TEXT-TO-IMAGE SYNTHESIS USING CONDITIONAL-VARIATIONAL AUTOENCODER

Ashir Raza, Jiwoo Kim, Naman Saxena & William Chen
Duke University

ABSTRACT

We designed and trained conditional variational autoencoders (CVAEs) for text-to-image generation, for both the FashionMNIST and the MS COCO. For FashionMNIST, the model generates images based on short text inputs. For MS COCO, the model leverages image-caption pairs during training, employing a pre-trained BERT encoder to transform captions into meaningful text embeddings. These embeddings then condition the CVAE image generation process. Overall, our CVAEs can handle both datasets and generate slightly meaningful images from different text inputs. Our code can be found here: https://github.com/wwillchen/text-to-image-cvae.

1 INTRODUCTION

Text-to-image generation has become a prominent area of research in deep learning, enabling models to create meaningful visual content based on textual descriptions. Conditional Variational Autoencoders (CVAEs) provide a framework for this task, allowing image generation to be conditioned on auxiliary information such as text or labels. In this project, we explore the application of CVAEs to two datasets, FashionMNIST and MS COCO. We outline our progression and implementation, as well as challenges and our solutions.

For the FashionMNIST dataset, we implemented a straightforward CVAE architecture that conditions image generation on class labels. By concatenating labels with the input data, the model learns to produce images that correspond to specific classes in the dataset, offering a clear baseline for evaluating the generative capabilities of CVAEs.

To handle the details of the MS COCO dataset, which pairs images with long, descriptive captions, we extended the Variational Autoencoder (VAE) architecture by integrating more complicated models. We first used a baseline CVAE model consisting of convolution layers for both encoding and decoding. We attempted additional modifications to in order to achieve better quality images, which we highlight in our ablation study. Specifically, we tried to incorporate a model with a ResNet backbone to enhance feature extraction and leveraged the VAE architecture utilized in Latent Diffusion Models (LDM)Rombach et al. (2021), known for its ability to generate high-resolution images. To better capture the intricate semantic relationships within the captions, we employed a pretrained Bidirectional Transformer (BERT) model. BERT effectively encodes the complex meanings of textual descriptions, providing robust embeddings. These embeddings were then fused with the visual features and processed through a Multilayer Perceptron (MLP) to learn a shared latent representation that bridges images and captions. The decoder was designed to integrate textual embeddings alongside the latent representation, enabling the generation of images that are semantically aligned with the provided captions.

2 PREREQUISITE

2.1 VARIATIONAL AUTOENCODER(VAE)

Variational Autoencoders (VAEs) are a class of generative models that leverage probabilistic inference to generate data. Unlike traditional autoencoders, VAEs learn a continuous probabilistic latent space, where each point represents a potential data instance. The encoder maps input data into a

Gaussian latent distribution, defined by a mean and variance. The decoder then samples from this latent distribution to reconstruct the input data. This probabilistic design enables VAEs to generate new, diverse instances by sampling from the latent space. Their ability to model complex data distributions makes VAEs a flexible model for tasks like data generation, representation learning, and interpolation.

2.2 Conditional Variational Autoencoder(CVAE)

Conditional Variational Autoencoders (CVAEs) extend the VAE framework to incorporate additional conditional information, such as class labels or additional features. By conditioning both the encoder and decoder on the additional information, CVAEs learn a latent space that is informed by the provided context. This allows for more controlled generation, rather than randomly generating as the traditional VAEs. For instance, when applied to image generation, a CVAE can produce images of a specific category by conditioning on a label. This advantage makes CVAEs more suitable for tasks requiring targeted generation or representation learning, such as image captioning, style transfer, and multimodal data synthesis.

2.3 Bi-directional Transformer (BERT)

Unlike traditional models that process text sequentially, BERT(Devlin et al. (2019)) utilizes a bidirectional architecture to capture the full context of a word by considering both its left and right surroundings simultaneously. This deep understanding of textual semantics enables BERT to excel in a wide range of NLP tasks, including sentiment analysis, question answering, and text classification. Pretraining BERT involves masked language modeling (MLM) and next sentence prediction (NSP), ensuring it learns rich contextual embeddings that can be fine-tuned for specific downstream tasks. Its versatility and effectiveness make it a critical component for tasks requiring deep textual understanding, such as caption generation and multimodal representation learning.

3 Method

In this section, we describe the datasets we have used and the different conditional VAE models used for training on those datasets.

3.1 Dataset

For this project two datasets were used: FashionMNIST and MS COCO. FashionMNIST dataset contains images of 10 different categories of objects: T-shirt, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, and Ankle boot. Each image is a grayscale image with dimension 28 x 28. The value of each pixel in the image range from 0 to 255. The dataset contains in total 70,000 images with 60000 being labeled as training set and 10000 labeled as test set.

The next dataset we train our model is Microsoft Common Objects in Context(MS COCO) dataset. This dataset is designed to test vision models on several tasks such as object detection, image segmentation, and image captioning. It contains more than 328,000 images divided across 80 object classes. For the purpose of object detection it contains bounding box coordinates of each object. Further, it contain pixel wise segmentation mask for segmentation related task. Finally, for the image captioning it contains 5 captions corresponding to each image.

3.2 FashionMNIST Model

For the FashionMNIST CVAE, since the posterior distribution $p_\theta(z \mid x, c)$ is computationally intractable, we hope to estimate it with a separately parametrized distribution $q_\phi(z \mid x, c)$, thus seeking to minimize $D_{\mathrm{KL}}(q_\phi(z \mid x, c) \| p_\theta(z \mid x, c))$

We know that as per Blei et al. (2011), $D_{\mathrm{KL}}(q_\phi(z \mid x, c) \| p_\theta(z \mid x, c)) = -ELBO + log(p(x|c))$. Thus, since $log(p(x|c))$ is an independent term, minimizing $D_{\mathrm{KL}}(q_\phi(z \mid x, c) \| p_\theta(z \mid x, c))$ is the same as minimizing negative $ELBO$, where $ELBO = \mathbb{E}_q[\log p_\theta(x, z|c)] - \mathbb{E}_q[\log q_\phi(z \mid x, c)]$.

Thus, we seek to maximize ELBO with respect to the label c:

$$\begin{aligned}
\text{ELBO} &= \mathbb{E}_q\left[\log p_\theta(x, z|c)\right] - \mathbb{E}_q\left[\log q_\phi(z \mid x, c)\right] \\
&= \mathbb{E}_q\left[\log p_\theta(x \mid z, c) p_\theta(z|c)\right] - \mathbb{E}_q\left[\log q_\phi(z \mid x, c)\right] \\
&= \mathbb{E}_q\left[\log p_\theta(x \mid z, c)\right] + \mathbb{E}_q\left[\log p_\theta(z|c)\right] - \mathbb{E}_q\left[\log q_\phi(z \mid x), c\right] \\
&= \mathbb{E}_q\left[\log p_\theta(x \mid z, c)\right] - \mathbb{E}_q\left[\log \frac{q_\phi(z \mid x, c)}{p_\theta(z|c)}\right] \\
&= \mathbb{E}_q\left[\log p_\theta(x \mid z, c)\right] - D_{\mathrm{KL}}(q_\phi(z \mid x, c) \| p_\theta(z|c))
\end{aligned}$$

We then employ the reparametrization trick by Kingma & Welling (2013) for $q_\phi(z \mid x, c)$, using $\epsilon \sim N(0, \mathbf{I})$, $(\boldsymbol{\mu}, \log \boldsymbol{\sigma}) = \text{EncoderNeuralNet}_\phi(\mathbf{x}, \mathbf{c})$, and $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$.

Thus, we can note that $q_\phi(z \mid x, c) \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$, a diagonal Gaussian distribution with no nonzero covariance terms. Thus, this can be percieved as a joint distribution of n independent univariate Gaussian distributions. Thus, we can write:

$$\begin{aligned}
D_{\mathrm{KL}}(q_\phi(z \mid x, c) \| p_\theta(z \mid c)) &= \frac{1}{2}\left[\mathrm{tr}(\Sigma_q) + \mu^\top \mu - k - \log\det \Sigma_q\right] \\
&= \frac{1}{2}\sum_{i=1}^{k}\left[\sigma_i^2 + \mu_i^2 - 1 - \log \sigma_i^2\right]
\end{aligned}$$

The full negative $ELBO$ loss can be written as:

$$-ELBO = \frac{1}{2}\sum_{i=1}^{k}\left[\sigma_i^2 + \mu_i^2 - 1 - \log \sigma_i^2\right] - \log p_\theta(x \mid z, c)$$

In applying this to FashionMNIST, we use a encoder module that took in the flattened 784 x 1 vectors of each image and concatenated a 10 x 1 one-hot encoded vector representation of each class conditioning. We then feed the augmented data through 2 fully connected layers, with three different sets of weights $\boldsymbol{W}_1, \boldsymbol{W}_{21}, \boldsymbol{W}_{22}$ to obtain a vectorized representation of $\boldsymbol{\mu}$ and $\log \boldsymbol{\sigma}$ respectively.

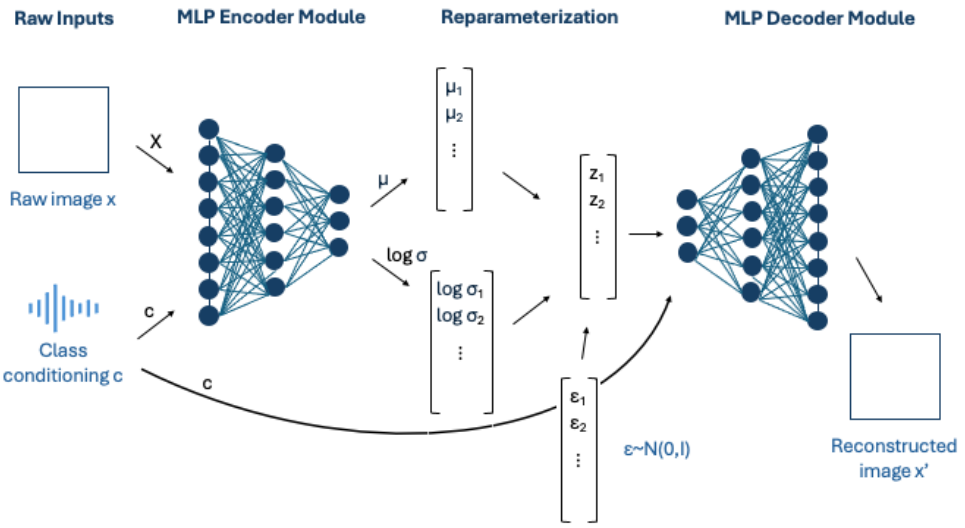

Figure 1: CVAE Architecture used for FashionMNIST Dataset

We then draw a vector $\boldsymbol{z} \sim N(0, \mathbf{I})$ and reparameterize it with $\boldsymbol{\mu}$ and $\log \boldsymbol{\sigma}$ as per described above. We then take $\boldsymbol{z}$ and concatenate it with the 10 x 1 one-hot encoded vector representation of the class conditioning. We then feed the augmented data through 2 fully connected layers, with two different

sets of weights $\boldsymbol{W}_3, \boldsymbol{W}_4$ to obtain the generated image. Figure 1 shows the overview of our data pipeline and model architecture.

To train the encoder and decoder modules of the CVAE, we backpropagate over the above negative ELBO loss with respect to each layer of weights in both modules. For the second term in the negative $ELBO$ loss, we assume pixel values are binary and thus use a Bernoulli distribution for each pixel. This leads us to the corresponding loss being the binary cross-entropy loss as we aim to minimize the negative log likelihood of the Bernoulli distribution.

This approximation simplifies the probabilistic modeling of image data. Intuitively, a pixel in a FashionMNIST image might represent whether a specific part of an article of clothing is visible. Treating the pixel as Bernoulli-distributed simplifies this representation and allows the model to focus on learning probabilities of binary patterns in the image.

3.3 MS COCO Model

Our encoder consist for 4 convolution layer each with filter of size 4. The number of channel increase in the following order: $3 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512$. The output of the final convolution layer is concatenate with the text embedding generated by BERT model. We have even tried adding instead of concatenation in the ablation studies. After concatenation, linear layer is used to generate mean and log standard deviation in the latent space for the image and text pair.

In the decode part, we first concatenate text embedding with the latent feature sampled from Gaussian distribution obtained from the encoder while training. The concatenated feature is passed through transpose convolution operator for image reconstruction. The number of channel while decoding decrease in the following order: $512 \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 3$. The entire architecture is given in Figure 2.

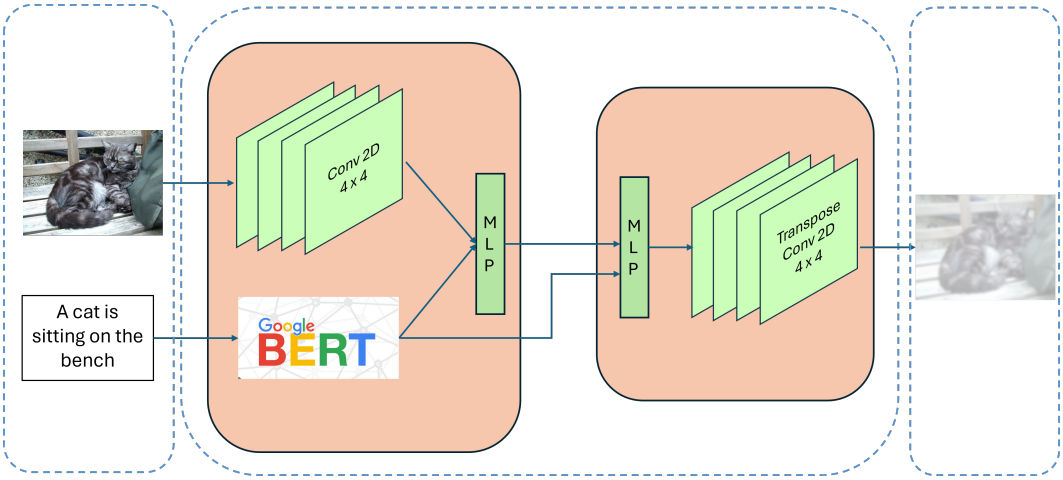

Figure 2: CVAE architecture used for MS COCO dataset

4 Experiment

4.1 CVAE for FashionMNIST

We used a learning rate of $\alpha = 0.001$, a hidden layer size of 400 for both $\boldsymbol{W}_1$ and $\boldsymbol{W}_3$, a batch size of 128, and a latent space dimension of 20 for $\boldsymbol{z}$. Upon training for 50 epochs, we visualized the latent space using sklearn.manifold.TSNE to map the $\boldsymbol{\mu}$ to 2D manifold, as can be seen in Figure 3.

The Conditional VAE has a connected continuous latent space, and it mixes the different classes thoroughly within the latent space than the vanilla VAE. We hypothesize this is because the Conditional VAE already takes class into account when encoding a given image, so it doesn't need to

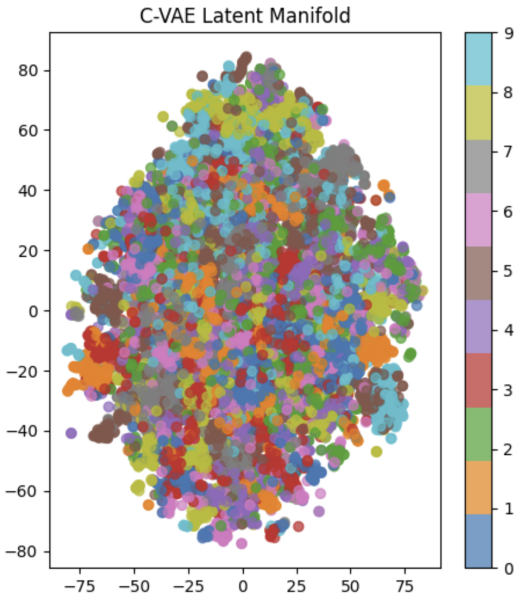

Figure 3: Latent Space of CVAE

separate them out based on class. We have the visualize the performance of the generation of the CVAE over the 10 different classes in Figures 4 and 5.

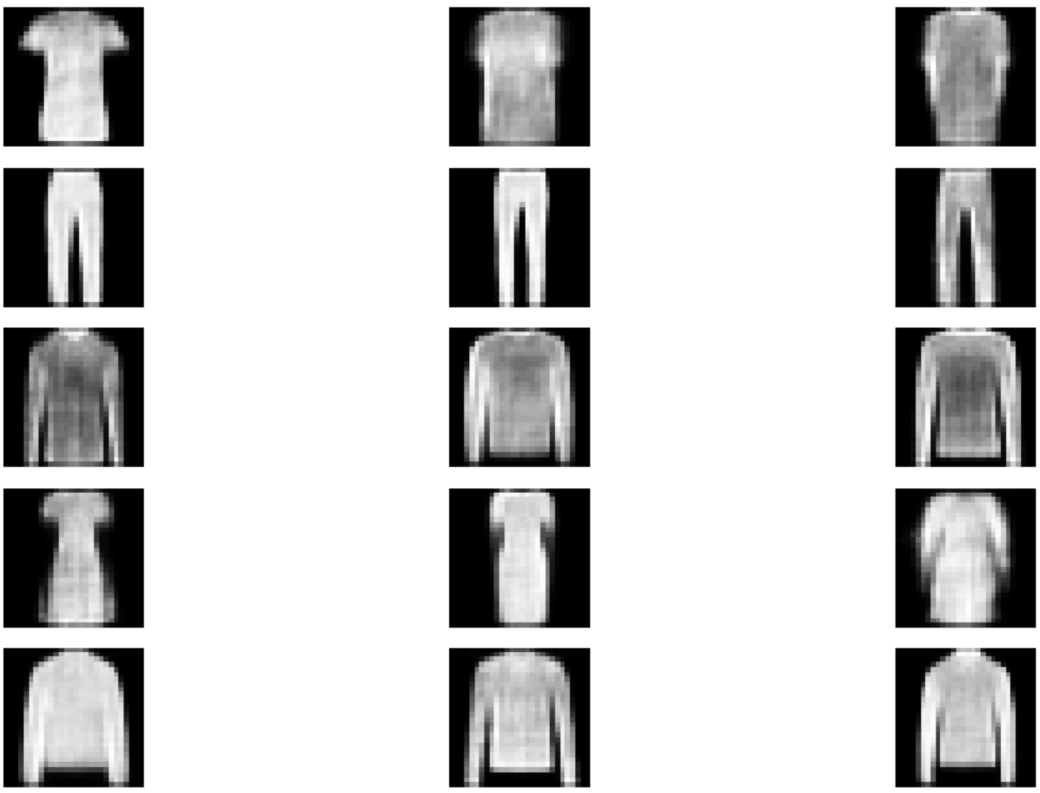

Figure 4: Visualizations for 0: T-shirt/top, 1 : Trouser, 2: Pullover, 3: Dress. 4: Coat

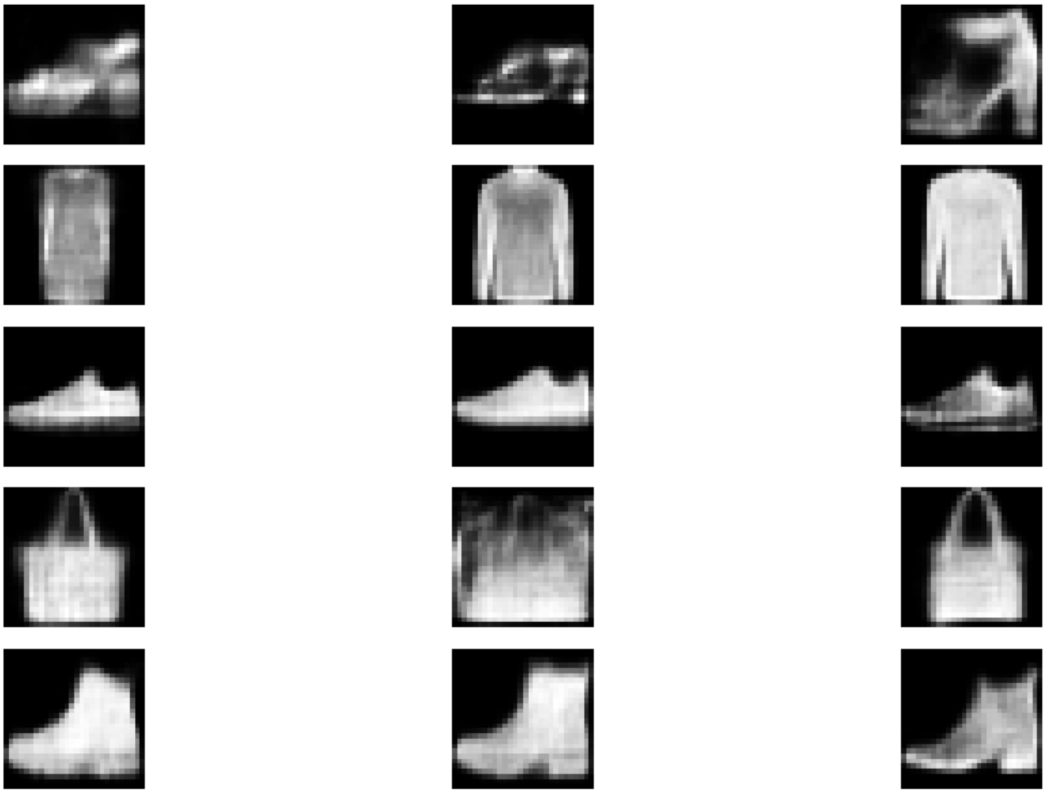

Figure 5: Visualizations for 5: Sandal, 6: Shirt, 7: Sneaker, 8: Bag, 9: Ankle boot

4.2 CVAE FOR MS COCO

We have provided the result of image generation for the MS COCO dataset in Figure 6. Before training we reduce the size of the images in the dataset to 64 x 64. We trained our model using Adam optimizer with learning rate $\alpha = 0.0001$. We used latent feature with 128 dimension and trained our model for 20 epochs on .30 of the dataset.

4.3 ABLATIONS

In an attempt to gain better results, various ablations have been tested.

In the original CVAE, the image features and text embeddings were concatenated and processed by an MLP. For the ablation study, we modified this approach by replacing concatenation with element-wise addition of the two embeddings (see Figure 8). The reason behind this change was to simplify the representation space for the MLP to process. However, experimental results demonstrated that this method underperformed compared to the original concatenation approach. This suggests that concatenation preserves richer feature information of each emebedding and better supports the model in capturing the complex relationships between images and text.

To leverage the robust feature extraction capabilities of ResNet blocks, we incorporated them into the encoder and decoder of the CVAE. Given ResNet's ability to learn high-dimensional features, this modification was expected to improve performance. However, training with the same dataset size and number of epochs yielded suboptimal results (see Figure 7). This limitation can be attributed to the model's increased complexity, necessitating more extensive data and longer training durations to fully exploit its potential. Despite its underwhelming performance in this setup, the inclusion of ResNet blocks remains a promising direction, contingent on adequate computational resources and data availability.

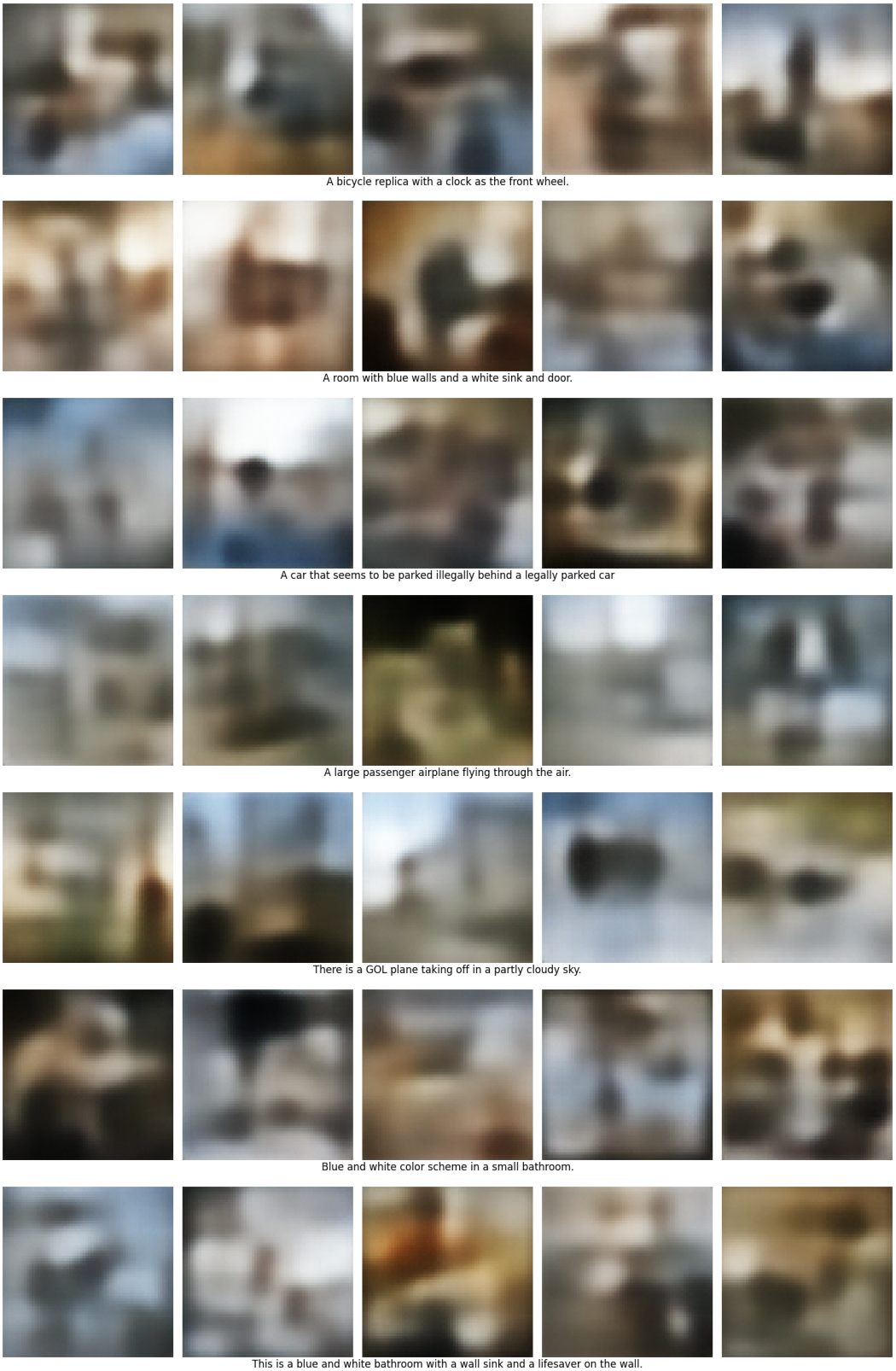

Figure 6: Image sampled from Conditional VAE

Unlike standard convolutional conditional VAEs, which often have challenges in capturing both high-dimensional image features and the complexity of diverse text embeddings, we adopted the Latent Diffusion Model (LDM) Rombach et al. (2021) VAE. This model is specially designed to generate high-resolution images by combining multiple convolutional layers, ResNet blocks, and attention mechanisms to efficiently handle long-range dependencies for spatial features. To tailor the architecture for our task, modifications were made by removing the original LDM loss, which incorporates the Learned Perceptual Image Patch Similarity (LPIPS) metric Zhang et al. (2018), with the traditional VAE loss. Moreover, to achieve optimal results, we utilized most of the original model's parameters during training. The results of the modified architectures did not surpass the performance

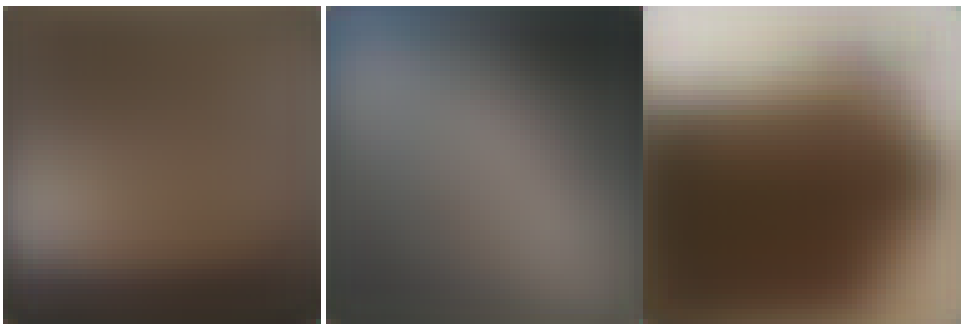

Figure 7: Image generation using ResNet based CVAE

of the base model. Training the full MS COCO dataset with this large architecture would typically require several days and substantial computational resources, including eight NVIDIA V100 GPUs. Due to resource constraints, we trained the model on only 5% of the MS COCO dataset. This limited training set resulted in underwhelming performance, as the reduced dataset size likely hindered the model's ability to fully capture the complex relationships between images and captions. Although the training loss decreased significantly, the model struggled to generalize effectively due to insufficient data diversity and volume. With access to adequate computational resources and the full dataset, we expect the model to achieve significantly better performance, showcasing its potential for high-quality image generation conditioned on textual descriptions.

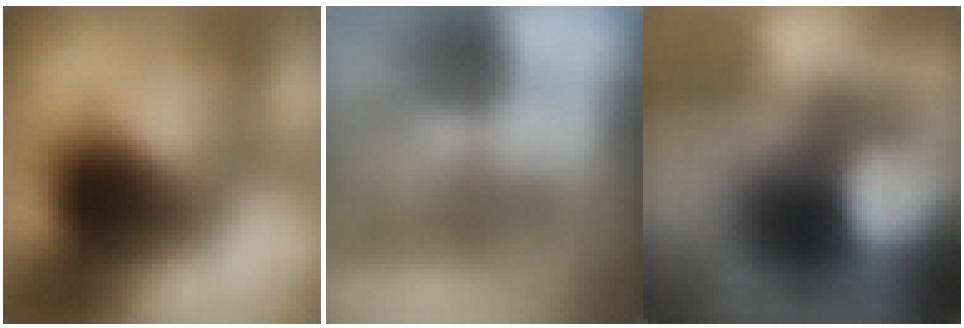

Figure 8: Image generation using addition instead of concatenation in CVAE

5 Conclusion

In this work, we developed a text-to-image synthesis model for both Fashion MNIST and MS COCO datasets. The model performed well on Fashion MNIST, generating clear images that aligned with the input text. However, the models performance on the more complex MS COCO dataset was limited due to resource constraints and a significantly reduced training dataset (5% of the total). While the architecture showed promise, the lack of sufficient data and computational power restricted the model's ability to fully capture the dataset's complexity. With adequate resources and training on the full dataset, the model has the potential to achieve significantly better results in generating high-resolution, contextually accurate images.

REFERENCES

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians, 2011. URL https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL https://arxiv.org/pdf/1312.6114.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.