Personalizing Text-to-Image Diffusion Models by Fine-Tuning Classification for AI Applications

Conference Paper · September 2023

CITATIONS
0

READS
1,864

5 authors, including:

Rafael Hidalgo
Montclair State University
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Nesreen Salah
Montclair State University
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Rajiv Chandra Jetty
Montclair State University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Anupama Jetty
Montclair State University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Personalizing Text-to-Image Diffusion Models by Fine-Tuning Classification for AI Applications

Rafael Hidalgo, Nesreen Salah, Rajiv Chandra Jetty, Anupama Jetty, and Aparna S. Varde

Department of Computer Science, Montclair State University, Montclair, NJ, USA
(hidalgor2 | salahn1 | jettyr1 | jettya1 | vardea)@montclair.edu

Abstract. Stable Diffusion is a captivating text-to-image model that generates images based on text input. However, a major challenge is that it is pretrained on a specific dataset, limiting its ability to generate images outside of the given data. In this paper, we propose to harness two models based on neural networks, Hypernetworks and DreamBooth, to allow the introduction of any image into Stable Diffusion, addressing versatility with minimal additional training data. This work targets AI applications such as augmenting next-generation multipurpose robots, enhancing human-robot collaboration, feeding intelligent tutoring systems, training autonomous cars, injecting subjects for photo personalization, producing high quality movie animations etc. It can contribute to AI in smart cities: facets such as smart living and smart mobility.

Keywords: ANN, Data Mining, Image Processing, Movie Animations, Photo Personalization, Stable Diffusion, Text-to-Image Creation

1 Introduction

In this paper, we address data mining in text-to-image generation via the paradigm of *Stable Diffusion* with fine-tuning using architectures based on artificial neural networks (ANN). It adds real-life perspectives to the images created (see Fig. 1) and can be useful in various AI applications such as movie animations.

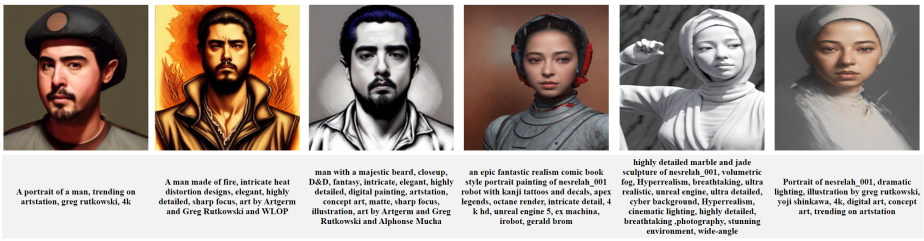

Fig. 1: Images created by Stable Diffusion after fine tuning via Hypernetworks (3 left) and DreamBooth (3 right); prompts to generate each image are below it.

From credit card fraud detection to TikTok, artificial neural networks are being used all around us. As is well-known today, an ANN is a machine learning paradigm modeled after the human brain. It is composed of many interconnected processing nodes, called neurons, which work together to process input data and make predictions or decisions based on that data. One of the main areas for neural networks usage is classification, i.e. the process of analyzing input data to estimate a target output as being in one of several predefined classes or categories. ANNs are particularly well-suited for tasks involving images.

A very interesting technology using ANNs for image classification and creation is the diffusion model. Specifically, we refer to *Stable Diffusion*, developed by Stability AI, CompVis LMU et al. [3] used for text-to-image creation. While it is a popular technology [10], [12], it incurs challenges [7] including the need for: exhaustive training data to generate images outside pretrained datasets; exploration of real-life applications especially with good social impact; and more diversity & inclusion (D&I) as per country, ethnicity etc.

Given this motivation and challenges, we address the problem in this paper, on "thinking outside the box" more specifically defined as follows.

- Encompass numerous real-life perspectives into Stable Diffusion
- Create new images with minimal training data, yet addressing versatility
- Incorporate more diversity & inclusion for global mass appeal

We propose a solution to this problem by exploring two advances, namely, Hypernetworks and DreamBooth in conjunction with Stable Diffusion that can be adapted to work with low volumes of training data while still producing a robust set of images in various contexts. We add real-life angles to this work by considering numerous facets for generating the images, and outlining several targeted applications. Furthermore, we take into account D&I from a global perspective, considering multiple countries, religions, subject names and ethnicity, aiming to stay away from stereotypes solely based on Western cultures (oft found in image searches). Our work yields high levels of user satisfaction, as evident from the experimentation.

While there is much literature in the area [10], [29], [12], to the best of our knowledge, our paper is early work on exploring Stable Diffusion with Hypernetworks and DreamBooth, contextualizing them with multiple real-life perspectives, using relatively less new training data, and leveraging D&I. This constitutes the novelty of our paper, contributing modestly to neural models and image processing, helping applications such as photo personalization and movie animations, and broadly making impacts on AI in smart cities, vis-a-vis smart living and other facets. We present the details of our work in the forthcoming sections.

The rest of this paper is organized as follows. Section 2 explains the models and methods harnessed in our work, namely, Stable Diffusion, Hypernetworks and DreamBooth. Section 3 describes the implementations of our proposed approaches along with algorithms. Section 4 synopsizes our experimental evaluation, presenting a discussion as well. Section 5 overviews related work in the area, placing our own work in context, and emphasizing its novelty. Finally, Section 6 states the conclusions and outlines prospective avenues for future research.

2 Models and Methods

In this paper, we focus on the paradigm of Stable Diffusion. We aim to add more real-life perspectives to the images in text-to-image creation via minimal additional data. This is explored via recent advances: Hypernetworks and DreamBooth. We discuss these main models, and the methods we apply on them.

2.1 Stable Diffusion

Considering the fundamental concept of diffusion models, we begin the process through a forward diffusion process using Markov chains (Fig. 2). Through Markov chains, we take an image and add some Gaussian noise to it. This step can be repeated an infinite number of times, but is usually terminated when enough noise is added to the picture such that it is not recognizable any more. Each image in the Markov chain from start to end can then be used to train a convolutional neural network (CNN) called U-Net to denoise the image [17, 28].

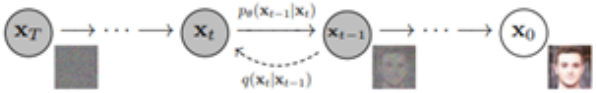

Fig. 2: Denoising Image Via Markov Chains [17].

Specifically U-Net is trained by receiving a noised image along with the number of iterations of noise applied to the noised image. Thereafter U-NET is instructed to calculate the total noise of the image so as to revert the image back to its original form. Depending on how well U-NET performs, it receives a negative or positive reward (in line with reinforcement learning). Eventually U-NET is able to calculate noise removal in all iterations of the Markov chain noising process and is trained to do this in a single step [17, 28].

However, many errors can occur if the network is allowed to remove noise in a single step, so the neural network instead is made to calculate all of the noise it believes is polluting an image, but only removes a fraction of that noise, just enough to undo the noising process by one iteration. The neural network then repeats this for an image until it reaches iteration 0 of the noising process, after which it yields the original image [17, 28, 12].

A GPT (Generative Pre-trained Transformer) is integrated to the model, enabling it to take text prompts to produce the desired image. After receiving text, the model receives a random noise image and the number of noise iterations the image underwent. The ANN denoises the image as per the process above [17, 28]. In order to ensure that our image is being well denoised, a technique called classifier free guidance (CFG) is applied. In CFG, the image is fed into the neural network twice. The first copy of the image is run with the text embedding, and the second copy without it. Noise is calculated from both the images, and

the difference between them is obtained. This is then used to guide the neural network into refining its denoising process to obtain an image as described by the text prompt [17, 28].

2.2 Hypernetwork Model

Hypernetworks are neural networks used to predict the weights of primary networks. The Hypernetwork consists of a lightweight feature extractor and a set of refinement blocks. Each refinement block is tasked with predicting the weights of a primary network. By training Hypernetworks over a large collection of data, the weights of the primary network are adjusted with specific inputs yielding more expressive models [16, 4]. Fig. 3 depicts a Hypernetwork.

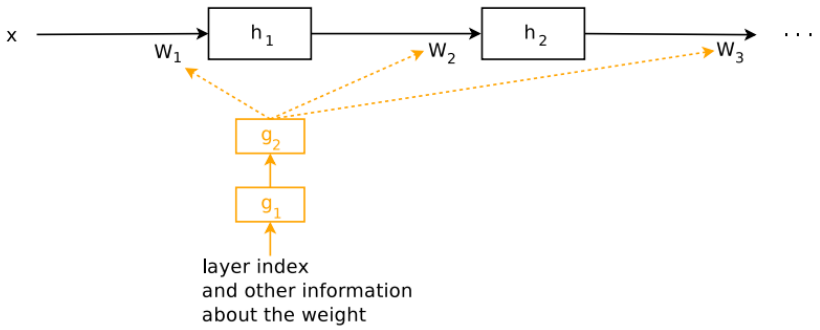

Fig. 3: Diagram of a Hypernetwork [16].

A major advantage of Hypernetworks is that they are tiny, mostly of the order of Megabytes (80 Mb per Hypernetwork). Another advantage is that they are modular and can be attached and detached from the main network as needed. A disadvantage is that the Hypernetworks change all the weights of the main networks. Hypernetworks have been deployed to applications such as 3D modeling, semantic segmentation, neural architecture search, continual weights and so on [10]. We therefore use them in conjunction with Stable Diffusion. The details are described in the next section on implementation.

2.3 DreamBooth Model

DreamBooth is a deep learning text-to-image diffusion model meant to fine-tune larger models. Prior to this, there existed large text-to-image models but they lacked the ability to generate realistic pictures of the subjects in the reference set, whereas in DreamBooth we find a new approach for "personalization" of text-to-image diffusion models. Using just a few pictures of the subject given as inputs, it fine tunes the pretrained text-to-image models (see Fig. 4) [14, 30].

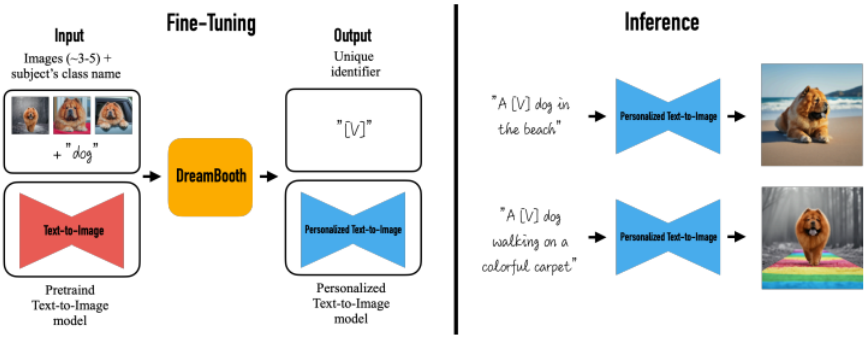

Fig. 4: Illustration of DreamBooth [30].

It operates in two steps. In the first step, it generates a low-resolution image from text-to-image models with input images and text-prompts containing unique identifiers, followed by the class name of the subject, to prevent overfitting and language drift (that causes the model to associate with the class name). In the following step, it fine-tunes the super-resolution component using the input images, and generates low-resolution as well as high-resolution versions of the input images [14, 30].

This allows the model to maintain high fidelity to small, but important, details of the subject. DreamBooth binds a unique identifier to the specific subject. Once the subject is embedded in the output domain of the model, it uses a unique identifier to generate fully novel photo-realistic images of the subject contextualized in different scenarios such as various poses, views, and lighting conditions that do not appear in the reference images. We can apply this technique to perform different tasks such as subject re-contextualization, text-guided view synthesis, appearance modification, and artistic rendering, while maintaining the key features of the subject [14, 30]. We now present the implementation of our approaches on DreamBooth and Hypernetworks along with Stable Diffusion.

3 Implementation of Approaches

In order to discover knowledge from data during text-to-image creation, as well as to explore a wide range of contexts for applications, we seek to incorporate our own personal images into Stable Diffusion. We use them for a variety of text-to-image generation scenarios. This is achieved by two approaches: adapting Hypernetworks, and deploying DreamBooth, as follows.

3.1 Hypernetwork Implementation

In our Hypernetwork adaptation, 24 photos of a given subject are taken. In the work shown in this paper, the subject is “Rafael Hidalgo” a Latino male student

with a family background spanning multiple countries. The photos are cropped to produce 1x1 pixel images, and converted to 512 x 512-pixel images [8] as required by the model. Fig. 5 depicts a sample of the images used here.

Fig. 5: Pictures for Hypernetwork Model Training.

An open-source Stable Diffusion Web UI (User Interface) developed by a GitHub user named "automatic1111" is installed here [5]. In this Web UI, the user "automatic1111" employs a Gradio library, Python, and PyTorch to use Stable Diffusion models for text-to-image generation. Besides utilizing this Web UI, for text-to-image generation, it is used to train the Hypernetwork. Stability AI, one of the main companies responsible for generating the Stable Diffusion model, has saved their copy of the model in the Hugging Face website, which contains a library of transformers, datasets, and demos of machine learning projects. We download a copy of this Stable Diffusion model onto a personal computer to use with the automatic1111 web UI. [3, 29]. With the Web UI and the model installed, we train the Hypernetwork using 24 images of the subject. First, we preprocess the pictures by running the subject's images through a Bootstrapping Language-Image Pretraining (BLIP) model. It generates relevant captions for the image, so it can associate pictures with the respective image captions.

Once the captions are generated for all images of the subject, the training is processed using a personal computer and 8GB of V-RAM. The Hypernetwork is trained for 10,000 steps. For instance, sample pictures are generated every 100 steps by the prompt "A portrait of a man, trending on artstation, greg rutkowski, 4k". After 10,000 steps, the Hypernetwork is tested via the use of different prompts. More about this appears in the section on experimental results with discussion. Algorithm 1 has the pseudocode on our Hypernetworks execution, and synopsizes its overall processing.

Algorithm 1 Training the Hypernetwork along with Stable Diffusion

Load pre-trained ANN model α
Initialize Hypernetwork HN
Preprocess data δ for subject σ
Use BLIP to generate captions γ for σ
Load modified subject data δ_m
Split δ_m into training set τ, validation set υ
for each iteration i **do**
    Have HN generate weights ω to influence α
    Train α using ω on τ
    Get results ρ from α similar to σ via associated class χ
    Update HN via backpropogation based on ρ
end for
Save the trained HN
Attach HN to α
Generate new subject data δ_n
Output δ_n for σ w.r.t. context

3.2 DreamBooth Implementation

Our deployment of DreamBooth is on similar lines as Hypernetwork, with a few variations. Multiple photos of our subject are taken. The subject in this task is "Nesreen Salah", a female student of Egyptian descent & American upbringing. The photos are adjusted to form 1x1 scale pixel images, and converted to 512 x 512-pixel images as befits the model. Fig. 6 has a snapshot of the images [8].

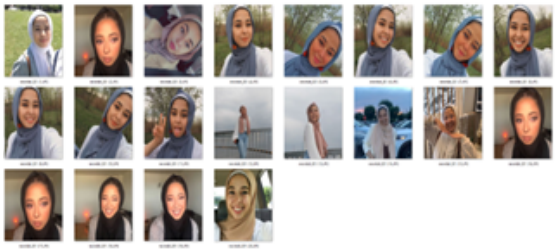

Fig. 6: Images for DreamBooth Model Training.

Analogous to the Hypernetwork, our DreamBooth model is trained on 20 pictures of our subject. The only caveat however, is that the process to train DreamBooth is resource-heavy, and it is recommended that at least 24 GB of V-RAM be available to train DreamBooth [10]. Therefore, we harness *Google Colab* (Google's Research Colaboratory) to procure a computer with the necessary amount of V-RAM. We adapt a DreamBooth implementation from a GitHub user with the username "ShivamShrirao" to train our modified model [31].

We take a copy of Stable Diffusion from the Hugging Face website for this task as well, however, this time we incorporate it into the DreamBooth implementation on Google Colab. Thereafter, we upload our dataset of pictures to the DreamBooth implementation. We then correlate the pictures with an instance-name of our choosing and with a class-prompt that already exists in the Vanilla Stable Diffusion Model. For our instance-name, we need to ensure that it does not exist in Stable Diffusion, so as to avoid confusion with another instance. It is best to stay away from conventional names, as yet another aspect of diversity. Hence, for our subject, we choose the name "nesrelah_001"; and we associate our subject with the class "person" [29].

Once the instance name and class name are established, we train our model using 4,040 steps. In order to generate these pictures, we use the prompt "nesrelah_001 person". After 4,040 steps, we download the new trained model.ckpt file from Google Colab onto a personal computer, to use it with the "automatic1111" Web UI as described earlier. The DreamBooth-modified Stable Diffusion model is tested via various prompts. More on this appears in our experimental results section [5]. Algorithm 2 has our pseudocode for the DreamBooth execution.

Algorithm 2 Training via DreamBooth in conjunction with Stable Diffusion

Load pre-trained ANN model β
Initialize DreamBooth DB
Preprocess data Δ for subject κ
Associate κ with class ζ related to it
Associate κ with unique identifier ι
Load modified subject data Δ_M
Split Δ_M into training set θ, validation set ψ
for each iteration j **do**
    Send input I to β using ι of κ
    Get predicted results ϕ from β similar to κ via ζ
    Calculate loss $\lambda = |\phi - \mu|$ where $\mu =$ actual results
    Use λ to update β via backpropagation
    Update DB using $ArgMin(\lambda)$
end for
Save the trained DB
Attach DB to β
Generate new subject data Δ_N
Output Δ_N for κ w.r.t. context

4 Experimental Results with Discussion

In order to test our models, we mostly use AI art search engines to help us decide the prompts to use. It is exciting to note that with a relatively small dataset, and two fine-tuning implementations, we can significantly harness the

power of Stable Diffusion to create prompts with any subject of our choosing. Hence, this facilitates "thinking outside the box". As seen here, we can put both the subjects on horses, render paintings of the subjects in various art styles, alter the subjects to look like Chibi figurines, and make a subject into a Jade statue. Accordingly, we evaluate our work (which at this point is a pilot study). We assess and compare the two fine-tuning models.

The main difference between these two fine-tuning technologies seems to be the following. In Hypernetworks, the class associated with the subject becomes affected by that very same subject. As can be seen in the prompts used for Hypernetworks, no specific name is selected for the subject. Instead, the subject is just referred to as "man". This essentially enables Stable Diffusion to create images of the subject, but at the cost of making all instances of "man" look like the subject. On the other hand, as can be seen with the prompts generated via DreamBooth, the subject is kept distinct from the class. Yet, the caveat with DreamBooth (as mentioned earlier) is that more V-RAM is needed to process the training, and the model itself is 2 GB large, versus the size of Hypernetwork, which is around 80 MB [10].

Based on our experimentation with Stable Diffusion alone, as well as using it in conjunction with Hypernetworks and DreamBooth, a summary of our results is presented here. The prompts in these experiments are chosen (guided by AI art search engines) as per the operations we aim to execute using Stable Diffusion with Hypernetwork / DreamBooth, anticipating various tasks that targeted users may perform with them.

4.1 Hypernetwork Experiments

The prompts for the Hypernetwork include:

1. "Man riding a horse, facing camera" (Fig. 7).
2. "Portrait of a man, trending on artstation, greg rutkowski, 4k" (Fig. 1)
3. "Man made of fire, intricate heat distortion designs, elegant, highly detailed, sharp focus, art by Artgerm and Greg Rutkowski and WLOP" (Fig. 1)
4. "Chibi man figurine, modern Disney style" (Fig. 8)
5. "Man with a majestic beard, closeup, D&D, fantasy, intricate, elegant, highly detailed, digital painting, artstation, concept art, matte, sharp focus, illustration, art by Artgerm, Greg Rutkowski, Alphonse Mucha" (Fig. 1)

4.2 DreamBooth Experiments

The prompts for DreamBooth include:

1. "nesrelah_001 on a horse" (Fig. 9).
2. "Highly detailed marble and jade sculpture of nesrelah_001, volumetric fog, Hyperrealism, breathtaking, ultra realistic, unreal engine, ultra detailed, cyber background, Hyperrealism, cinematic lighting, highly detailed, breathtaking ,photography, stunning environment, wide-angle [cgi, 3d, doll, octane,

Fig. 7: Prompt: Man riding a horse, facing camera.

Fig. 8: Prompt: Chibi man figurine, modern Disney style.

render, bad anatomy, blurry, fuzzy, extra arms, extra fingers, poorly drawn hands, disfigured, tiling, deformed, mutated]" (Fig. 1)
3. "Portrait of nesrelah_001, dramatic lighting, illustration by greg rutkowski, yoji shinkawa, 4k, digital art, concept art, trending on artstation" (Fig. 1)
4. "An epic fantastic realism comic book style portrait painting of nesrelah_001 robot with kanji tattoos and decals, apex legends, octane render, intricate detail, 4 k hd, unreal engine 5, ex machina, irobot, gerald brom" (Fig. 1)
5. "Chibi nesrelah_001 figurine, Disney style" (Fig. 10)

Fig. 9: Prompt: nesrelah_001 on a horse.

Fig. 10: Prompt: Chibi nesrelah_001 figurine, modern Disney style

4.3 Overall Assessment with Comparison

Considering these experiments, Table 1 synopsizes our evaluation, the base case being that of Stable Diffusion without fine-tuning; the others being its fine-tuning with Hypernetwork and DreamBooth respectively. The evaluation is fairly simplistic at this point because this is a pilot study. Users evaluating this work are asked to mention whether they are satisfied with the created images (comments optional). We have a small group of 20 student users. It is evident that Hypernetworks and DreamBooth both yield higher levels of user satisfaction than Stable Diffusion alone. Yet the training data used for both these models is substantially low, compared to the original Stable Diffusion dataset, indicating that we can achieve good results in text-to-image generation in various scenarios, "thinking outside the box", without much additional data.

Table 1: Summary of Assessment Outcomes in Pilot Study

Approach	**Avg. Accuracy**	**Synopsis of Users' General Comments**
Base Case	Approx. 80%	Users not satisfied with $\sim 20\%$ of images
Hypernetwork	Approx. 90%	Users partly satisfied with $\sim 10\%$ of images
DreamBooth	Approx. 95%	Users find $\sim 5\%$ of images slightly below expectation

Table 2: Comparative Observations of Fine-Tuning Models in Pilot Study

Model	**V-RAM**	**Main Source**	**Train-Time**	**Other Observations**
Hypernetwork	$8GB$	Personal Comp.	10000 steps	Class affected by subject
DreamBooth	$24GB$	Google Colab	4040 steps	Class distinct from subject

As per quantitative and qualitative observations, the time needed to train these models is of the order of minutes (not hours) while the resources used are

mainly a personal computer and Google Colab. More specifically, the Hypernetwork needs 8GB of V-RAM and is trained on a PC, the training requiring 10,000 steps; while DreamBooth needs 24 GB of V-RAM and is trained mainly using Google Colab, the training occurring in 4,040 steps. This is summarized in Table 2. Thus, the Hypernetwork is less resource-consuming and is trained using more steps while DreamBooth is more resource-intensive and is trained with relatively fewer steps. Also, the Hypernetwork makes the class get directly affected by the subject associated with it, while DreamBooth keeps the subject and class distinct from each other (as explained at the beginning of this section). Both the models are effective for fine-tuning classification in Stable Diffusion, requiring reasonable training time and resources, as evident from the Tables.

4.4 Targeted Applications

The potential applications of these technologies are quite substantial. One can certainly leverage such advancements to be able to create art, despite one's skills. An interesting application that could be considered is the generation of new pictures of a dearly departed family member, friend, or famous personality, e.g. movie actor. Another application can be to generate different styles of clothing to figure out what looks best on a given person. One can even save substantial amounts of money on a photo-shoot by tweaking several parameters. These applications entail subject injection in photographs to encompass reality.

On the whole, the real-life image curation can have many benefits due to infusing more personalization, potentially useful in contexts such as:

- Photo adjustments good for clothes comparison, formal photo-shoots etc.
- Autonomous vehicles for automated driving across various regions
- Human-robot collaboration / interaction with enhanced image classification
- Intelligent tutoring systems and mobile apps with personalized icons for user interaction to achieve worldwide outreach in a variety of scenarios
- Movie animations with high quality images & videos entailing mass appeal
- Next-generation multipurpose robots with improved object detection and versatile behavior

Some of these applications can make positive impacts on various facets of AI in smart cities in line with earlier work in our research lab, e.g. improved autonomous vehicles for smart mobility [25], enhanced human-robot collaboration for smart manufacturing [11] etc. Much of the work in this paper can be used for future studies in our lab across various projects spanning visualization, textual data, mobile app development, smart city applications and so forth [21] [20], [6], [38], [26]. Hence, our work in this paper is applied research with implementation and experimentation, contributing to neural networks and image processing, making positive impacts on numerous targeted applications.

5 Related Work

There is a much work on Stable Diffusion and fine-tuning, as elaborated in the literature [10], [29], [12]. There is also a myriad of research on image mining in

general [35], [18], [37], [34], [22]. Our work in this paper is orthogonal to such research. Yet, one must be cognizant of the technological and ethical limitations. The creators of the Stable Diffusion paradigm have stipulated the following. "While the capabilities of image generation models are impressive, they can also reinforce or exacerbate social biases. Stable Diffusion v1 was trained on subsets of LAION-2B(en), which consists of images that are primarily limited to English descriptions. Texts and images from communities and cultures that use other languages are likely to be insufficiently documented. This affects the overall output of the model, as Caucasian and Western cultures are often set as the default. Further, the ability of the model to generate content with non-English prompts is significantly worse than with English-language prompts [29]."

Other researchers have shown that there are true societal biases. Bianchi et al. [7] compiled images on specific prompts to see if they catered to a specific race, ethnicity, sexual orientation, etc. For instance, in "An American man and his car" vs. "An African man and his car", the picture of the American man is often portrayed as affluent, while the African man is usually shown as impoverished. Perhaps future implementations of Stable Diffusion can address this problem, or perhaps fine-tuning technologies such as Hypernetworks and DreamBooth can help remove societal biases inherent in current models [7]. In connection with this, issues such as public opinion can often be important; just as people voice their concerns on environmental matters [13], they can also express their reactions to artistic pursuits, thereby offering the scope for multidisciplinary research spanning data science, analogous to a few other works [36], [32].

Furthermore, some limitations of the overall technology include using an artist's name to generate an image in that artist's style. We used artist-names such as Greg Rutkowski and Alphonse Mucha to generate our chosen art in styles reminiscent to those they produce. By using an artist's style, the artistic quality of the image increases. However, if someone were to profit from the image, it then begs the question: Should the artist be credited, and even compensated for the given image? Also, would future work by the artist be devalued if AI is able to generate the same art-style for free? These and other limitations indicate that Stable Diffusion has to be taken with a grain of salt. Much of this presents the potential for future work.

Other related work entails infusing commonsense knowledge (CSK) into image generation [15], object detection [9] and more aspects [2], [19]. There is a plethora of work on CSK as outlined in tutorials [33], [27] some of which can be relevant for image creation. Cultural commonsense knowledge is being studied [1], [23] and can be explored further to enhance text-to-image and image-to-text (automated captions) generation. This is another aspect of D&I. Our paper does not address CSK, however, it addresses societal context and global diversity to some extent. Future work in this area can consider facets of CSK as well.

On a final note, we can mention that recent advances such as ChatGPT [24] might possibly be correlated with some of our work in this paper. Much of ChatGPT thrives on reinforcement learning. Hence, the text and images used in this study can be used to train ChatGPT and other such systems in order

to make them more versatile and globally-oriented, propelling additional work on the lines of thinking outside the box. We make just a modest contribution here, helping to augment text-to-image creation with fine-tuning classification, thereby being advantageous to many AI applications.

6 Conclusions

In this paper, we aim to generate novel photo-realistic images from text prompts, using a given reference set, laying much emphasis on subject personalization. This is achieved through Stable Diffusion, DreamBooth, and Hypernetworks. Comparing the results, there definitely is a better degree of control with DreamBooth vs. Hypernetworks. Also, DreamBooth seems to generate higher quality images. However, the Hypernetwork model is less resource-heavy and can be implemented completely on a typical PC [30, 16].

Main Contributions: Briefly, our contributions are highlighted as follows.

1. Investigating Stable Diffusion with Hypernetworks and DreamBooth for text-to-image generation
2. Getting high user satisfaction for image creation with low training data
3. Addressing novel and versatile contexts, producing good quality images, and outlining various targeted applications
4. Leveraging diversity & inclusion as per various real-life perspectives
5. Making broader impacts on AI in smart cities in a modest manner

Limitations and Future Work: Various limitations of Stable Diffusion seen in the literature and corroborated by our experiments, present the scope for future work in the area. The notion of diversity & inclusion for image creation from text needs more attention. Though we have focused on some of it, there are still open avenues, e.g. as noticed in the related work. Racial, ethnic and other types of diversity in text-to-image generation calls for further research. Exploring these aspects while also addressing privacy-preserving issues and confidentiality concerns, can pose more challenges. These can open up avenues for future work.

Furthermore, qualitative performance of the models used in this work can be judged on a wide variety of tuning datasets. In this paper, two diverse datasets have been used to tune each of the two models, while in the future more heterogeneous data can be considered. Likewise, quantitative performance in terms of "user preference" on images obtained by the two models trained with the same training data can be judged as well, e.g. users should choose the "best image" between any pair of images generated by the two models. Such detailed experiments can be carried out with larger datasets as well as bigger user study groups such as those on AMT (Amazon Mechanical Turk). Permissions for such work need to be obtained from the respective IRB (Institutional Review Board) since there are human subjects involved in the study, and this can be rather time-consuming. Hence, a small scale study has been conducted in this paper with informal evaluations, which puts forth the scope for further work.

Additionally as future research, we can explore details from an application-standpoint, e.g. human-robot collaboration, autonomous vehicles, and intelligent tutoring systems, blending that with relevant projects in our labs. It is also important to address ethical issues such as artist compensation which might be beyond the scope of our own work but can be addressed by other researchers working across the concerned applications. Finally, we aim to investigate more specific roles that commonsense knowledge can play with respect to the overall theme of this research. Our work in this paper modestly contributes to ANN and image processing, making potential impacts on various AI applications.

Acknowledgments & Disclaimer

Aparna Varde acknowledges NSF grants 2018575 "MRI: Acquisition of a High-Performance GPU Cluster for Research & Education", and 2117308 "MRI: Acquisition of a Multimodal Collaborative Robot System (MCROS) to Support Cross-Disciplinary Human-Centered Research & Education at Montclair State University". She is a visiting researcher at Max Planck Institute for Informatics, Germany (ongoing from sabbatical). She is an Associate Director of CESAC: Clean Energy & Sustainability Analytics Center, Montclair State University. We make a disclaimer that the opinions presented here are extracted from online sources; and the content of this paper including the images is not meant to offend / hurt any national, ethnic, cultural, racial, religious and other groups. The images produced here are taken with the consent of the respective subjects. Any resemblance to anyone else is coincidental. This is a pilot study.

References

1. Acharya, A., Talamadupula, K., Finlayson, M.A.: An atlas of cultural commonsense for machine reasoning. In: AAAI Conference on Artificial Intelligence (2021)
2. Aditya, S., Yang, Y., Baral, C., Fermuller, C., Aloimonos, Y.: From images to sentences through scene description graphs using commonsense reasoning and knowledge. arXiv preprint arXiv:1511.03292 (2015)
3. A.I., S.: Stable diffusion public release, https://stability.ai/blog/stable-diffusion-public-release
4. Alaluf, Y., Tov, O., Mokady, R., Gal, R., Bermano, A.H.: Hyperstyle: Stylegan inversion with hypernetworks for real image editing. arXiv:2111.15666 [cs] (03 2022), https://arxiv.org/abs/2111.15666
5. AUTOMATIC1111: Stable Diffusion Web UI (11 2022), https://github.com/AUTOMATIC1111/stable-diffusion-webui
6. Basavaraju, P., Varde, A.S.: Supervised learning techniques in mobile device apps for androids. ACM SIGKDD Explorations **18:2**, 18–29 (2017)
7. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., Caliskan, A.: Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. arxiv.org (11 2022). https://doi.org/10.48550/arXiv.2211.03759, https://arxiv.org/abs/2211.03759
8. Birme: Birme - bulk image resizing made easy 2.0 (online free), https://www.birme.net/

9. Chernyavsky, I., Varde, A.S., Razniewski, S.: CSK-Detector: Commonsense in object detection. In: IEEE International Conference on Big Data. pp. 6609–6612 (2022), https://doi.org/10.1109/BigData55660.2022.10020915
10. Cheung, B.: Stable diffusion training for personal embedding (11 2022), https://bennycheung.github.io/stable-diffusion-training-for-embeddings
11. Conti, C.J., Varde, A.S., Wang, W.: Human-robot collaboration with commonsense reasoning in smart manufacturing contexts. IEEE Transactions on Automation Science and Engineering **19**(3), 1784–1797 (2022)
12. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. arXiv:2105.05233 [cs, stat] (06 2021), https://arxiv.org/abs/2105.05233
13. Du, X., Kowalski, M., Varde, A.S., de Melo, G., Taylor, R.W.: Public opinion matters: Mining social media text for environmental management. ACM SIGWEB **Autumn**, 1–15 (2020)
14. Face, H.: Dreambooth fine-tuning example, https://huggingface.co/docs/diffusers/training/dreambooth
15. Garg, A., Tandon, N., Varde, A.S.: I am guessing you can't recognize this: Generating adversarial images for object detection using spatial commonsense. In: AAAI Conference on Artificial Intelligence. vol. 34, pp. 13789–13790 (2020)
16. Ha, D., Dai, A., Le, Q.V.: Hypernetworks. arXiv:1609.09106 [cs] (12 2016), https://arxiv.org/abs/1609.09106
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. arXiv:2006.11239 [cs, stat] (12 2020), https://arxiv.org/abs/2006.11239
18. Hsu, W., Lee, M.L., Zhang, J.: Image mining: Trends and developments. Journal of intelligent information systems **19**(1), 7–23 (2002)
19. Ilievski, F., Szekely, P., Cheng, J., Zhang, F., Qasemi, E.: Consolidating commonsense knowledge. arXiv preprint arXiv:2006.06114 (2020)
20. Kaluarachchi, A., Roychoudhury, D., Varde, A.S., Weikum, G.: SITAC: Discovering semantically identical temporally altering concepts in text archives. In: Intl. Conf. on Extending Database Technology (EDBT), ACM. pp. 566–569 (2011)
21. Karthikeyan, D., Shah, S., Varde, A.S., Alo, C.: Interactive visualization and app development for precipitation data in sub-saharan africa. In: IEEE Intl. IOT, Electronics and Mechatronics Conf. (IEMTRONICS). pp. 1–7. IEEE (2020)
22. Karthikeyan, D., Varde, A.S., Wang, W.: Transfer learning for decision support in covid-19 detection from a few images in big data. In: IEEE International Conference on Big Data. pp. 4873–4881 (2020)
23. Nguyen, T.P., Razniewski, S., Varde, A., Weikum, G.: Extracting cultural commonsense knowledge at scale. In: WWW, the ACM Web Conference (2023). https://doi.org/10.48550/ARXIV.2210.07763, https://arxiv.org/abs/2210.07763
24. OpenAI: Introducing ChatGPT. https://openai.com/blog/chatgpt (2022)
25. Persaud, P., Varde, A.S., Robila, S.: Enhancing autonomous vehicles with commonsense: Smart mobility in smart cities. In: 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI). pp. 1008–1012. IEEE (2017)
26. Puri, M., Varde, A., Du, X., De Melo, G.: Smart governance through opinion mining of public reactions on ordinances. In: 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI). pp. 838–845. IEEE (2018)
27. Razniewski, S., Tandon, N., Varde, A.S.: Information to wisdom: Commonsense knowledge extraction and compilation. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. pp. 1143–1146 (2021)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. arXiv:2112.10752 [cs] (04 2022), https://arxiv.org/abs/2112.10752

29. Rombach, R., Esser, P.: Compvis/stable-diffusion-v-1-4-original · hugging face (11 2022), https://huggingface.co/CompVis/stable-diffusion-v-1-4-original
30. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv:2208.12242 [cs] (08 2022), https://arxiv.org/abs/2208.12242
31. ShivamShirao: diffusers/examples/dreambooth at main · shivamshrirao/diffusers (09 2022), https://github.com/ShivamShrirao/diffusers/tree/main/examples/dreambooth
32. Suchanek, F.M., Varde, A.S., Nayak, R., Senellart, P.: The hidden Web, XML and the semantic Web: Scientific data management perspectives. In: International Conference on Extending Database Technology (EDBT), ACM. pp. 534–537 (2011)
33. Tandon, N., Varde, A.S., de Melo, G.: Commonsense knowledge in machine intelligence. ACM SIGMOD Record **46**(4), 49–52 (2017)
34. Theisen, W., Cedre, D.G., Carmichael, Z., Moreira, D., Weninger, T., Scheirer, W.: Motif mining: Finding and summarizing remixed image content. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1319–1328 (2023)
35. Varde, A., Rundensteiner, E., Javidi, G., Sheybani, E., Liang, J.: Learning the relative importance of features in image data. In: IEEE ICDE (International Conference on Data Engineering), workshops. pp. 237–244 (2007)
36. Varde, A.S.: Challenging research issues in data mining, databases and information retrieval. ACM SIGKDD Explorations **11:1**, 49–52 (2009)
37. Varde, A.S., Rundensteiner, E.A., Ruiz, C., Maniruzzaman, M., Sisson Jr, R.D.: Learning semantics-preserving distance metrics for clustering graphical data. In: Proceedings of the 6th international workshop on Multimedia data mining: mining integrated media and complex data. pp. 107–112 (2005)
38. Varghese, C., Pathak, D., Varde, A.S.: SeVa: A food donation app for smart living. In: 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC). pp. 0408–0413. IEEE (2021)