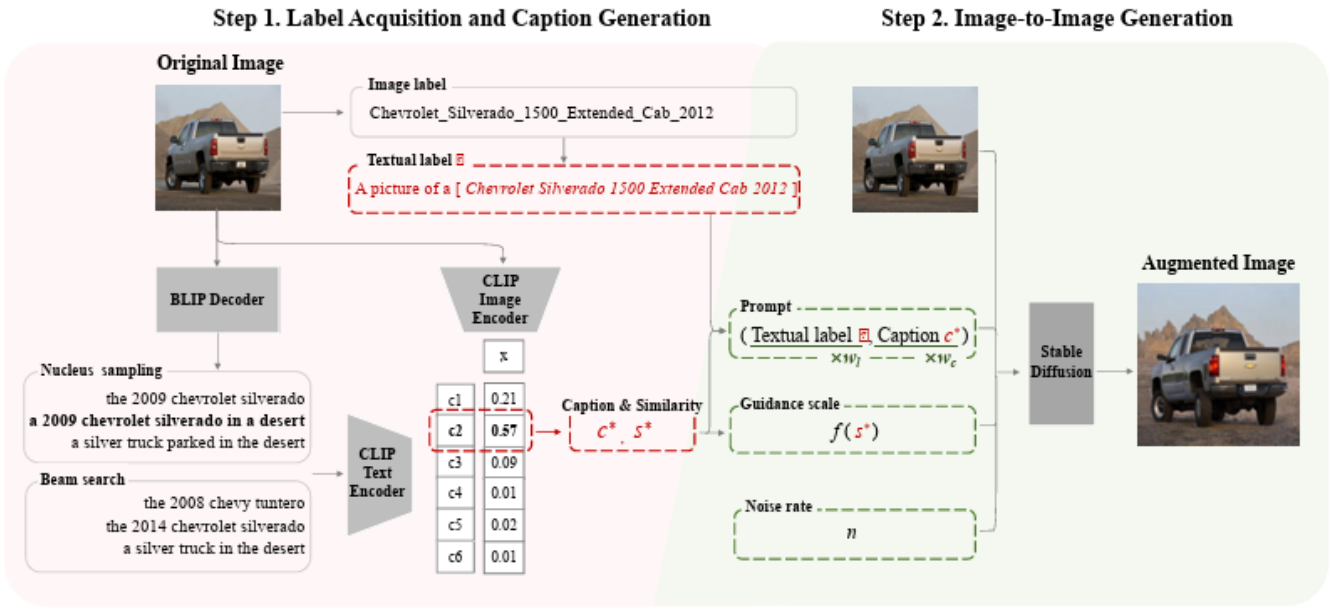

Figure 1: SGID Architecture

Figure 3: Generated image with cosine similarity of 0.756

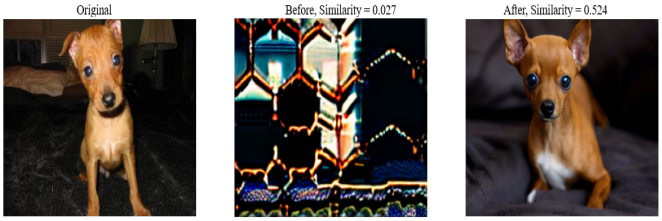

Figure 4: Comparison of images after pre-processing

Comparison of Augmentation Techniques and Model Training

In this section, we evaluate the impact of various augmentation techniques on the classification accuracy of our model. We employ the pre-trained 'vit-base-patch16-224-in21k' model [2] and fine-tune it on the dataset under consideration. We use the Oxford Pets and the Flowers dataset for our evaluation.

All experiments utilize the Adam optimizer with a constant learning rate of 0.0001. The augmentation techniques compared are as follows:

- **Baseline:** No augmentations are applied during the fine-tuning process.
- **CutMix:** We adopt the official CutMix-PyTorch implementation by Yun et al. 2019, with $\alpha = 1$ and a CutMix probability of 0.5.
- **RandAugment:** We utilize the RandAugment transformation provided by PyTorch, retaining the default parameters (N = 2, M = 9).
- **SGID:** This augmentation technique involves supplementing the training set with additional images generated via SGID. We experiment with varying proportions of SGID images to assess their impact.

For SGID, different proportion of generated images (25% till 100%) were also added to the training set to analyze its effects of the model performance. The same training set-up and hyperparameters were used as above. The figure 5 depicts a comparison of different augmentation techniques against SGID.

Performance Metrics

In Tables 1 and 2 we have mentioned accuracy achieved with our code implementation V/S the benchmarks which are actually mentioned in the research paper by the authors. In the subsequent sections of the report we have critically analyzed the reasons behind the observed results as well.

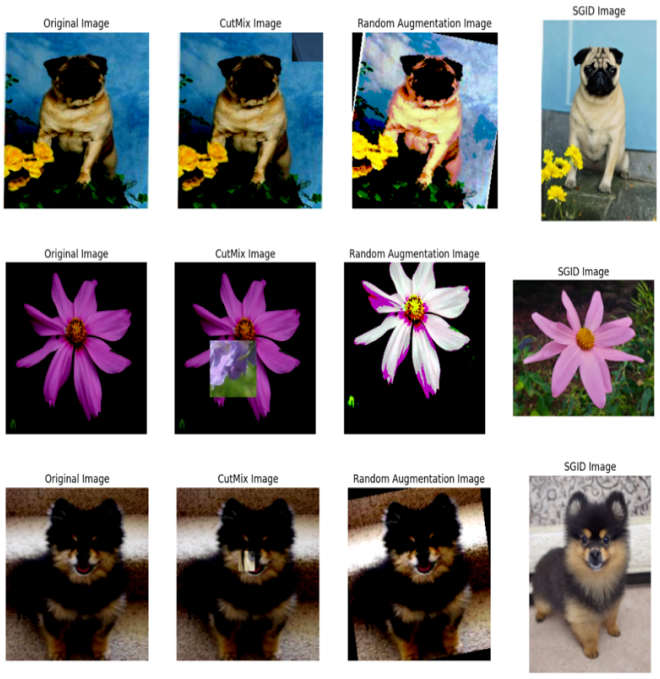

Figure 5: Visual comparison of different augmentations

	Pets - Ours	Benchmark in Paper
Baseline	0.9209	0.9204
Cutmix	0.9199	0.9245
RandAugment	0.9220	0.9291
SGID(25%)	0.9049	-
SGID(50%)	0.9049	-
SGID(75%)	0.9062	-
SGID(100%)	0.9000	0.9338

Table 1: Classification accuracy of different augmentation techniques on Pets dataset

	Flowers- Ours	Benchmark in Paper
Baseline	0.9822	0.9570
Cutmix	0.9847	0.9653
RandAugment	0.9886	0.9666
SGID(25%)	0.9804	-
SGID(50%)	0.9603	-
SGID(75%)	0.9490	-
SGID(100%)	0.9219	0.9717

Table 2: Classification accuracy comparison of different augmentation techniques on Flowers dataset

Critical Analysis

During our experimentation, a noteworthy observation was that the inclusion of SGID did not give any improvement, instead it led to a deterioration in the results. This surprising discovery makes us think that our model might be overfitting, making it struggle when dealing with real images in the test set.

The paper we're following for our experiment gives a lot of details, but there are still some things not mentioned, like the specific values used for certain elements. This makes it tricky for us to exactly replicate the setup. Also, there is variability in nucleus sampling and beam search which produces different captions and hence we obtain different results for the same image. As such, perfect reproducibility of the original paper is difficult as seen in Figure 6

Even though the paper provides extensive implementation details, there might be some gaps. There is variability and randomness in the steps listed in the paper, thus it doesn't give the exact same results. Looking at the pictures in the original paper, we see that after using SGID, the background changes, but the main subject of the image stays the same or changes slightly. In our results, though, we're seeing bigger changes in the both subject as well as background. We were able to resolve this issue by adjusting the weights for captions and labels when we're creating the prompts for stable diffusion.

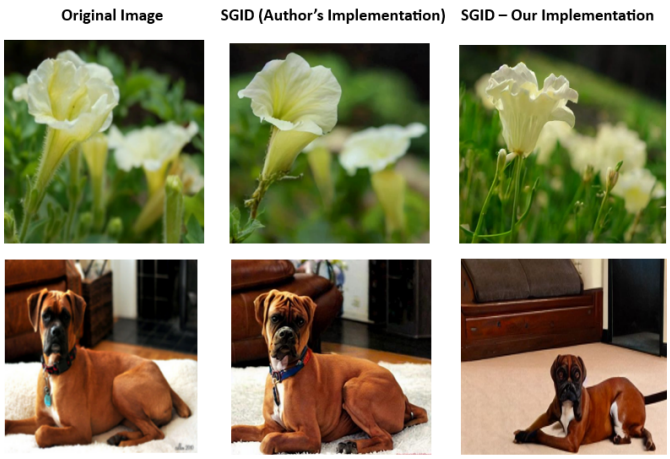

Figure 6: Comparison of our SGID Images V/S Author's implementation

Shortcomings in Traditional Perturbation Based Augmentation:

- **Limited diversity:** Perturbation-based methods often only introduce predefined modifications to original images, such as flipping, cropping, and rotating. This can limit the diversity of the augmented images and may not be sufficient to prevent overfitting.
- **Local variations:** Perturbation-based methods focus on introducing local changes to the image, which may not capture the global semantic structure of the image. This can lead to augmented images that are semantically inconsistent with the original image.

Shortcomings in Proposed SGID Model:

Although, SGID generates good quality images which can potentially improve classification model accuracy, it also comes with certain limitations which are enumerated below:

- **Increased computation cost:** Generative methods like SGID are generally computationally more demanding than perturbation-based methods. This is because they require training and running a generative model, which can be time-consuming and resource-intensive. Moreover, SGID requires 3 different models (BLIP, CLIP, and stable diffusion) to generate a single image.
- **Limited control over augmentation:** While SGID provides guidance to the generative model through prompts and captions, there is still a degree of randomness involved in the augmentation process. This can make it challenging to precisely control the types of augmentations that are generated. However, there are some progress in the field of making controllable generation, with one example being ControlNet [8]. This could improve the reproducibility of generated images.

Possible Improvements:

- To ensure consistency of generated images, additional images could be generated from one image and cosine similarity can be used to compare the generated and original image. If the similarity does not meet a threshold value, the generated image is discarded. By doing so, one can ensure that generated images are similar and reduce the amount of noise from these images during training.
- To reduce overfitting to generated images, another image classifier could be trained to predict whether the image is real or fake from the sample of training and generated images. Only the generated images which the classifier predict as False would be added to the extended training dataset, rather than all the generated images indiscriminately. In this case, only the generated images that look authentic enough or similar to the distribution of real images would be used for training.

Conclusion

While we expected that SGID would enhance the performance of our image classification model, our experiments revealed that it actually hindered the model's accuracy. This unexpected outcome suggests that the model may have become overly reliant on the augmented images generated by SGID.

Further investigation is needed to understand the underlying reasons for this overfitting phenomenon and to determine potential strategies to mitigate its effects. While SGID holds promise as an augmentation technique, careful optimization is necessary to ensure its effectiveness in practical applications since the computation costs involved are quite high. However, it can prove to be quite beneficial in scenarios where peak accuracy is desired or beating a benchmark is more crucial.

References

[1] Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2019. RandAugment: Practical automated data augmentation with a reduced search space. *arXiv preprint arXiv:1909.13719*.

[2] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[3] Li, B.; Xu, X.; and Wang, X. 2023. Semantic-Guided Generative Image Augmentation Method with Diffusion Models for Image Classification. *arXiv preprint arXiv:2302.02070*.

[4] Li, J.; and Li. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*. Submitted on 30 Jan 2023.

[5] Radford, A.; and Kim. 2021. Clip: Learning Transferable Visual Models From Natural Language Supervision. In *arXiv preprint arXiv:2103.00020*. Submitted on 26 Feb 2021.

[6] Robin Rombach, D. L. P. E. B. O., Andreas Blattmann. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv preprint arXiv:2112.10752*.

[7] Yun, S.; Han, D.; Oh, S. J.; and Chun, S. 2019. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6023–6032.

[8] Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.05543*.