Fighting Class Imbalance with Textual Inversion

Anton Kutsenko, Bahareh Najafi

December 2024

1 Abstract

The main task of this project is to use Textual Inversion to generate high-quality synthetic images for underperforming or underrepresented classes from the Oxford 102 Flower Dataset and utilize them to improve classifier performance. The focus of the analysis is on the Textual Inversion part and generation quality assessment. The results of our work show that Textual Inversion can help to generate high-quality data for the given concept and can be useful for classification tasks with small amount of classes.

2 Introduction

Class imbalance often leads to poor performance for underrepresented classes, while usual data augmentation techniques are very limited. In recent applications, generative techniques showed their ability to make these data augmentations more diverse, but still relevant to the problem thanks to guidance possibilities. For the chosen Oxford Flower dataset the usage of the Textual Inversion technique with a pre-trained Stable Diffusion model can help to generate diverse and high-quality synthetic images for specific underrepresented and underperforming classes. In this project, we employ Textual Inversion to learn special embeddings that will help to address the stated issue and also analyze generation quality.

3 Related work

Our project relies on the techniques that were proposed in recent papers. Textual Inversion (Gal et. al, 2022) is the main concept we explore. The authors proposed an approach to fine-tune a pre-trained text-to-image model for learning a special token to represent new concepts. This idea exploits pre-trained model capabilities with lightweight and fast training with a few image examples. This makes it relevant to our problem.

Since we use the Stable Diffusion model it is important to introduce the concept of the Latent Diffusion Model (Rombach et. al, 2022). The main idea is to switch from an expensive Diffusion process in the image space to a latent, compressed space, using a Variational Autoencoder. This architecture keeps all the main advantages of previous Diffusion Models while being more efficient.

The last concept we explore is CLIP from Radford et. al (2021). CLIP is a model that aligns text and image embeddings in a shared space, enabling text-to-image matching. This is a very important part of Textual Inversion as we will try to learn new embeddings that have to match our concept from images.

These three main papers in combination with the HuggingFace diffusers library will be the core of our work.

4 Dataset and Metrics

The Oxford 102 Flower Dataset consists of 102 flower species, with a total of 8189 images. The dataset is split into:

Training set	**Validation set**	**Test set**
1020 images (10 per class)	1020 images (10 per class)	6149 images

Evaluation Metrics for the project are:

1. **Frechet Inception Distance (FID)** - main metric to evaluate the quality of generated images by comparing the distribution of features extracted from a pre-trained Inception network for both real and generated images.
2. **Precision, Recall, F1-score** - to evaluate classification results and choose classes for Textual Inversion

5 Proposed Method

The proposed methodology consist of several main steps:

1. Identify target classes for Textual Inversion. We fine-tuned the ResNet-50 model and checked classwise performance to select target classes. We tried to use classes that are not performing well and at the same time represent rare flower concepts (to make Textual Inversion relevant to the problem)

2. Select Images for Textual Inversion training. Selected images represent train data for the chosen class.
3. Train Embedding for special "<token>" representing the class using Textual Inversion.
4. Sample Images: Generate new images and evaluate their quality. Several assessments were made: comparison of the trained token embedding similarity with already existing tokens from vocabulary, FID calculation between real images, images from Stable Diffusion before and after Textual Inversion. More on this in the Experimental results section.
5. Use generated images to retrain/finetune the classifier with augmented data.

We also provide high-level description of the main step - Embedding training with Textual Inversion technique. We start with a pre-trained Stable Diffusion model that uses textual prompts to guide Diffusion process. Then We define a special token (e.g., "<flower_class_name>"), which will be associated with our concept of the flower class. This token is added to the model's vocabulary. The model is fine-tuned on our images from the underrepresented class, while all the weights except trained token embedding are kept frozen.

6 Results and findings

6.1 Classification results

Our main findings from the classification part gave us an understanding that there are some limitations to our approach with the given data. We realised that the dataset was very easy due to its good quality. We achieved ~ 0.8 validation accuracy/F1 after 5 epochs of fine-tuning a fully connected layer of ResNet-50 (fig. 1). Considering such metrics there was no need for advanced data augmentation. Also, training one concept with Textual Inversion takes around 2 hours with 2000 training steps, so we could train only 3 tokens for the dataset with 102 classes, which is a small number to improve classification metrics. As a result, we decided to focus mainly on the generative part and its evaluation.

6.2 Raw Model Generation

Initially, we explored using Stable Diffusion without Textual Inversion to generate synthetic images, which we will subsequently call the raw model generations. In this setup, the model was given a simple prompt "Photo of a "<flower_class_name>"". Since many of the flowers are well-known, the model was familiar with them since initial training (fig. 2).

However, several issues arose during the generation process. The generated images had significant variability, which made them unsuitable for augmenting the given dataset - Oxford flowers have high-quality close photos inside. Moreover, some flower names were not recognized by the model or were known by different names, leading to inaccurate image generations. For example, "Sword Lily" was known as "Gladiolus" (fig. 3) and "Flower Prince of Wales Feathers" was getting good generations only under the official species name "Amaranthus hypochondriacus".

These results indicate that the raw model generation can be used for some augmentations, but is not very relevant for rare and hard concepts.

6.3 Textual Inversion and comparisons

To address the limitations of the raw model generation, we utilized Textual Inversion to train specific embeddings for several flower classes. We trained three tokens for different flowers: azalea, sword lily and canna lily.

Two different initial placeholder tokens were used for training these concepts. for Sword Lily and Canna Lily we started with the token "flower" as the model has no good representation for them, while for Azalea, the embedding was trained from the token "azalea" directly as the model had it in the vocabulary. Also, the first approach is more fair as often we want to train concept that we don't know exact name for.

To evaluate the quality we made a comparison between images generated from raw Stable Diffusion, after Textual Inversion, and the test and validation images from the original dataset. The results showed the following:

The images generated using Textual Inversion embeddings showed a reduction in appearance variability and were closer in visual characteristics to the original images from the dataset (fig. 4). FID between real images and images generated using trained tokens was significantly lower in comparison with raw generations.

	Azalea	**Sword lily**	**Canna lily**
FID raw - real	198.1	206.6	210.3
FID trained - real	**127.7**	**146.6**	**161.9**

We also compared the trained embeddings for each of the three tokens with existent embeddings from the CLIP vocabulary. For all three tokens, the closest embeddings were the other two due to the similarity of training data: close and high-quality photos of flowers. For the "Azalea" trained token original "azalea" token was close as it was an initial token, while the initial "flower" token that we started with for the other two approaches was not similar. Additionally, other plant-related tokens such as "planted", "hibiscus" or "flourish" were found to be close to the trained tokens, showing that trained embeddings represent some plant-related semantic features.

7 Conclusion and Future work

The use of Textual Inversion allowed for a more controlled and accurate generation of synthetic images for flower classes. By training specific embeddings for each flower, we improved the consistency and quality of the generated data, leading to better augmentation of the Oxford Flower dataset.

However, our project has limitations in some parts. We realised that fighting class imbalance with Textual Inversion is a good approach if dataset classes represent rare and very specific concepts - otherwise raw model can be used. Moreover, it is useful only if classes do not have much variability of object appearance and the number of underrepresented classes is small (i.e. binary classification).

Future work on the topic might include several perspectives. It can be interesting to compare results not only with the target class but with all the classes from the chosen dataset to see if generated images represent the target class better. Additionally, improvements in prompt quality may increase diversity and help to find the perfect image structure for the given dataset. The other idea is to train tokens starting from empty or random initial tokens but for a longer time to give the model more freedom. Also, the use of other techniques like LoRA or DreamBooth can be beneficial.

For more detailed results and visualizations refer to the project repository.

References

[1] Gal, R., Alaluf, Y., Patashnik, O., Bermano, A. H., Maron, H., & Cohen-Or, D. (2022). An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. `https://textual-inversion.github.io/`

[2] Hugging Face. Textual Inversion with Diffusers Library. `https://huggingface.co/docs/diffusers/v0.3.0/en/training/text_inversion`

[3] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*. `http://arxiv.org/pdf/2103.00020`

[4] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv preprint arXiv:2112.10752*. `https://arxiv.org/pdf/2112.10752`

8 Appendix

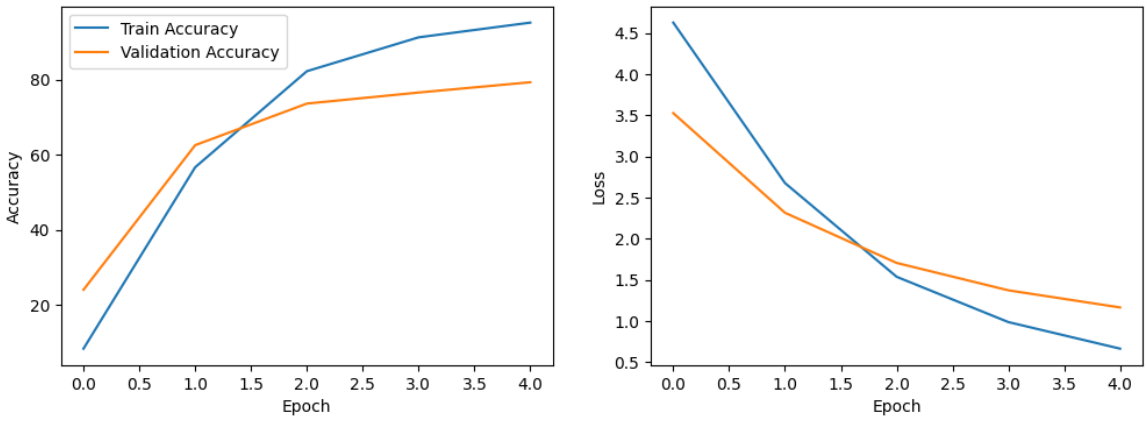

Figure 1: Accuracy vs. loss plot for classification

Figure 2: Raw model generations

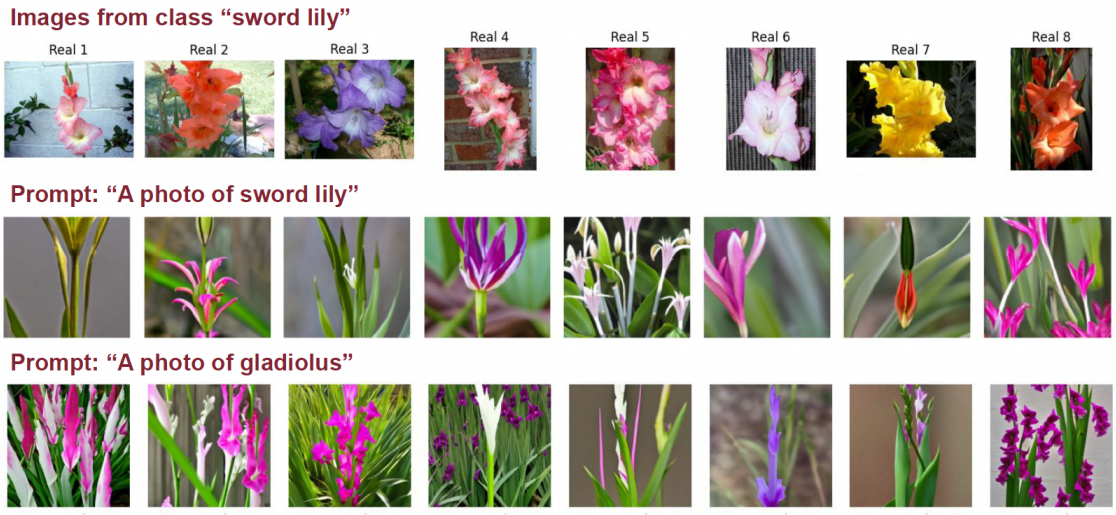

Figure 3: Prompts comparison for sword lily with raw model.

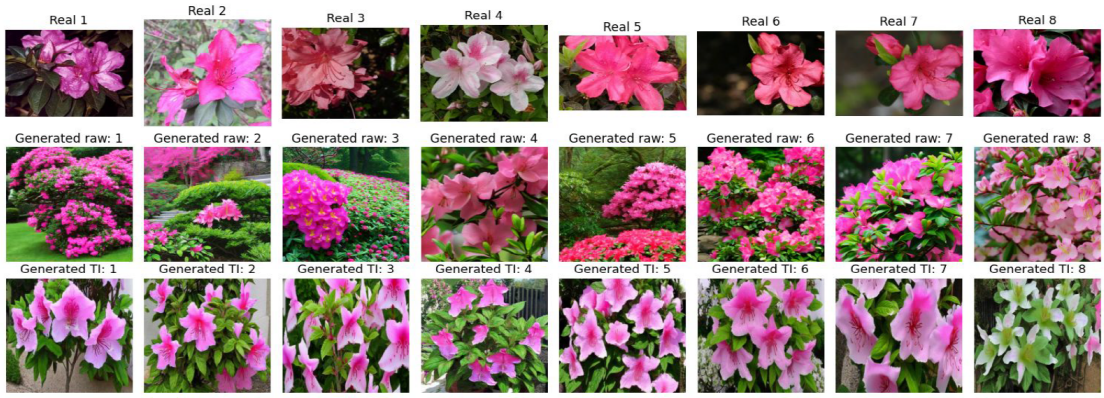

Figure 4: Azalea: real-raw-trained comparison