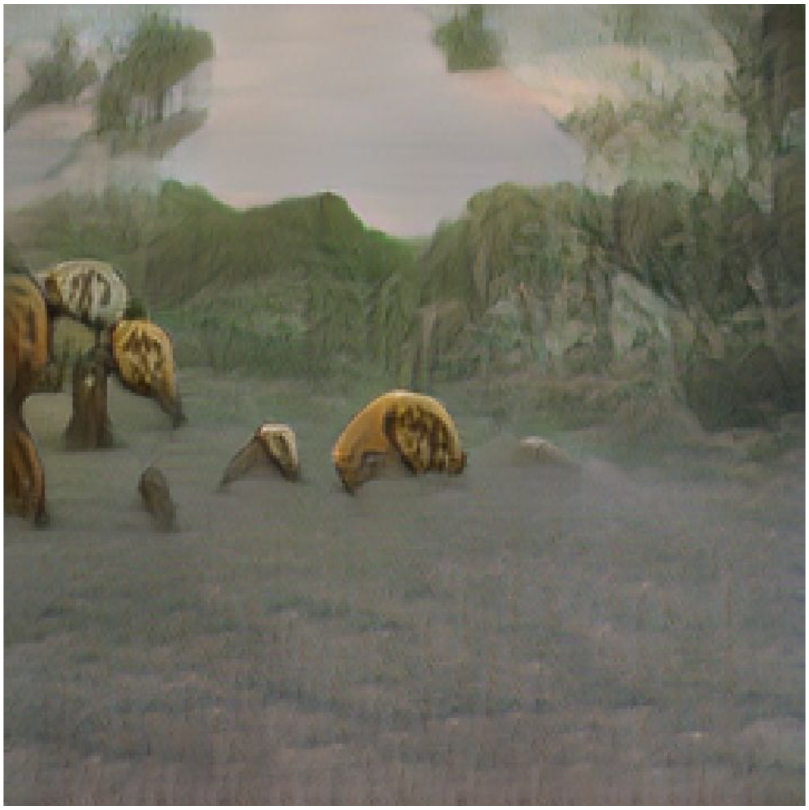
ART AND MACHINE LEARNING
CMU 2019 SPRING
FINAL PROJECT

Poem2Image

Poem :
Three rings for the pasta, under the sky
Four for the trees, in their halls of stone
And nine for the mortal fish, doomed to leaves
In the land of Mordor, where the shadows lie

Generated Art:



Prajwal Prakash Vashist , Sunil Kumar , Yashovardhan Chaturvedi

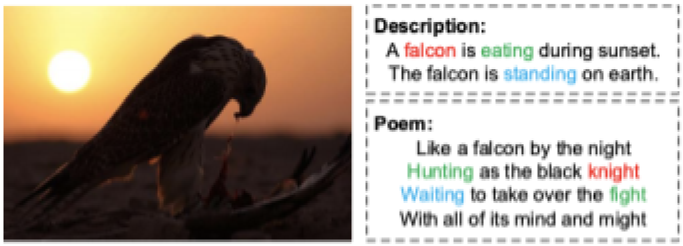
Concept:

We wanted to explore if GANs can be trained to illustrate abstract concepts that hold artistic value. Our task consists of generating images from snippets of Poetry. This hasn't been addressed in current research and neural networks have conventionally not performed well on the task of interpreting natural language. We also wanted to build an interface that would make it convenient for people to interact with our model through a website where a user can enter his own poetry and generate results. We feel this interface serves an additional purpose of helping people understand better the inner workings of a GAN especially people who are not well versed with implementation details of GANs.

Related Work

While planning and researching for this project, we relied heavily on the work done in the following two papers: Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training by Liu et al. [1] and AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks by Xu et al. [2].

Liu et al. utilized GANs as a way to process images and then generate text in a poetic form that also reflected the content of the image. Specifically, they broke up the problem into two distinct tasks. First, a deep CNN network was used to perform image captioning on the input. The caption is then fed to a GAN that utilizes an RNN-based generator to convert these captions from plain text to a poetic form. To this end, the authors collected a dataset of image to poem pairs. In addition, they also utilized a dataset of unpaired poems. The discriminator in the GAN was then asked to classify the output of the generator as a paired poem, a generated poem, or an unpaired poem. Another discriminator was also used to classify the poetic-ness of the output. An example for this process can be seen in Figure 1.



Description:
A falcon is eating during sunset.
The falcon is standing on earth.

Poem:
Like a falcon by the night
Hunting as the black knight
Waiting to take over the fight
With all of its mind and might

Figure 1 [1]

The work done by Xu et al. was crucial in the development of our project. Their paper introduced the AttnGAN model, which utilized attention to attend to various portions of the image based on the specific words in the input. Their work focused primarily on the creation of images based on highly descriptive texts, as seen in Figure 2, as opposed to the highly abstract forms we normally expect from poems.

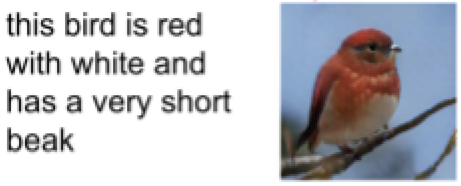

this bird is red
with white and
has a very short
beak

Figure 2 [2]

Technique:

Because the dataset used by Liu et al. already came with poem and image pairings, it was an ideal and convenient choice for us to use with our model. Thereafter, we simply had to adapt the AttnGAN model proposed by Xu et al. to work for our specific case. The structure of the AttnGAN model can be seen in Figure 3.

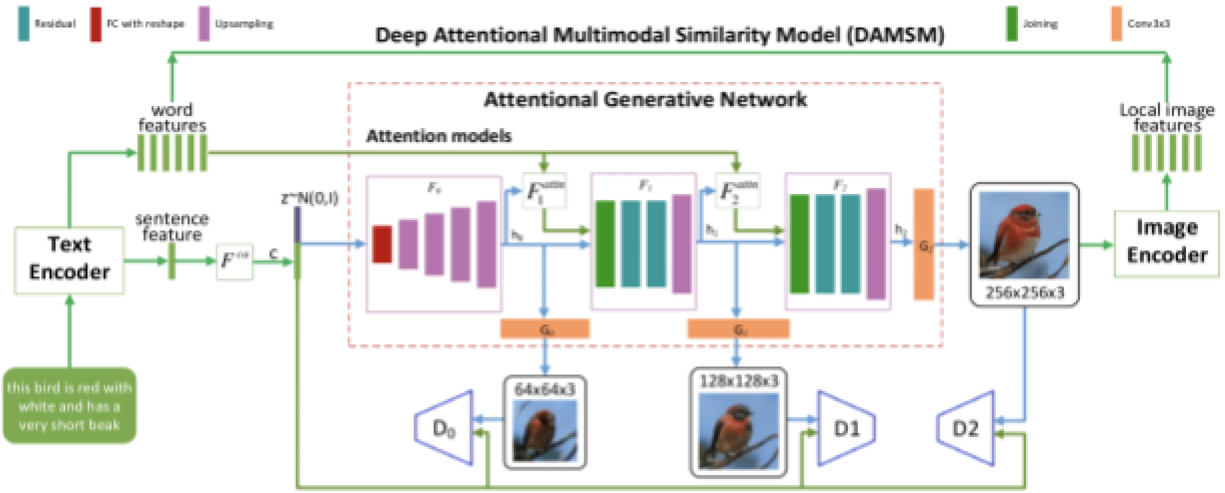


Figure 3

Here, we can see that the input text is first fed to an RNN-based text encoder. This encoder extracts both sentence level and word level features from the input. These features are then fed to the GAN. The sentence level features are used first to create a rough image that vaguely represents the expected output. This can be seen in the 64x64x3 image fed to the D0 discriminator. At this point, the word level features are then fed to the model as attention states, allowing the model to attend to various regions within the image based on the word features. These states are fed to the next stage of the generator along with the previous image, and model is able to create an image with more specific features. For example, elements like small beak are not present in the initial image, and instead we see just a vague outline of a bird.

However, after feeding the attention states, the model has learned to attend to specific segments of the input text, such as ‘short beak’, or ‘red and white’, and we can see that the next 128x128x3 image that is produced has these features more clearly distinguished. This process is applied again to further increase the resolution of these features and obtain a generated image that captures the essence of the input text. In addition to the text encoder and GAN, we also utilize an CNN-based image encoder to process the final output and extract salient features from the images. The final component of the model is the DAMSM (Deep Attentional Multimodal Similarity Model). This is specifically used during training to measure the similarity between the text features and the image features, ensuring that the final image is still relevant to the input text.

Process:

Luckily, due to the nature of the AttnGAN model, adapting this network to work with poems did not require a massive overhaul of the model. Instead, we simply had to get the model to learn the interactions between components of the poems and their related images. Since the text encoder is RNN based, we do not need to worry about the varying lengths of these poems.

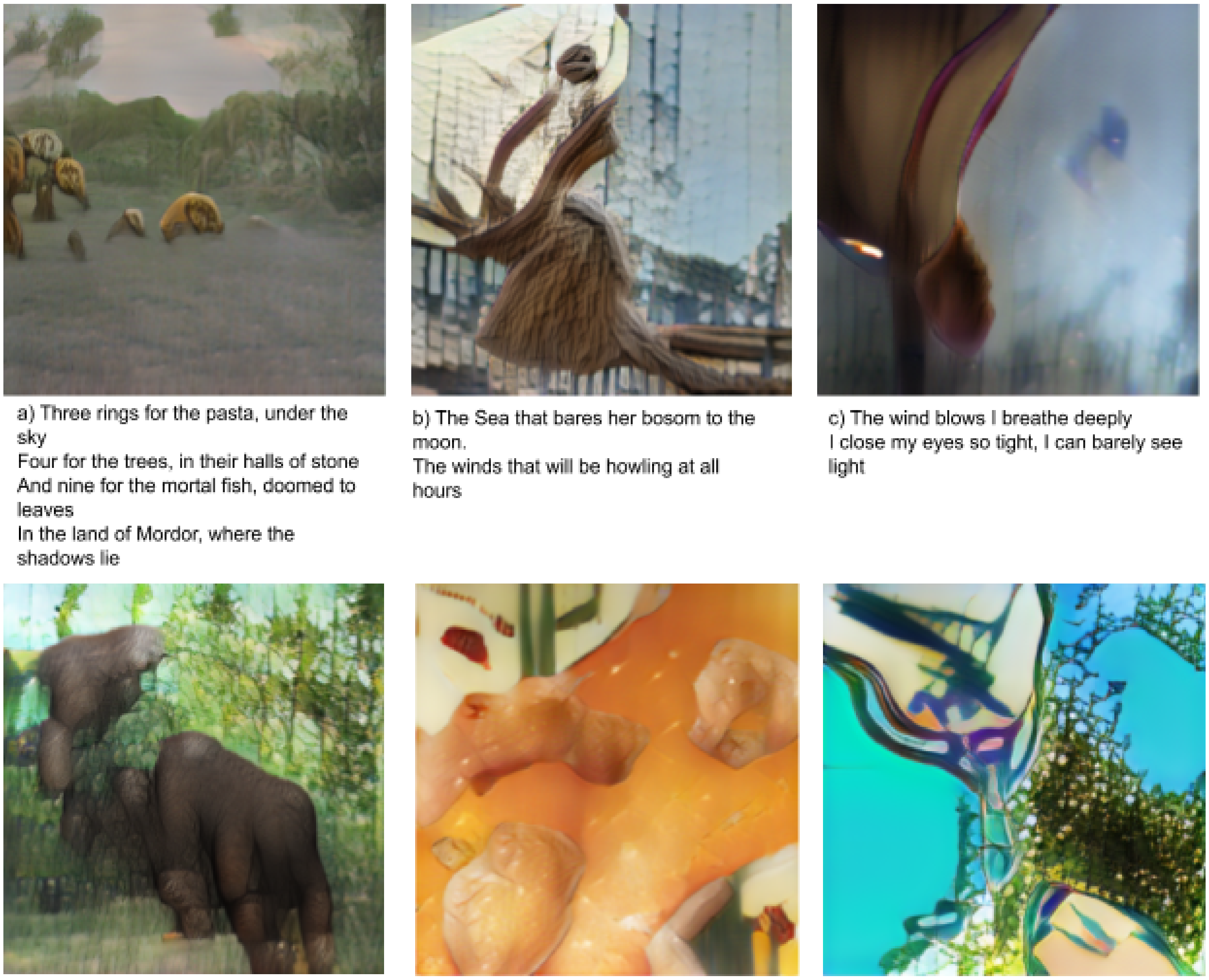
We first experimented with training the network from scratch. After training for three days we were not getting any interesting results at all and realized that the network will probably take more than a week to train from scratch.

We then decided to start training for the AttnGAN using pre-trained weights for the network which were obtained from training on the COCO dataset. One of the reasons we felt that would work is because the COCO dataset contains 91 classes that represent real-world everyday objects. As a result, we felt that these concepts must be learned by the network and would have been harder to learn if we had trained the network from scratch. However, in order to extract the text features, we trained the text encoder from scratch. To obtain the best results, we experimented with training several different text encoders by changing the underlying architecture for the RNN cells, with LSTM, GRU, and normal RNN cells. In the end, we found that LSTM’s worked the best, proceeded to conduct the rest of our experiments utilizing the encoder built around LSTM’s.

The AttnGAN was designed to generate fine grain images. However, we felt that training fine-grained images from poems would not achieve the desired effect of generating “visually appealing” images from poems. We also wanted to obtain a more abstract result such that it would allow users to more easily relate or interpret the images on their own. As a result, we wanted the network to focus more on the colors and overall visual aesthetics of the concepts present in the poems and not focus on getting the details of the concepts right, leading to more abstract images. In order to achieve this, we augmented our training dataset by blurring the images and then training the model. We felt that blurring the images would allow the network to pick up more abstract concepts, such as associating colors with words. For example, a bright sunny day would be associated with yellow colors, while a moon would be associated with a white circular patch and not on the exact details or spots on the surface of the moon. As a result, the network became easier to train, as we felt we had simplified the underlying distribution to be learnt by the model. We have explored the learning of these abstract concepts in the results section below.

Result:

The images in Figue-4 were generated using the stock AttnGAN.



a) Three rings for the pasta, under the sky
Four for the trees, in their halls of stone
And nine for the mortal fish, doomed to leaves
In the land of Mordor, where the shadows lie

b) The Sea that bares her bosom to the moon.
The winds that will be howling at all hours

c) The wind blows I breathe deeply
I close my eyes so tight, I can barely see light

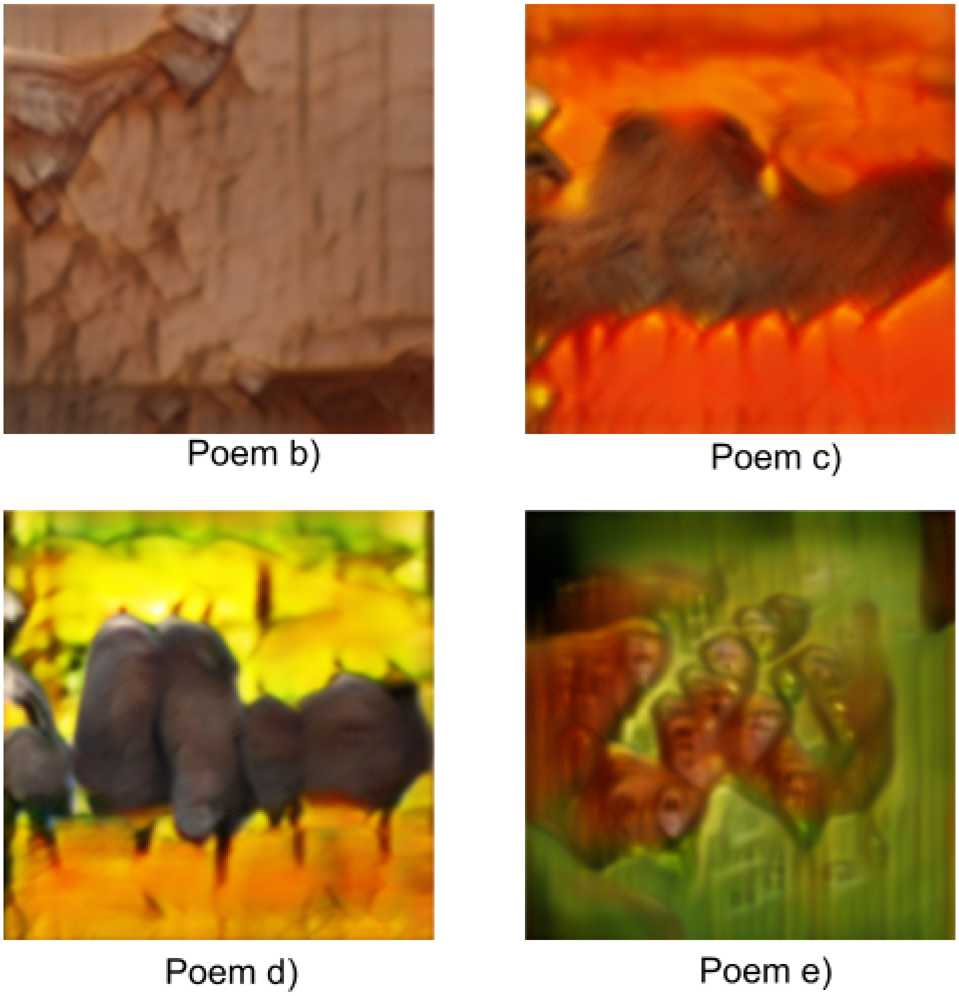
d)The elephant is a thing to behold,
with colors more beautiful than gold.

e)They say it's sweet that we laugh because our bodies literally can't contain the joy.

f) The butterfly is a thing to behold,
with colors more beautiful than gold.

Figure-4

The images in Figure-5 were generated by the model trained using transfer learning on blurred images on some of the same poems as above as indicated below the generated images.



Poem b)

Poem c)

Poem d)

Poem e)

Figure 5

Reflection

The model that generated the images in Figure-4 learns to map keywords in the poems to specific concepts/objects. For example, trees in poem a), features of an elephant in poem d).
On the whole, the generated images are somewhat related to the poem and are not very abstract.

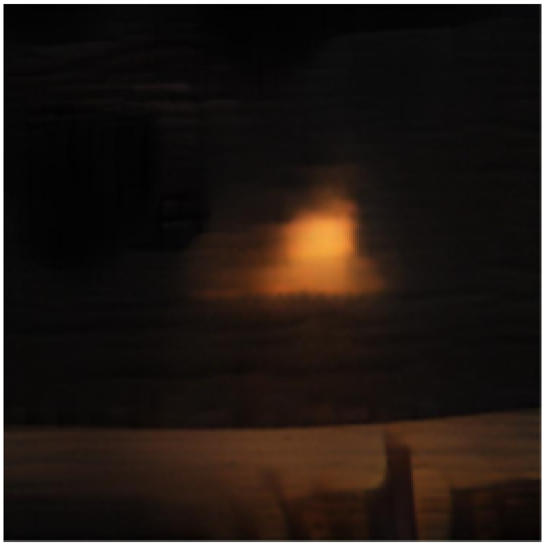


Figure 6 - A stormy night on a beach is dark and gloomy but not to worry because it is about to get sunny and sunny

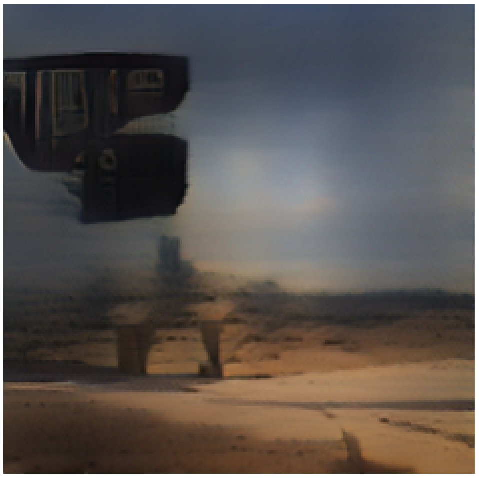


Figure 7 - A stormy day on a beach is dark and gloomy but not to worry because it is about to get sunny and sunny

The concepts of sun, night and beach are clearly depicted in Figure-6.
In figure-7, the poem differs from that of figure-6 by a single word, “A stormy **day**” instead of “A stormy **night**”. This difference in apparent in the color and tone of the generated images.

The idea for training the network with blurred images was to try to prevent the model from learning specific features of the images in the dataset but rather learn the colors/ moods associated with the text. This, as a result, generated more colorful and abstract art as portrayed in Figure 5.

In the end, we decided to use the images generated by the stock model as we felt they were more aesthetically pleasing and relevant to the poems associated with them. These images also seemed to match the expectations of the users of our system during the exhibition.

Reference:

[1] Liu, Bei, Fu, et al. (2018, October 10). Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training.

[2] Xu, T., Zhang, P., Huang, et al. (2019). AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks.

[3] Lin, Maire, Michael, Belongie et al. (2015, February 21). Microsoft COCO: Common Objects in Context. Retrieved from https://arxiv.org/abs/1405.0312

[4] AttnGAN code - https://github.com/taoxugit/AttnGAN

CODE: https://github.com/prajwalppv/Poetry2Image