# Generative CLIP-Conditioned Variational Autoencoder

Ariana Azarbal, Ayman Benjelloun Touimi, Sofia Tazi

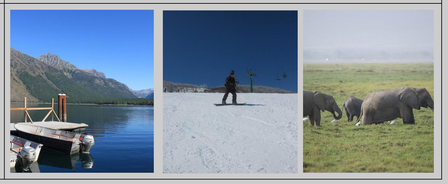
BROWN

## Introduction

In recent years, the intersection between language and imagery in Deep Learning has garnered increasing attention, particularly with the development of CLIP by OpenAI. CLIP maps text and image data into a shared latent space, opening new avenues for multimodal research. Our project was inspired by OpenAI's 2022 paper, "Hierarchical Text-Conditional Image Generation with CLIP Latents," which explores text-to-image generation using CLIP embeddings. Instead of implementing a computationally intensive diffusion model, we have conditioned a Variational Autoencoder (VAE) with CLIP embeddings. We explore conditioning with both image and text embeddings, leveraging CLIP's encoding of semantic and structural visual/textual information.

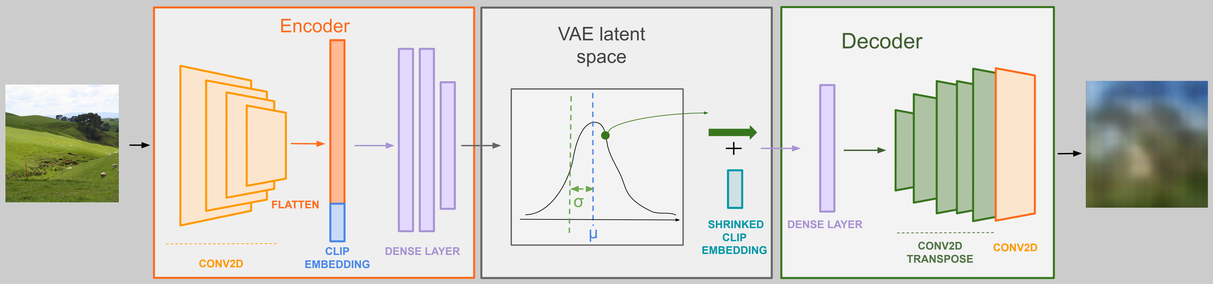
## Dataset

We refined a subset of the MS COCO dataset by selectively parsing captions for bucolic and natural themes, such as "cow," "sheep," "grass," and "mountains". The images were formatted in 64x64 or 128x128 resolutions, enhancing the visual distinction between elements.

| | |
|---|---|
| Small dataset *(cow, sheep)* |  |
| Big dataset (*cow, sheep, mountain, grass, forest, landscape)* |  |

## Architecture

Our model is a **Conditional Variational Autoencoder (CVAE)**, a type of generative model which learns a dataset's distribution to produce new samples. It has **≈32 million parameters.**




Our **encoder** consists of two parts:

**4 Conv2D Layers** (256 3x3, 256 5x5, 128 3x3, 128 3x3), strides of 2, same padding, Kaiming-initialized weights, Leaky ReLU & Batch Norm.

**3 Dense Layers** (output: 2048, 2048, 1024), Kaiming-initialized weights, Leaky ReLU and Batch Norm. The first layer takes, as an input, the output of the convolutional layers **concatenated with a CLIP embedding.**

For our **latent representation:**

**2 Dense layers** with output size 128 and Kaiming-initialized weights compute two vectors, **μ**, and **σ**. A **Dropout,** with rate of 0.3, is applied to the obtained sample **z.**

**1 Dense Layer** with output size 32 and Kaiming-initialized weights, shrinking the CLIP embedding before **concatenation**.

Our **decoder** consists of two parts:

**1 Dense Layer** with output size 16,384, Kaiming-initialized weights, Leaky ReLU and Batch Norm.

**5 Conv2D Transpose Layers** (256 3x3 stride=2, 256 5x5 stride=2, 128 5x5 stride=2, 128 3x3 stride=1, 32 5x5 stride=2) with Kaiming-initialized weights, Leaky ReLU, Batch Norm, same padding.

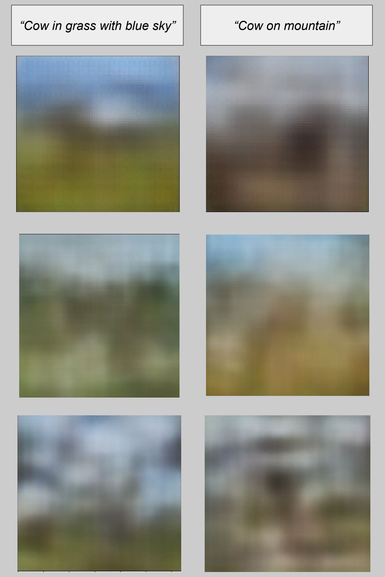
**1 Conv2D Layer** (3 5x5 stride=1), with Kaiming-initialized weights, same padding, and sigmoid activation.

**Loss Function:** We used a combination of **Binary Cross Entropy** to determine reconstruction loss, **KL-Divergence,** and **Structural Similarity.**

$$L = L_{\text{BCE}}(x, \hat{x}) + 0.5 D_{\text{KL}}(N(\mu, \sigma), N(0,1)) + 0.1(1 - \text{SSIM}(x, \hat{x}))$$

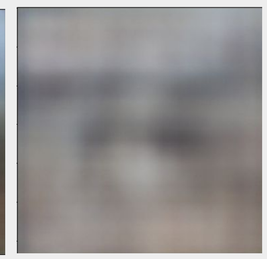
## Results

| Model | Best Validation Loss | Qualitative Observations |
|---|---|---|
| **64x64 Im. Embed-Conditioned** | 2376 (Epoch 25) | Pixelated, but decent with features. |
| 64x64 Text Embed-Conditioned | 2526 (Epoch 8) | Poor. |
| 128x128 Im. Embed-Conditioned | 9227 (Epoch 25) | Effective with colors, shapes, features like "sky", "grass", "sheep". |
| **128x128 Text Embed-Conditioned** | 9701 (epoch 9) | Decent with colors, shapes. |
| **128x128 Im Embed-Conditioned (No Dropout)** | 9386 (Epoch 10) | Effective at high-level features. Blurry. |
| 128x128 Text Embed-Conditioned (No Dropout) | 9085 (Epoch 10) | Worse images despite lower loss. |

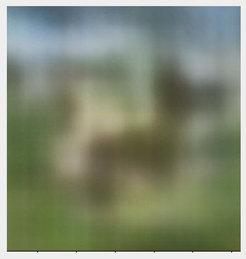



**Most effective model : 128x128 Im. Embed-Conditioned**

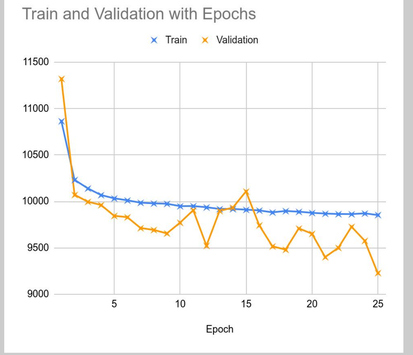
*"White sheep on a mountain under blue sky"*



*"Cows playing in the snow"*



*"Cow in green grass under blue sky"*




## Discussion

**Lessons Learned:** We were impressed by the richness of the CLIP latent space and its effective ability to bridge the gap between textual and graphic representations and introduce semantic understanding fairly easily: our model performs better when trained on image embeddings and asked to predict on text embeddings, as opposed to being trained on text embeddings directly.

**Limitations:** Images produced by our model are still very blurry, despite significant improvements throughout the course of the project. The model also struggles to generate fine-grained features, like specific objects. The dataset is also fairly limited and specific, making it difficult to generalize to broader image generation.

**Future Work:** Complexifying the decoder could help with image sharpness and generation of fine-grained features. Experimenting with different weights for the loss function may also lead to more consistent results.

## References

**Reference paper:** "Hierarchical text conditional image generation with CLIP Latents" Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen

**Dataset:** MS COCO : common objects in context.

## Acknowledgements

We would like to thank our professor **Ritambhara Singh,** our mentor TA, **Michael Lu** and **Calvin Luo** for giving us tremendous support and guidance.