Variational Autoencoder for Image Generation

Algot Larsson Eskilsson
alges694@student.liu.se
Anton Bergman
antbe028@student.liu.se
Cajsa Schöld
cajsc235@student.liu.se
Pontus Ferm
ponfe408@student.liu.se
Shamil Limbasiya
shali220@student.liu.se

Abstract

This paper explores the development and advancement of Variational Autoencoders (VAEs) for image generation tasks. VAEs have emerged as a powerful tool in generative modeling, capable of learning latent space representations of data and generating new samples. In this paper we develop our own VAE utilizing the CIFAR-10 dataset, explore architecture and performance metrics of VAEs, comparing them with diffusion models. Evaluation metrics include visual inspection of generated images and the structural similarity index, where results indicate that the VAE can preserve structural information but struggles with sharpness and color range compared to original images. We discuss potential improvements such as larger datasets, hyperparameter tuning, and validation sets to mitigate overfitting. Our findings suggest that training on single classes yields better results, and there is room for optimization to enhance VAE performance.

1 Introduction

In recent years, the advancement of deep learning techniques has led to remarkable progress in various fields, including computer vision. Among these techniques, Variational Autoencoders (VAEs) have emerged as a powerful tool for generative modeling. VAEs are capable of learning rich latent representations of data and generating new samples from these representations. In this paper, we present our project which focused on developing a Variational Autoencoder tailored for image generation tasks. Our primary objective was to construct a model capable of synthesizing images from random noise, leveraging the expressive power of deep learning architectures.

To accomplish this goal, we utilized the CIFAR-10 dataset, a widely-used dataset comprised of labeled color images across ten distinct classes (1). The diversity of the dataset allow for a robust training of generative models like VAEs and facilitate thorough evaluation of the model performance.

Through this project, we aim to contribute to the understanding of VAEs for image generation and explore their potential applications. Our approach involves a comprehensive investigation of model architecture and evaluation metrics to assess the efficiency of the models capability in generating images. To further assess the performance our model, we have conducted a comparative analysis with a diffusion model introduced in lab2 of the TDDE70 Deep Learning course at Linköping University.

In the following sections, we provide a comprehensive overview of related work, detail our problem formulation, describe our methodology, present experimental results, and conclude with a discussion on the implications of our findings.

Submitted to 1st Seminar on Deep Learning Projects (TDDE70).

2 Related Work

2.1 Variational autoencoders

Variational autoencoders (VAEs) were first introduced in 2013 by Kingma and Welling in their paper titled “Auto-Encoding Variational Bayes”(2). The objective of the VAEs, like regular autoencoders, is to learn a low-dimensional representation of the input data. However, VAEs introduce probability to the model by encoding the data into a probability distribution over latent variables. This characteristic makes them suitable for e.g. generative tasks. Kingma and Welling introduced the Stochastic Gradient Variational Bayes (SGVB) approach to obtain a simple and unbiased estimator of the evidence lower bound, which is indirectly used to approximate the true posterior distribution. This SGVB estimator is utilized in the AutoEncoding Variational Bayes Algorithm (AEVB) to efficiently optimize the parameters of the VAE. Compared to methods like MCMC sampling, this approach introduced by Kingma and Welling involves simpler and fewer calculations and allows for cheaper training of the model.

The paper “Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images”(3) illustrates the application of deep variational autoencoders for image generation tasks. They trained very deep (defined as networks with more layers than previous models) VAEs on several image datasets, such as CIFAR-10, with an architecture that enables the use of fewer parameters than previous models, while achieving better results. Negative log-likelihood is used as a metric to compare their network’s performance against existing models, including autoregressive models (e.g. PixelCNN++), a flow-based model (Flow++) and previous VAEs with fewer stochastic layers. The experiments show that deeper VAES can outperform other generative models in terms of likelihood while using fewer parameters, thus improving efficiency.

2.2 Diffusion models

Diffusion models are latent variable models in which the forward process gradually adds Gaussian noise to the data following a Markov chain of transitions, the models can then learn to reverse this process and gradually remove the Gaussian noise to generate new output (4). During training the model learns to predict the Gaussian noise added to the input at the current step. When generating new output, it starts with a Gaussian distribution of noise where the model iteratively predicts the previous step. While diffusion models have some resemblance to VAEs, the process of adding noise and learning how to reverse the process is intuitively easier to follow.

Ho et.al illustrates the application of diffusion models for image generation in their paper “Denoising Diffusion Probabilistic Models”(4), where they present a model capable of generating high quality images. Their model utilizes a U-Net structure for the backward process and is trained using a weighted variational bound, which combines diffusion probabilistic models and denoising score matching. When training the model on the CIFAR-10 dataset and using inception and FID scores to evaluate the model performance, they find that their model outperform many previous models found in the literature, as eg. Gated PixelCNN, Sparse Transformer and SNGAN.

3 Problem formulation

The aim of the project was to build and evaluate a VAE for image generation tasks. A VAE is a neural network that renders images based on a latent representation. It consists of two main components: encoder and decoder. The encoder maps the input data to a probabilistic latent space. This means that instead of mapping the input data to one point in the latent space it instead maps it to a distribution in the latent space. The decoder then uses that distribution in the latent space to recreate the original data.

During the course of the project, we aimed to achieve the following:

1. Build a variational autoencoder that could synthesize images from noise.
2. Train the autoencoder using the CIFAR-10 dataset, which is described more in depth below.
3. Evaluate the performance of the variational autoencoder compared to a diffusion model.
4. Identify the positive and negative aspects of a variational autoencoder by seeing if it could outperform the diffusion model.

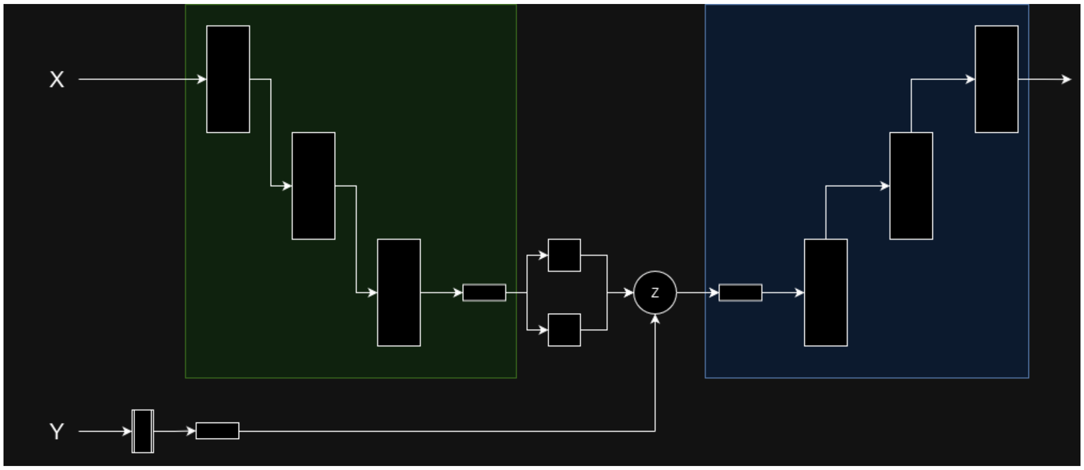

Figure 1: The variational autoencoder consisting of encoder (green) with 3 convolutional blocks with downsampling, represented by downward arrows, and decoder (blue) 3 convolutional blocks with upscaling, represented by upward arrows.

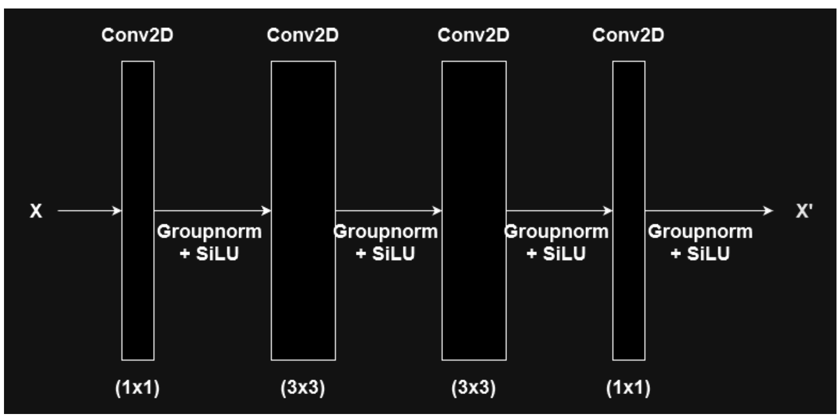

Figure 2: Visualization of a single convolutional block in the variational autoencoder.

3.1 Data

The data that was used to train was the CIFAR-10 dataset (5), which consists of 60 000 32x32 labeled color images. The images are split up into 10 different classes. Each class containing 5 000 training images and 1 000 test images. The different classes are airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck. An important note is that the images only contain one prominent instance of the classes object. Another thing to note is that there are no overlaps across different classes. For example, the class automobile doesn't contain any images of pickup trucks as there could be some confusion with regards to the class truck.

3.2 Expected outcome

We anticipate that the VAE will be able to synthesize images that resembles the original input image during training. But due to the short timespan of this project and our limited amount of resources, we don't expect it to create close to exact copies of the input image. We do however believe that we will have somewhat similar results when comparing the VAE to a diffusion model.

4 Method Description

4.1 Model

The final model, see figure 1, was built up of blocks with four CNN:s as described by R. Child (6) which can be seen in figure 2. Contrary to the paper the model did not make use of residual connections. After the input is encoded it is sent through two multilayer perceptrons that represent the mean and variance. These values are then used to derive the Kullback-Leiber divergence for the loss function, as well as creating a representation in the latent space together with introduced noise and the class representation. The input was then passed through the decoder part which mirrors the encoder to create an output with the same dimensionality as the input image. When trained the encoder part was put aside and only the decoder was used to generate new images with the class representation and uniform noise as input.

The initial VAE model was designed to only learn one class of images to evaluate the dimensions of the model. Once the model successfully created output similar to the training data the model was expanded to encompass all classes of the dataset. This was done by feeding the class label through an embedding layer and a multilayer perceptron. The output was then weighted and added to the latent representation.

The diffusion model used was the U-Net and cddpm model provided and implemented in Lab 2 in this course, TDDE70. The model was altered to handle the CIFAR-10 dataset.

4.2 Training

The training was done on a subset of the CIFAR-10 dataset, the subset being a single class or a smaller number of images from each class. This was mainly done due to hardware limitations to speed up the training. The subset consisted of 5000 images of a single class and 100 images from each class when multiple classes were used. The training of the model is simple and similar to the training done in the labs. The data was divided into batches of size 46. The other hyperparameters are the number epochs which were 200 and the learning rate of 0.001. To evaluate the predictions during training of the VAE two loss functions were used: binary cross entropy and Kullback-Leibler divergence, these two were then added together for a total loss. For the diffusion model only the mean squared error loss was used. The optimizer used on the VAE was the Adam optimizer and the diffusion used the RAdam optimizer.

5 Result

5.1 Testing

The VAE model was trained on the classes "plane" and "frog" separately but also multiple classes. Besides looking at the generated images we also looked at the structural similarity index which compares each pixel between two images and returns a value between -1 and 1. One indicates that the pictures are similar in structure and -1 complete dissimilarity. The test images consisting of 1000 examples for each class were than used to calculate the structural similarity index. The results are presented below.

5.1.1 Planes

In Figure 3 the original test images are displayed where Figure 4 shows the images generated by the model. When calculating the structural similarity index we got 0.99872 which indicates that the generated image closely preserves the structural information.

5.1.2 Frogs

When calculating the structural similarity index we got 0.99969 which indicates that the generated image closely preserves the structural information. The visual result can be seen in the appendix.

Figure 3: Testing images

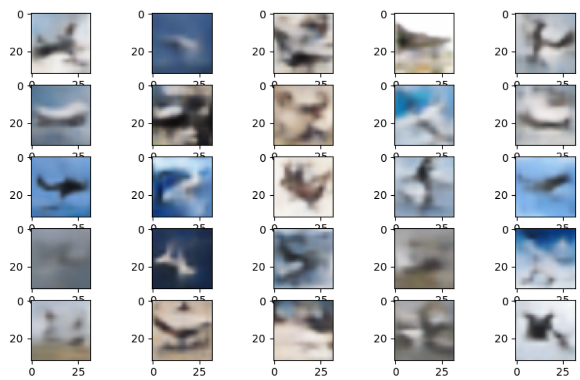

Figure 4: Generated images by the model

5.1.3 Multiple classes

When training the model on multiple classes the model was initially trained on one image from each class to show that each class could successfully be sampled from the latent space. When trained on 100 images from each class, the generated results became less coherent than for the single class model. The classes become less distinguishable but still have some unique properties, e.g. the water in class 8 (ships). The visual result can be seen in the appendix.

5.1.4 Diffusion Model

For the diffusion model we received varying results across different training cycles, as can be seen in the appendix. Through visual inspection it can be observed that the model sometimes generated outputs resembling something similar to the images in the CIFAR-10 dataset, while at other times it generated images consisting of only random noise. The images that resembled something from CIFAR-10 have more complex structures than the objects in dataset.

6 Conclusion/Discussion

6.1 Results VAE

When visually comparing our generated images with the original images we see a couple of key differences. First of all we see that the sharpness of the picture differs a lot. In our original dataset we clearly see our objects in almost every picture whereas with the generated ones there are several cases where it's hard to see the motive. An other thing that stands out is that the color range is a lot smaller with the generated images, in many of the images there is a light background and with a darker object seen in the picture, this became more apparent when looking at frogs rather than planes. With that said it has done a great job considering it's small training dataset that only consists of 5000 images. In the case of planes we in many instances see the resemblances to objects that look similar to planes. However, when training it on multiple classes it is almost impossible to make out the objects in the generated images. As mentioned in the result some key characteristics could still be made out like the water when the model was asked to generate boats. So in all the cases the model seems to keep track of the structures in the images.

One improvement that could be done is a larger data set for training either by augmentation of the existing data or with more examples. Another improvement is larger picture, the CIFAR-10 data set consists only of pictures with size 32x32 by increasing the size we also increase the amount of data the model can learn and sample from. The hyperparameters are another point of improvement, turning the training of the model we didn't experiment with the hyperparmeters to find the optimal ones so by experimenting a bit perhaps a better model can be found. One last major improvement that could be done is including a validation set to see if the model is showing signs of overfitting. Currently the model is trained for 200 epochs but by using validation we can use early stopping to avoid overfitting.

The conclusions we draw from this are that the model is better when trained on one class rather than several at once. In both the cases of training with one and many classes the model seems preserve much of the structural information that is then used when generating new images. We also draw the conclusion that there are plenty of improvements that could be done that might yield a better model in the end.

6.2 Diffusion models vs VAE

In comparison to diffusion models, VAEs offer a distinct approach to generative modeling. While both methods aim to generate realistic images, they differ significantly in their underlying principles and training methodologies.

Diffusion models, as exemplified in the work of Ho et al. (4), follow a process where Gaussian noise is progressively added to the data, and the model learns to reverse this process to generate new samples. This approach is intuitive and conceptually straightforward, as the model directly learns the addition and removal of noise.

On the other hand, VAEs, as introduced by Kingma and Welling (2), employ a probabilistic framework to learn latent representations of data. VAEs map input data to a distribution in latent space, enabling them to capture complex data distributions and generate novel samples. They utilize techniques like variational inference and stochastic gradient descent to optimize the model parameters efficiently.

In terms of performance, diffusion models have shown remarkable capabilities in generating high-quality images, as demonstrated by Ho et al. (4). They excel in preserving fine details and producing visually appealing results. However, diffusion models might require more complex training procedures and architectures to achieve optimal performance. On the contrary, VAEs offer a more versatile and scalable approach to image generation. They are relatively simple to implement and train, making them more accessible for various applications. While VAEs may not always match the exact image fidelity achieved by diffusion models, they provide a solid foundation for exploring generative modeling across different domains.

In our results, which are detailed in the report above, we observed a high structural similarity index for VAE-generated images. However, upon visual inspection, we noticed a lack of fine details and sharpness in these images. On the other hand, the diffusion model-generated images exhibited slightly better detail and sharpness, but with potentially compromised structural accuracy. These observations align with the strengths and limitations described for both diffusion models and VAEs. While VAEs offer structural accuracy and ease of implementation, diffusion models excel in detail preservation at the cost of increased complexity. Overall, our results provide insights into the trade-offs between structure, sharpness and complexity in generative modeling techniques.

7 Ethical Considerations

As we delve into the development and applications of variational autoencoders (VAEs) for image generation, it is equally as important to address the ethical implications associated with this technology. While VAEs offer promising opportunities for innovation and advancement, they also pose several ethical challenges.

7.1 Privacy and Data usage

One of the most ethical concerns is with regards to privacy and the responsible usage of data. The datasets used to train VAEs often consist of vast amounts of images collected from various sources. Ensuring the privacy and consent of individuals whose images are included in these datasets is paramount. Recent discussions, as highlighted in a Scientific American article (7), emphasize how personal information is widely used in training generative AI models.

Without proper safeguards, there's a risk of unauthorized access to personal information, potentially leading to privacy breaches and violations of individual rights. As responsible individuals, we must follow data protection policies, obtain informed consent where necessary, and prioritize the anonymization of sensitive information to safeguard the privacy of individuals.

7.2 Biases and Fairness

Another critical consideration is the potential for bias and discrimination in VAE-generated outputs. Biases present in the training data, whether implicit or explicit, can manifest in the generated images which may result in unfair, harmful or discriminatory outcomes. It is essential to address biases at every stage of the model development process, from data collection and preprocessing to algorithm design and evaluation. By employing techniques such as bias detection, mitigation, and fairness-aware learning, we can strive to minimize the impact of biases and promote fairness and inclusivity in our models.

For example, a recent incident concerning Google (8) highlights AI models' bias and fairness challenges, where Google apologized after its AI model, Gemini, generated racially diverse images of Nazis.

7.3 Misuse of Generated Content

The ability of VAEs to generate highly realistic images raises concerns about the potential misuse of this technology. In the wrong hands, VAE-generated content could be used to create deceptive or harmful material, such as deepfakes, which can have serious consequences for individuals and society at large. To mitigate this risk, it's essential to promote responsible usage of generative models and advocate for ethical guidelines governing their deployment. Transparency regarding the origin of generated content and mechanisms for content verification can help mitigate the spread of misinformation and prevent malicious use of VAE-generated images.

In conclusion, while VAEs hold immense potential for innovation and creativity, they also pose significant ethical challenges that must be addressed proactively. By prioritizing privacy, fairness, responsible usage, transparency, and accountability, we can harness the power of VAEs for positive societal impact while mitigating potential harms. Together as a society, we have a collective responsibility to navigate these ethical considerations thoughtfully and ensure safe future AI development.

8 Statements of Contribution

During the project, a majority of the programming was done together during meetings. Therefore all the project members played a part in creating the initial model and it might be hard to pinpoint who did what.

8.1 Algot

My contribution in the implementation of the project was working on the model architecture during the programming sessions done together. Outside these sessions I did troubleshooting and training for the initial VAE model, some helper functions, as well as extended the VAE model to work with several classes of the CIFAR-10 dataset. For the report I wrote the Method Description and the results for the multiple classes VAE, as well as a bit on the Diffusion models in Related Work,. Overall, I have obtained a good understanding of both generative models used in the project.

8.2 Anton

My contribution to the project and this paper encompassed writing the Conclusion, Abstract and Introduction, as well as the discussion on Ethical Considerations. Additionally I actively participated in all parts of both the development of the VAE model and writing of this project paper. In the development of the VAE model most progress was made during group sessions and meetings where I actively contributed, building a comprehensive understanding of the model development ant functionality. Overall, my involvement demonstrates a well-rounded grasp of the project objectives and methodologies.

8.3 Cajsa

My contribution to this paper mainly included writing the related work section for the VAE and diffusion model. Regarding the programming I preformed the initial implementation of the diffusion model for the CIFAR-10 dataset. In addition to that I actively participated at our group sessions and meetings where we, as mentioned above, built the initial model. Writing the project paper and participating in these sessions have deepened my understanding of these models and how they can be used for image generation.

8.4 Pontus

My contribution to the project and the paper has been writing the Project Formulation as well as contributing a bit to the information about diffusion models. I also ran some of the training for the diffusion model after it had been adjusted for the CIFAR-10 dataset. In regards to the development, I actively took part in our group coding sessions and meetings, where I amongst other things created the first version of the decoder for the model. During these meetings, the entire group collaborated on creating the model architecture. These sessions, along with the project paper has given me a better understanding of the model and the concept of Variational Autoencoders.

8.5 Shamil

My contribution to the project has been writing the results of the VAE as well as the discussion about VAE. I also ran some training for the model for the class frog. I was involved during the whole development process and participated actively during our group sessions where most of the progress was made. Additionally I implemented the evaluation method in terms of structural similarity index to be able to compare our generations to the test data. Overall through the project and this paper I have gained a better understanding of different types of generative models and learned their respective strengths.

References

[1] A. Krizhevsky, V. Nair, and G. Hinton, "The cifar-10 dataset." `https://www.cs.toronto.edu/~kriz/cifar.html`.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013, revised 2022.

[3] R. Child, "Very deep vaes generalize autoregressive models and can outperform them on images," 2021.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[5] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[6] R. Child, "Very deep vaes generalize autoregressive models and can outperform them on images," *arXiv preprint arXiv:2011.10650*, 2020.

[7] L. Leffer, “Your personal information is probably being used to train generative ai models.” https://www.scientificamerican.com/, 2023.

[8] A. Robertson, “Google apologizes for ‘missing the mark’ after gemini generated racially diverse nazis.” `https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical`, 2024.

A Additional test results

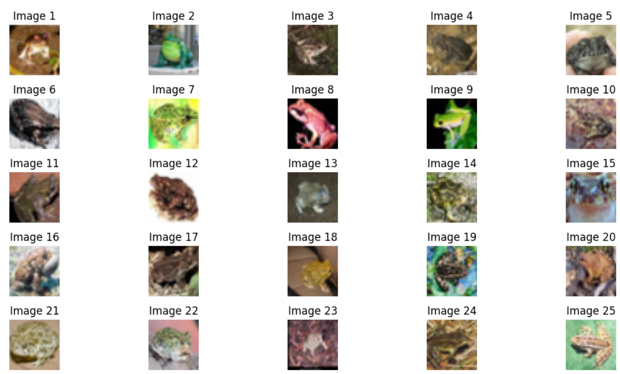

Figure 5: Testing images

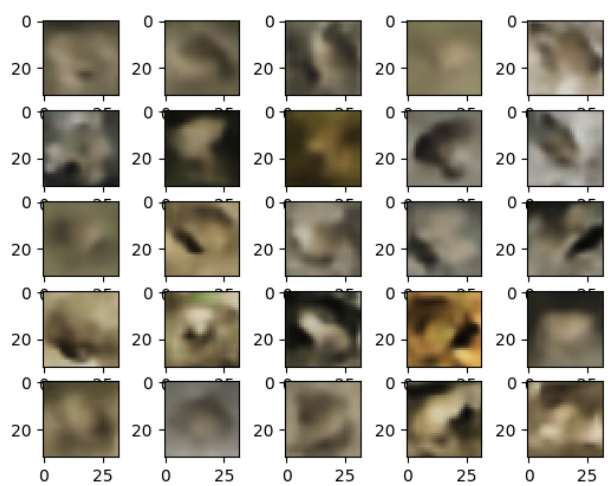

Figure 6: Generated images by the model

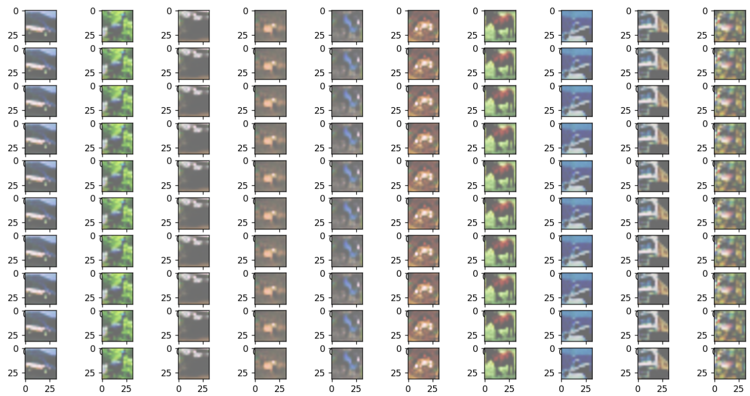

Figure 7: 10 generated samples in each class for model trained on 1 image from each class, class 1-10 from left to right.

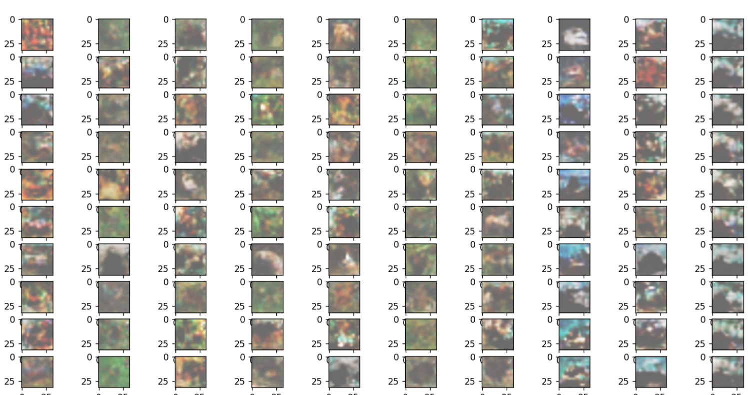

Figure 8: 10 generated samples in each class for model trained on 100 images from each class, class 1-10 from left to right.

Figure 9: The results from training the diffusion model on the CIFAR-10 dataset for 200 epochs

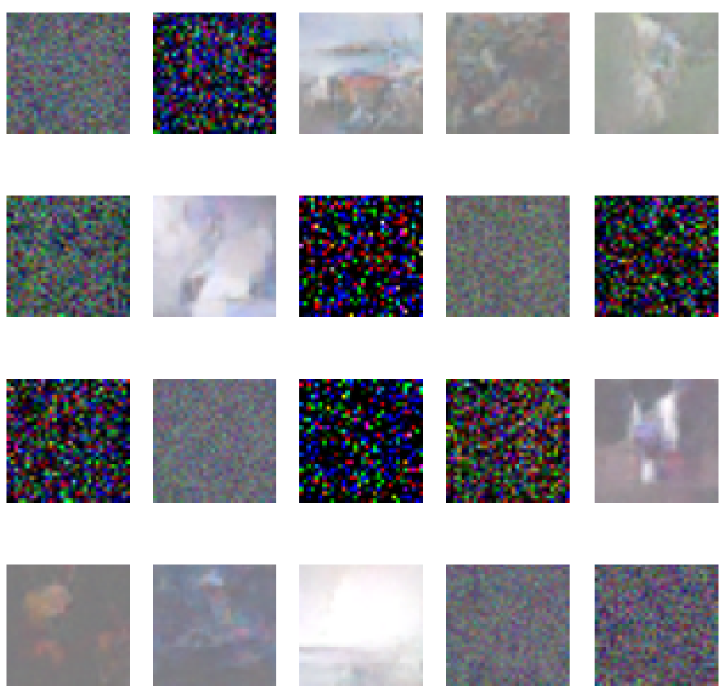

Figure 10: The results from a second attempt of training the diffusion model on the CIFAR-10 dataset for 200 epochs

Figure 11: The results from training the diffusion model on the CIFAR-10 dataset for 400 epochs