VISION TRANSFORMERS FOR GENERATIVE MODELING

Rasmus J. Pedersen (s164564), Sarah M. Cournane (s210291), Miguel A. R. Alvarez (s212780)

DTU

ABSTRACT

The recent rise of attention-based architectures such as Vision Transformers (ViT) in computer vision tasks due to their computational efficiency, has not yet reached image generation tasks. Deep generative models (DGM) are still dominated by convolution based architectures, and only limited research has been done on the use of Vision Transformers in deep generative models. In this project, we propose a DGM design using ViT as the encoder in a Conditional Variational Autoencoder (ViT-CVAE). This ViT-based CVAE and a CNN-based CVAE were compared and shown to have similar performance both in reconstructing images and in generating diverse enough new samples to improve the performance of a classifier when used for data augmentation.

***Index Terms*—** ViT, VAE, Conditional VAE, GAN, Data augmentation.

1. INTRODUCTION

Deep generative models (DGM) are deep neural networks trained to approximate complicated, high-dimensional probability distributions. When trained successfully, they can be used to create new samples from the underlying distribution. Research and literature on DGMs is growing rapidly and some advances have even reached the public sphere with applications such as image super resolution, photo inpainting or face aging. [1]

In the past years, Convolutional Neural Networks (CNN) have been the way to go in the fields of computer vision and image processing. CNN-based Variational Autoencoders (VAE), and its Conditional version (CVAE), have become one of the most popular architectures for image generation. However, Vision Transformers (ViT) have recently achieved highly competitive performance in benchmarks for several computer vision applications, such as image classification and object detection. The aim of this project is to explore the potential of Vision Transformers within the field of generative modelling by comparing the image reconstruction results and performance as generators of CNN-based CVAE and ViT-based CVAE models combined with a GAN training for the data augmentation process in an image classification task on CelebFaces Attributes Dataset (CelebA).

2. BACKGROUND

2.1. Variational Autoencoder

An autoencoder is a neural network architecture consisting of two parts: an encoder to encode the input into a compressed (lower dimensionality) representation, commonly known as the latent space, and the decoder to decode it back such that the reconstructed input is as similar as possible to the original one [2]. For image data, convolutional layers are typically used in both the encoder and the decoder. In generative modelling of images, the decoder of a trained convolutional autoencoder can be used to generate new images from points in the latent space.

However, content generation from autoencoders is limited by the lack of continuity in the latent space. Variational Autoencoders (VAE's) present a continuous latent space by encoding inputs as a probabilistic distribution, usually a Gaussian distribution characterized by a vector of means and another of standard deviations. The decoder must learn to reconstruct the inputs from sampled values of the distribution estimated by the encoder. This is known as the "*re-parameterization trick*".

This is achieved by adding to the reconstruction loss a regularization term expressed as the Kullback-Leibler divergence between the returned distribution and a standard Gaussian.

Conditional Variational Autoencoder (CVAE) is an extension of Variational Autoencoders in which a condition is imposed on both the encoder and decoder inputs. The CVAE model offers an approach to control the data generating process of a VAE since the latent space is distributed under a conditional probability distribution.

2.2. Vision Transformer

Transformers are a self-attention-based architecture[3], which have revolutionized natural language processing (NLP) since its inception due to the architecture's computational efficiency and scalability allowing the training of massive models.

Vision Transformers (ViT), as proposed by Google Research Brain team[4], is an architecture that makes it possible to use the standard Transformer[3] for computer vision tasks, a field that is dominated by convolutional architectures. Transformers, originally used for NLP, use word-tokens as

input. For ViT to be able to use the standard transformer, its input also needs to have this token shape. To achieve this, ViT first divides the input image into patches of equal size, and then performs a patch embedding, which flattens the patches and maps them to D dimensions where D is the size of the latent vector of the transformer. Using this approach, the ViT architecture creates a token for each patch, which can be fed into a standard Transformer and can thus utilize the Transformers computational efficiency and scalability. Additionally a positional embedding with the same dimension D as the patch embeddings is learned and summed to the patch embeddings to retain positional information.

2.3. Generative Adversarial Network

Generative Adversarial Networks (GAN)[**?**] are a class of machine learning frameworks where two neural networks contest with each other:

- `Generator,` **G**: It tries to estimate the distribution of the training data. Its goal is to create "fake" images that look like the training images.
- `Discriminator,` **D**: It tries to estimate the probability that a data sample comes from the original training data and not from **G**.

During simultaneous training, both components play a min-max game and minimize their losses: the generator is constantly trying to outsmart the discriminator by generating better and better fakes, while the discriminator is working to become a better detective and correctly classify the real and fake images. The equilibrium of this game is when the generator is creating perfect fakes that look as if they came directly from the training data, and the discriminator is left to always guess at 50% confidence that the generator output is real or fake.

3. METHOD

3.1. Methodology

A CVAE using ViT in the encoder was designed, and a traditional fully convolution based CVAE was implemented. Initially, both the CNN-based CVAE and ViT-based CVAE generative models were trained. Then, the best performing model of each architecture evaluated on the test split was chosen. The chosen models were then used as the generator in a GAN structure and trained in an attempt to improve the visual quality of the generators. A pre-trained CNN-based classifier was fine-tuned on the training set as a baseline, additional classifiers, one for each generator, were fine-tuned with synthetic data from sampling the latent space distribution of the generators. The synthetic data was sampled in a manner such that each mini-batch had balanced classes, i.e. more images were generated of classes with low representation in the training set. The performance of the classifiers were evaluated on overall and individual class accuracy to determine the suitability of Vision Transformers within generative modelling.

Class \Split	**Train**	**Validation**	**Test**
Male Blond Hair	1.33%	1.35%	1.50%
Male Brown Hair	11.0%	12.48%	8.14%
Male Black Hair	20.87%	20.68%	21.95%
Female Blond Hair	23.31%	23.47%	20.85%
Female Brown Hair	24.14%	28.44%	23.02%
Female Black Hair	19.33%	13.57%	24.53%

Table 1: Class distribution for train, validation and test split.

3.2. Dataset

The dataset used was CelebFaces Attributes Dataset (CelebA) [5], a large-scale dataset with 202,599 colored celebrity face images, each with 40 binary attribute annotations. For the experimentation, 6 classes were defined based on the binary attribute annotations male, blonde, brown and black haired. Images labeled with none or more than one hair color were discarded to avoid ambiguity and to ensure distinction among classes. Therefore, from the original 162,770 training, 19,867 validation and 19,962 testing images, only 94,447 training, 11,711 validation and 11,510 testing images were used during experimentation. The dataset has an imbalance of the chosen classes, with particular blond males being underrepresented as can be seen in table 1.

3.3. Architecture

3.3.1. CNN-based CVAE

The encoder receives as input an image with its embedded label concatenated as an additional channel and it is composed of a stack of N=5 convolutional layers with 8^n filters of size 3x3, batch normalization and Leaky ReLU activation.

The hidden layer of the VAE contains two separated fully connected layers of the size of the latent space: one representing the mean vector (μ) and the other the standard deviation vector (σ). Then a vector with the dimensionality of the latent space is sampled following the actual values of μ and σ and the label is concatenated, the vector is then embedded using a learned embedding to get the correct dimension for the decoder.

The decoder is composed of a stack of N=5 transposed convolutional layers with 8^n filters of size 4x4, batch normalization and ReLU activation and a final transposed convolutional layer with 3 filters of size 4x4 and hyperbolic tangent activation to output the reconstructed image.

3.3.2. ViT-based CVAE

The architecture remains the same except for the encoder and the hidden layer. The encoder is a Vision Transformer that receives as input an image with its embedded label concatenated as an additional channel. The image is then reshaped into a sequence of flattened 2D patches, where (H, W) is the resolution of the original image, C is the number of channels (including the label channel), (P, P) is the resolution of each image patch and $N = HW/P^2$ is the resulting number of patches. Since the Transformer uses constant latent vector size D through all layers, the patches are flattened and mapped to D dimensions with a trainable linear projection to obtain patch embeddings [6]. Two learnable embeddings are concatenated to the sequence of embedded patches, whose states at the output of the Transformer's encoder serve as the mean vector (μ) and the standard deviation vector (σ). Finally, standard learnable 1D position embeddings are added to the patch embeddings to retain positional information. The resulting sequence of vectors is the final input to the encoder.

The Transformer encoder is composed of a stack of N = 7 identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism and the second is a fully connected feed-forward network. A residual connection is employed around each of the two sub-layers, followed by layer normalization. To facilitate these residual connections, all sub-layers in the model, as well as the embedding layers, produce outputs of same dimension D. [7]

The hidden layer simply samples a vector from the values of the μ and σ output tokens and the label is concatenated and embedded similar to the CNN-based CVAE.

3.3.3. PatchGAN

ViT-based CVAE is used as the Generator of the GAN structure. For the Discriminator, first a version inspired by the DCGAN was implemented. But influenced by the input of the ViT, the PatchGAN[8] Discriminator was finally used. PatchGAN is a discriminator which penalizes local image patches rather than the whole image. Discriminator tries to classify N × N patches as real or fake. Classifying patches rather than the whole image enables the Discriminator to run and train much faster.

The PatchGAN discriminator receives as input either a generated image from ViT-CVAE or a real one from the CelebA dataset. The Discriminator is formed by four blocks: within each one of them, there is a 2D Convolutional layer with kernel size 4, stride 2 and padding equal to 1, a normalization layer (except for the first block) and finally a LeakyReLU layer. After the four blocks, a final padding and a 2D Convolution are applied.

3.3.4. CNN-based Classifier

The classifier consists of a feature extractor followed by a fully-connected layer with 6 output neurons for classification. The feature extractor is a ConvNeXt Tiny model[9] pretrained on ImageNet with the classification layer removed and its weights frozen, thus only the added fully-connected classification layer is trained.

4. EXPERIMENTATION

4.1. Training

ViT-based and CNN-based VAEs together with the PatchGAN discriminator were trained with the AdamW optimizer (which is said to yield models that generalize much better and is thus able to compete with SGD while training much faster [10]). `ReduceOnPlateau` was the chosen learning rate scheduler and early stopping depended on the validation loss, therefore the model would stop training if no improvements were seen within 15 epochs. Training was done on a Nvidia A100 GPU. VAEs and GANs use different error measures during training which are describe below.

4.1.1. VAE

VAE's loss function consists of two parts:

- **Kullback-Leibler divergence:** It measures the difference between $\mathcal{N}(0, 1)$ and $\mathcal{N}(\mu, \sigma)$
- **Reconstruction loss:** Mean squared error loss between the output of the VAE and the real image.

4.1.2. PatchGAN

In addition to the generator's errors, the model was further trained by using the GAN structure. The output of PatchGAN's Discriminator is a 5×5 matrix where each position of the matrix corresponds to a patch in the input image and it determines whether the corresponding patch looks real or generated. Therefore the loss function was implemented as follows:

- **Discriminator Loss:** If the input image is real, the loss function is a Binary Cross Entropy loss within the Discriminator's output and a matrix full of ones of the same size. Contrarily, if the input image is fake then the BCE will have as target a zero matrix.
- **Generator Loss:** The Generator, in addition to the VAE loss defined above, also had a BCE loss between the output of the Disciminator when it received a generated image and a matrix full of ones.

(a) Conditional VAE

(b) Conditional ViT-VAE

Fig. 1: Generated faces conditioned on the different classes one for each class. Conditioned class from left to right male blond, male brown, male black, female blond, female brown, female black. See appendix C for more samples.

Model	ELBO	Reconstruction loss	KL-Divergence
CVAE	489.6	343.2	146.4
ViT-CVAE	490.2	347.5	142.7

Table 2: Comparison of evaluation metrics on the test split of the best performing models, for the CNN-based and ViT-based architectures.

	Baseline	CVAE	ViT-CVAE
Male Blond Hair	0.00%	**20.81%**	13.87%
Male Brown Hair	**90.50%**	84.63%	85.06%
Male Black Hair	59.00%	69.53%	**71.63%**
Female Blond Hair	**93.42%**	90.79%	90.46%
Female Brown Hair	**77.96%**	72.72%	71.28%
Female Black Hair	42.65%	61.03%	**65.92%**
Overall	68.21%	73.04%	**74.30%**

Table 3: Comparison of overall and class accuracy of classifiers using synthetic data from generators, and the baseline classifier with no synthetic data.

4.2. Results

Comparing the performance of the CNN-based and ViT-based CVAE's, we were able to achieve a similar performance using the ViT-based architecture compared to the traditional CNN-based architecture, as seen in table 2. Furthermore, comparing the samples in figure 1 we can see the conditioning is working well for both models, and both models can generate a variety of passable faces.

The attempt to improve the quality of the generators using the GAN structure did not yield acceptable results and its use was discarded. The simultaneous training of two competing models is inherently unstable and, despite the large amount of empirical trial and error, the training of a stable GAN model was not achieved. Thus no classifier was trained using the GAN generators.

The performance of the classifiers seen in table 3 show the baseline classifier is unable to classify men with blond hair, and perform poorly on especially black hair for both sexes. We also observe that by using synthetic data from the VAE's during training, the classifiers are able to classify some men with blond hair, and generally perform better on black hair of both sexes. It does this with only a slight decrease in performance on the baseline classifiers best classes and achieves a higher overall accuracy on the test dataset. The performance of CVAE and ViT-CVAE is again comparable with CVAE facilitating a bigger improvement on blond males and ViT-CVAE a bigger improvement on black hair for both sexes.

5. DISCUSSION AND CONCLUSION

In this work, we have explored the application of Vision Transformers in generative modelling of images. Concretely, we presented a generative model consisting of a Conditional Variational Autoencoder integrating a Vision Transformer as an encoder and additionally using this generator in a GAN structure, however, the later did not result in any improvements to the models quality and was dropped.

The results obtained by the ViT-CVAE are on par with the still dominant convolutional architectures. Our trained model was able to both reconstruct images successfully and capture variations and patterns in training data to generate diverse enough new samples to improve the performance of a classifier when used for data augmentation. The presented results, together with the computational efficiency and scalability of Vision Transformers, seem to indicate a promising future for self-attention-based architectures within generative modelling.

As noted in the original ViT paper[4], for classification tasks ResNet outperformed ViT with smaller pre-training datasets and the ViT only started to perform better with larger pre-training. Since we did not utilise pre-training for our generative models, it would be interesting to see further work investigate if this relationship between the architectures performance and size of pre-training dataset is the same for generative models.

6. REFERENCES

[1] Lars Ruthotto and Eldad Haber, "An introduction to deep generative modeling," 2021.

[2] Dor Bank, Noam Koenigstein, and Raja Giryes, "Autoencoders," 2020.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz

Kaiser, and Illia Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," 2017.

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, "Image-to-image translation with conditional adversarial networks," 2016.

[9] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," *CoRR*, vol. abs/2201.03545, 2022.

[10] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," 2017.

A. MODEL PARAMETERS

Parameter	ViT-CVAE	CVAE
depth	7	–
dim	1024	1024
mlp_dim	1024	–
dim_head	64	–
heads	16	–
patch_size	16	–
num_classes	6	6
ngf	48	32
lr	1e-4	1e-4
image_size	(128, 128)	(128, 128)

Table 4: Hyper-parameters of best performing models for ViT-CVAE and CVAE.

B. CODE

Link to code repository:
https://github.com/RasmusJuul/Vision-transformers-for-generative-modeling

C. GENERATED SAMPLES

C.1. CVAE

Fig. 2: Samples generated by CVAE

C.2. ViT-CVAE

Fig. 3: Samples generated by ViT-CVAE