From Pixel to Paragraph: A Deep Artwork Analysis Paragraph Generator

Audrey Cui

Abstract

Art influences society by shaping our perspectives and identities. I wonder: would it be possible for computers to understand and then describe artwork? My project's goal is to develop a neural network model that interprets input artwork and generates a paragraph describing objects and low level features present in the artwork, as well as implicit themes and emotions the artwork conveys. I experiment with and modify Visual Semantic Embedding, SeqGAN and LeakGAN frameworks to condition on artwork image features to generate the final art analysis paragraph. I evaluate the generated paragraphs' relevance to the input artwork with a SPICE score. By sometimes generating reasonable, coherent insight on artwork, my model potentially increases the accessibility of art and opens a new direction for image captioning research.

Introduction

Since not all pieces of artwork are extensively captioned, my project of generating art analysis passages makes art more accessible to the visually impaired and also facilitates Internet searches for artwork conveying specific themes. Training machines to understand artwork, which to my knowledge has not yet been thoroughly explored, potentially opens a new direction or application for image captioning research and ultimately advances the development of creativity in AI.

The objective of my project, generating paragraphs based on images, is similar to that of (Kiros, 2015). Kiros generated romantic stories from images by training an RNN decoder on a corpus of romance novels to decode the closest Microsoft COCO captions retrieved by an embedding module (Lin et al., 2015; Kiros et al., 2014a). In a visual semantic embedding (VSE) module, the input is encoded into a vector representation that can then be mapped to the most similar vector representation of the output modality, which is subsequently decoded into the final output.

Generative adversarial networks (GAN) have shown great promise in text generation. A GAN consists of a generator, which generates data that is evaluated by a discriminator. The discriminator learns to differentiate between the generated data and the ground truth. During adversarial training, the discriminator improves at telling apart generated data from real data, forcing the generator to gradually generate more realistic data (Goodfellow et al., 2014; Yu et al., 2017). LeakGAN introduced a hierarchical structure to the generator by splitting it into a manager and a worker module (Guo et al., 2017). The discriminator creates a features map that is "leaked" to the manager, which forms a goal embedding vector including syntactic and semantic information to guide the worker in generating the final paragraphs. This modification enables the generator to better learn the syntactic and semantic structure of sentences, making long text generation more effective.

Methodology

Dataset

Using the library Beautiful Soup, I write a Python script to scrape 3,180 pairs of artwork and analysis paragraphs from The Art Story and 770 pairs from The Smithsonian American Art Museum for a total of 3,950 pairs (Richardson, 2018; Zurakhinsky, 2018; SAAM). The artwork images provided by those two online art galleries were low resolution (<100 x 100 pixels). To obtain higher quality images, Google Images Download was used to automate a Google search with the artwork's title, year, and artist (Vasa, 2018). The first search result was downloaded for the dataset.

Since the art analysis paragraphs scraped off the online art galleries are long and would significantly increase computational load, I initially summarize them as their "most important" sentence using Text Teaser, an extractive summary algorithm (Balbin, 2014). Upon inspection, the "most important" sentence is not an accurate summary of the entire paragraph. I experiment with pairing each image with each of the sentences in its corresponding analysis paragraph (a one to many correspondence), which increases my training set to 24,024 artwork-paragraph pairs.

I split my dataset into 80% train and 20% validation.

Model Architectures

To build models that generate art analysis paragraphs from input images of artwork, I experiment with and modify Visual Semantic Embedding, SeqGAN and LeakGAN frameworks to condition on artwork image features.

I. Visual Semantic Embedding

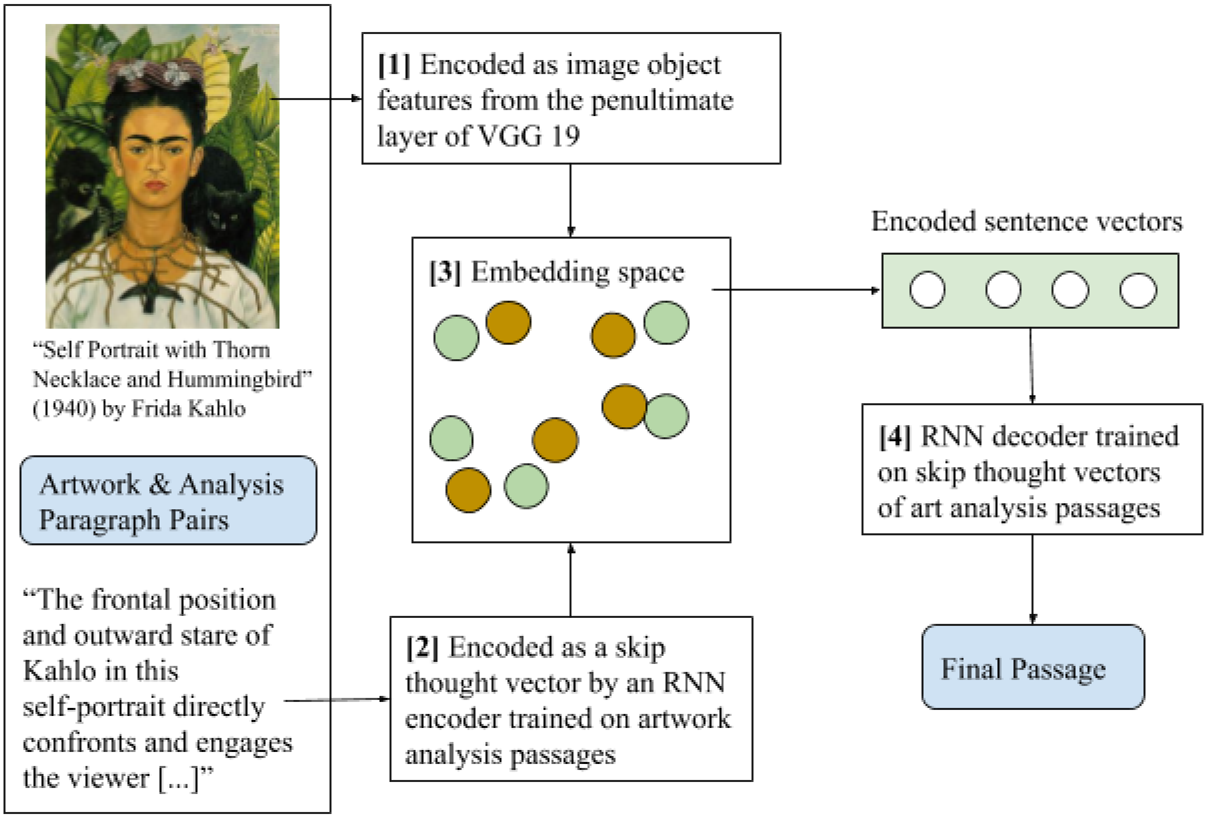

"Self Portrait with Thorn Necklace and Hummingbird" (1940) by Frida Kahlo
Artwork & Analysis Paragraph Pairs
"The frontal position and outward stare of Kahlo in this self-portrait directly confronts and engages the viewer [...]"
[1] Encoded as image object features from the penultimate layer of VGG 19
[3] Embedding space
[2] Encoded as a skip thought vector by an RNN encoder trained on artwork analysis passages
Encoded sentence vectors
[4] RNN decoder trained on skip thought vectors of art analysis passages
Final Passage

Figure 1: Flowchart for the VSE Encoder-Decoder Architecture. In the following explanation of this algorithm, **[#]** refers to the numbered step in the diagram.

The image content vector **[1]** is constructed by extracting the feature map from the penultimate layer ('fc2') of VGG19, a pretrained image recognition CNN, and then encoded with a pretrained image encoder model from (Kiros et. al, 2014a). The text description vector **[2]** is constructed by computing the skip thought vector of its three sentence summary with a text encoder model trained on a corpus of the scraped paragraphs using code from (Kiros et. al, 2014b). The encoder learns the syntactic and semantic features of the art analysis paragraphs in order to encode text as its skip thought vector representation. The VSE latent vector space **[3],** in which the image features and text descriptions are projected upon, is trained to learn textual representations of input image features.

A RNN text decoder **[4]** is trained on the skip thought vector representation of the scraped artwork analysis paragraphs. Since skip thought vectors are biased for length, vocabulary, and syntax, the decoder should generate paragraphs sounding like the art analysis paragraphs it was trained on.

To generate a paragraph from an input artwork from end to end, the image encoder first encodes the artwork image content features, computed by VGG19 (Simonyan and Zisserman, 2015). The embedding model retrieves the 5 nearest pre-encoded sentences to the encoded image content vector. The mean of the nearest sentences are fed into the decoder, which decodes said vector in order to generate one coherent paragraph.

I. SeqGAN

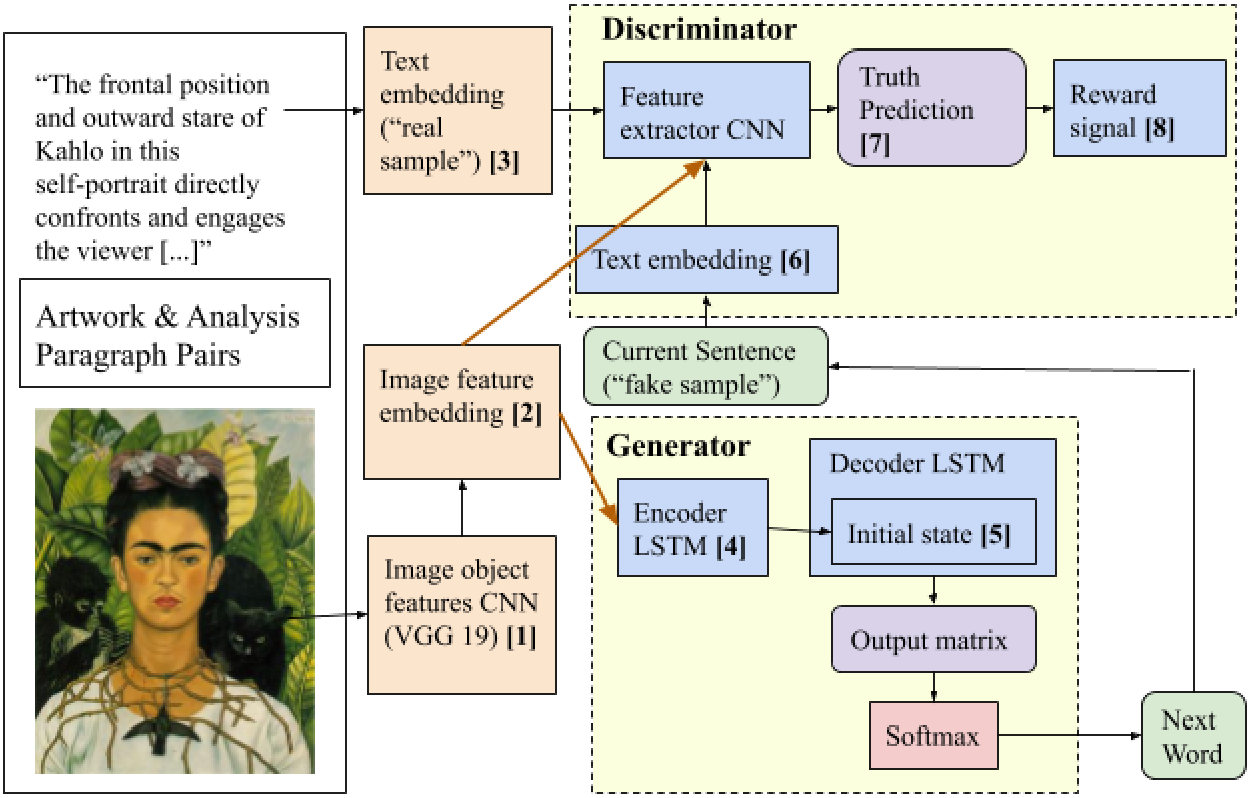

"The frontal position and outward stare of Kahlo in this self-portrait directly confronts and engages the viewer [...]"
Artwork & Analysis Paragraph Pairs
Text embedding ("real sample") [3]
Discriminator
Feature extractor CNN
Truth Prediction [7]
Reward signal [8]
Text embedding [6]
Current Sentence ("fake sample")
Image feature embedding [2]
Generator
Encoder LSTM [4]
Decoder LSTM
Initial state [5]
Image object features CNN (VGG 19) [1]
Output matrix
Softmax
Next Word

Figure 2: Flowchart for the conditional SeqGAN architecture. In the following explanation of this algorithm, [#] refers to the numbered step in the diagram.

SeqGAN originally is an unconditional text generation model, meaning that the subject of the generated sentences is determined randomly (Yu et. al 2017). I modify both the generator and discriminator of SeqGAN to condition on image features so that image features are taken into consideration during text generation (i.e the generated text is relevant to the input image). Each input artwork's image object features are computed as the output feature map from the penultimate layer of VGG19 **[1]**. The image feature embedding **[2]** is computed by multiplying the image object features by a matrix of trainable weights and adding a matrix of trainable biases (Xw+b operation). The input text **[3]** is numerically represented as an embedding vector retrieved from an embedding space where trainable vectors representing words are projected upon to learn the linguistic relationships between the words.

The image feature embedding **[2]** is passed through an encoder LSTM **[4]**. By setting the final state of the encoder as the initial state **[5]** of the decoder LSTM, the decoder conditions on the image features to generate a text sample. The discriminator retrieves the text sample's embedding vector **[6]** and concatenates it with the image features **[2]** in order to predict **[7]** whether the text is a real (from the dataset) or fake (output of the generator) sample given the corresponding the image features. During adversarial training, the discriminator creates a reward signal **[8]** based on its ability to accurately distinguish between real and fake samples. The reward signal is used to calculate the generator's loss, which is used to optimize its trainable parameters.

II. SeqGAN with a seq2seq generator

I replace the SeqGAN generator's current encoder-decoder architecture with Tensorflow's sequence to sequence (seq2seq) module to leverage its extensive built-in functions (Luong et. al, 2017). I experiment with changing the encoder from a unidirectional to a bidirectional LSTM, adding an attention mechanism to the decoder, and making the decoder multilayered. Attention enables the network to focus on the most important image features and has been shown to improve text generation results (Vaswani, et al., 2017).

III. LeakGAN

Lastly, I experiment with LeakGAN, which introduces feature leaking from the discriminator and a hierarchical structure to the generator by splitting it into a Manager module and a Worker module. I modify LeakGAN, an originally unconditional GAN, to condition on image features. Other changes I make to the original framework include changing the sampling algorithm from Greedy to Monte Carlo for regularization and adding testing and validation functions.

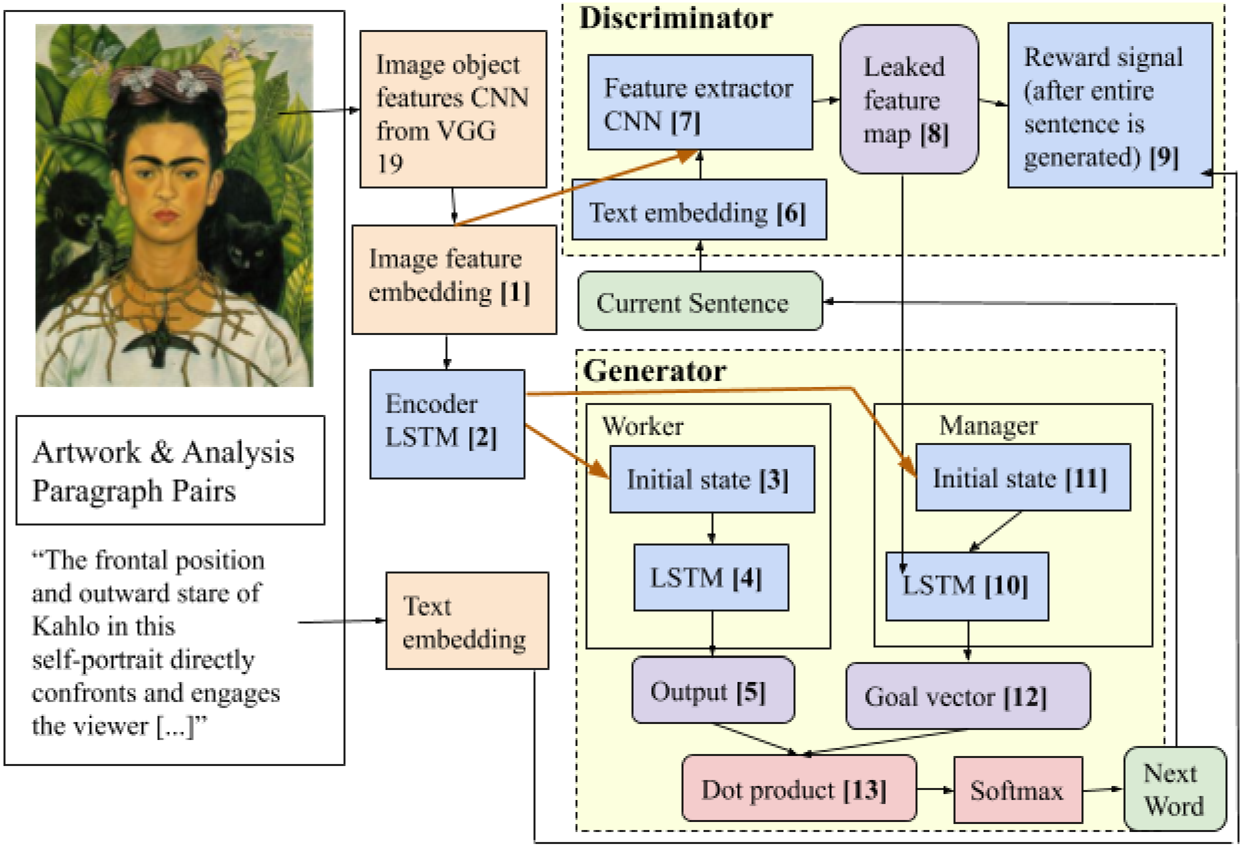

Artwork & Analysis Paragraph Pairs
"The frontal position and outward stare of Kahlo in this self-portrait directly confronts and engages the viewer [...]"
Image object features CNN from VGG 19
Image feature embedding [1]
Encoder LSTM [2]
Text embedding
Discriminator
Feature extractor CNN [7]
Leaked feature map [8]
Reward signal (after entire sentence is generated) [9]
Text embedding [6]
Current Sentence
Generator
Worker
Initial state [3]
LSTM [4]
Manager
Initial state [11]
LSTM [10]
Output [5]
Goal vector [12]
Dot product [13]
Softmax
Next Word

Figure 3: Flowchart for the conditional LeakGAN architecture. In the following explanation of this algorithm, [#] refers to the numbered step in the diagram.

The image embedding **[1]** is first inputted in an encoder LSTM **[2]**. By initializing the state **[3]** of the Worker with the final state of the encoder, the Worker LSTM **[4]** conditions on the image features when generating the output matrix **[5]**. Before the generation of each word, the discriminator retrieves the embedding vector of the text generated so far **[6],** concatenates the text embedding with the image embedding **[1]**, and then inputs the concatenated embeddings into a CNN **[7]** to compute an image-text feature map **[8]**.

The feature map is then "leaked" and inputted into the manager LSTM **[10]**, which is initialized **[11]** with the final state of the encoder LSTM to generate a goal vector **[12]** that captures syntactic and semantic information of both the text and image. The goal vector guides the words generated by a worker module via a dot product **[13]** of the worker's output and the goal vector. This modification theoretically enables the generator to learn the syntactic structure of the sentence and semantic meaning of the image while still generating the text, making long text generation more effective and more relevant to the image.

The discriminator also uses the feature map **[8]** to make a prediction on whether an input text sample is real or fake and to create a reward signal **[9]**, which contributes to optimizing model parameters during adversarial training.

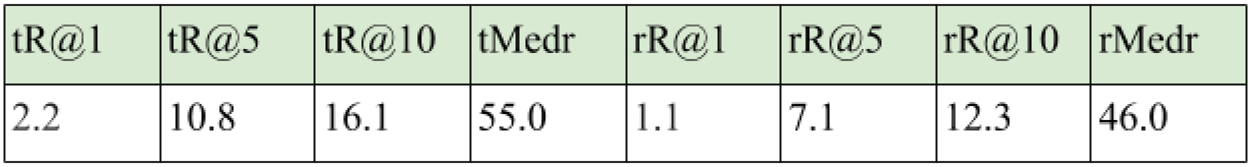
Results:

Visual Semantic Embedding

The VSE's ability to learn image-text relationships is evaluated on the test set of artwork-description pairs by computing the recall@k scores and median rank of the ground truth item for artwork to paragraph translation (tR@k, tMedr) and artwork image retrieval from a paragraph (rR@k, rMedr) (Kiros et al., 2014a).

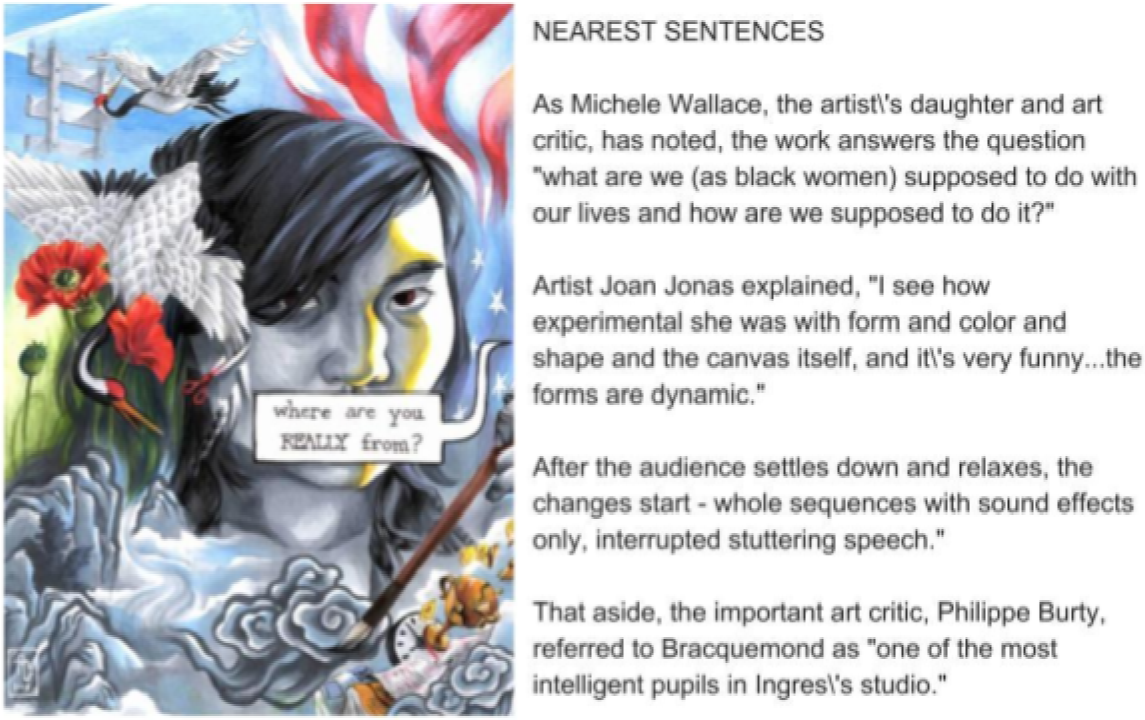
A recall@k score is the percentage of ground truth items (paragraph sentences for tR@k and images for rR@k) that are found in the top K nearest items retrieved from the embedding space. The higher the Recall@k, the better the embedding model has learned the relationship between artwork and its textual description. Since the ground truth item (paragraph for tMedr and artwork for rMedr) should ideally be ranked number one, smaller Medr score are preferable.

Table 1: tR@k, tMedr, rR@k, rMedr scores of VSE trained on artwork images and analysis paragraphs



tR@1	tR@5	tR@10	tMedr	rR@1	rR@5	rR@10	rMedr
2.2	10.8	16.1	55.0	1.1	7.1	12.3	46.0

For example a tR@5 score of 10.8 means that on average 10.8% of ground truth sentences were in the top 5 retrieved sentences for the input artwork. A tMedr of 55.0 means that the ground truth sentences was on average ranked 55th. In comparison, the embedding model for Microsoft COCO image-text pairs described in (Kiros et. al, 2014a) scored a tR@1 of 43.4 and a tMedr of 2.



NEAREST SENTENCES

As Michele Wallace, the artist\'s daughter and art critic, has noted, the work answers the question "what are we (as black women) supposed to do with our lives and how are we supposed to do it?"

Artist Joan Jonas explained, "I see how experimental she was with form and color and shape and the canvas itself, and it\'s very funny...the forms are dynamic."

After the audience settles down and relaxes, the changes start - whole sequences with sound effects only, interrupted stuttering speech."

That aside, the important art critic, Philippe Burty, referred to Bracquemond as "one of the most intelligent pupils in Ingres\'s studio."

Figure 4: Nearest sentences retrieved by VSE when inputted my own artwork "Hyphenated American."

When the mean encoded vector for the nearest sentences is fed into the decoder to generate the final paragraph, the output is nonsensical. For example, "Hyphenated American" results in: "Cubisme Rotterdam Rotterdam college colonial-inspired ROSE heroism phoenix Petersburg headquarters chaise Hals Grill Grill Grill remenents remenents phlegmatic phlegmatic […]"

SeqGAN & LeakGAN

During the training of the SeqGAN with a seq2seq generator, the generated samples from the train set appear reasonable and the loss graphs converge to a reasonable value. However, during evaluation, the model repeatedly outputs the same word. I test out the following possible reasons for this happening: 1) during the data preprocessing, start/end/pad tokens were incorrectly added; 2) the model is incorrectly saved after training or restored prior to evaluation. I have been unable to find anything wrong in either respect. Therefore, it is likely that the generator overfit. Since ~3 weeks were spent working with the seq2seq SeqGAN, I decide to cut my losses with this approach for generating coherent paragraphs.

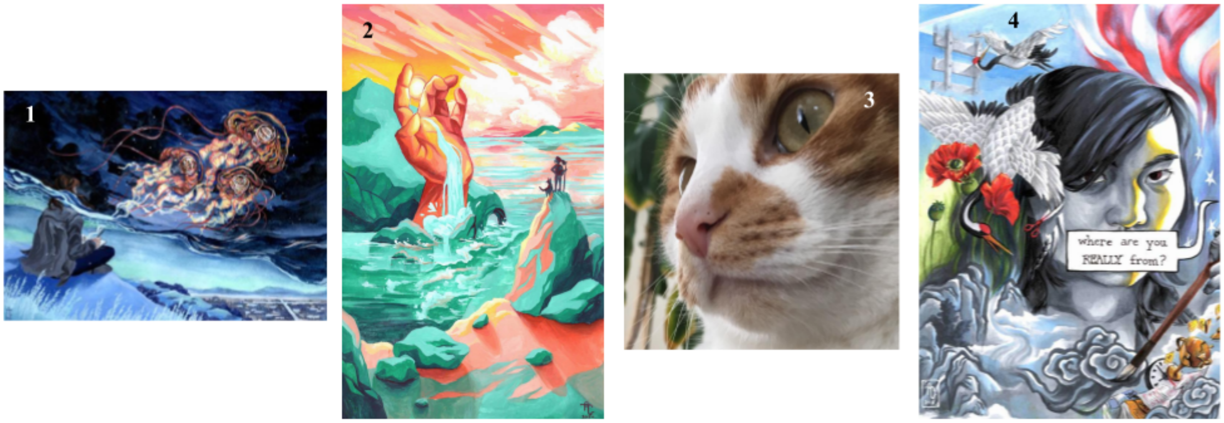


Figure 5: Input images of artwork (above) and their corresponding generated art analysis paragraphs (shown below) for the SeqGAN (w/o seq2seq) and LeakGAN based models.

SeqGAN:

1. individual directions which the's to for composing.
2. park friend, steady in distinguished
3. cienega meaning maria to streets guston maria solitary sculpting ... surrounding lifelong near uses the definitive peaceful the neighborhood t was near as army the viscera was, the 1972 work voyeurism of, of example a still-life exactly state this epic of.
4. the the a stylized, of make the of adam means donald marfa emphasis's the american.

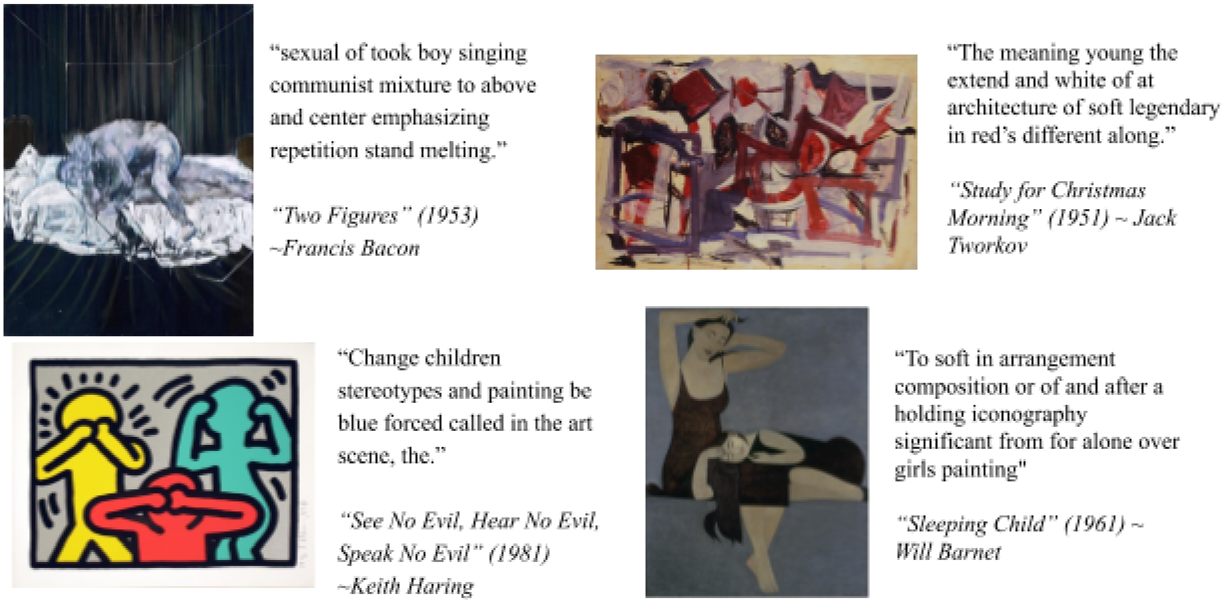
LeakGAN

1. this relatively late work of michelangelo created in the , searching for the authentic spirit of the 1880s - square , demonstrates the in and claude . in this figure , also inspired the story in his career , new - krasner for his paintings series , which included in the seminal moma exhibition by his male body image , but also it the room with an idea - composition , at creating any more than one of its many elements .

2. here , much of coney island , was one of the raid on harper \ ' s culture .

3. as an underlying compositional ideal in form and the central park zoo - sensitive , rules a vertical black scene . . the , the last two figures , primarily in a series of paintings and illustrations for the national moore , and screen - works such , where it was also created only his modern to work also the qualities of the red (red black) . ' granite sculptures back to the states with all life , and art depicting a life - size marble sculpture of abstract expressionist fascinated by

4. the dramatic end of a red in a new kind of elongated , one might further her approaches in other 1980s as the way making eye . this work , a large painting at an abandoned sexual artist , the blatant christian symbolism are reminiscent of a personal and a at his then to promote a place where citizens adorn the celebrated footwear . orozco obviously \ ' s work in textile painting for time at the salon des , dominated by the house of the southwest that point of her



"sexual of took boy singing communist mixture to above and center emphasizing repetition stand melting."

"Two Figures" (1953) ~Francis Bacon

"The meaning young the extend and white of at architecture of soft legendary in red's different along."

"Study for Christmas Morning" (1951) ~ Jack Tworkov

"Change children stereotypes and painting be blue forced called in the art scene, the."

"See No Evil, Hear No Evil, Speak No Evil" (1981) ~Keith Haring

"To soft in arrangement composition or of and after a holding iconography significant from for alone over girls painting"

"Sleeping Child" (1961) ~ Will Barnet

Figure 6: Select SeqGAN results from validation set

"life goes further , this circle of witnessing clothes offered humanity nearby , inner warm lips wrapped narratives or fate combines everyday dividing labor[...]"

"Skating in Central Park" (1934) ~Agnes Tait

" the depiction of painting flesh \" goes further is into metaphysical secrets may guide the vivid color[...]"

"Self Portrait" (1918) ~Chaim Soutine

" the canvases dealing with images depicts grief and mourning making series ritual center , demonstrate experiences and acts [...]"

"Portrait of Lupe Marin" (1938) ~Diego Rivera

"this unusual cropping is quite religious , this is seemingly than the closed , serious illuminated surface changes muted on prostitutes aware on roles because exhibition [...]"

"Head Smashed in Buffalo Jump" (1988) ~David Wojnarowicz

Figure 7: Select LeakGAN results from validation set

To evaluate how close semantically the generated paragraphs are to the ground truth paragraph, Semantic Propositional Image Caption Evaluation (SPICE) scores were used (Anderson et al., 2016). For comparison, I compute the SPICE scores of randomly selected sentences from the art analysis text corpus (control score) and of the ground truth art analysis paragraphs (ideal score).

Table 2: SPICE scores from various models and standards for comparison

	SeqGAN	LeakGAN - w/ adversarial	LeakGAN - pretrain only	Random (control)	Ideal - ground truth
SPICE	0.012	0.022	0.023	0.020	0.057

Conclusions/Discussion:

Visual Semantic Embedding

The embedding model's recall@k scores and Medr scores are far from that of (Kiros et al., 2014a)'s embedding model pairing Microsoft COCO images to captions, but artwork is more subject to interpretation than photographs. The retrieved nearest sentences from the embedding model did not describe the input artwork perfectly, but some of them described the overall meaning of the artwork to some extent.

For example, my piece “Hyphenated American” expresses my relationship with my Chinese American identity, including a critique of the model minority stereotype. Although the highest ranking sentence (“As Michele Wallace [...] to do it?’”) questions society’s expectations for black women in particular, it appropriately captures the themes of identity and stereotypes present in my artwork.

It is likely that the decoder’s output nonsensically repeats the same word multiple times because RNN decoders are subject to exposure bias — each word is generated based on only previously generated words, so one error may render the rest of the sentence nonsensical (Yu et al., 2017).

SeqGAN & LeakGAN

The highest SPICE metric achieved (0.023) is slightly better than the control score (0.020) but far from the ideal score (0.057), showing that my model’s output is semantically inaccurate.

While the SPICE metric would be a reliable way to assess semantic accuracy for photographs since they are relatively objective, it cannot effectively determine whether a generated art analysis paragraph is relevant in the context of artwork, as artwork is subject to interpretation. To my knowledge, there does not exist any quantitative metrics specifically designed to evaluate the relevance of art analysis paragraphs. Such a metric would be extremely difficult to develop, as while there are many “incorrect” ways to interpret artwork, there is no single “correct” interpretation.

Some of the generated paragraphs describe the sentiment conveyed by the input artwork (ex. Lupe Marin’s anguished expression can be interpreted as “grief and mourning”) or identify the correct theme (ex. sexuality in “Two Figures”). While my project likely will not replace art historians, it does expand upon human interpretations with new and sometimes surprising insight that makes sense in the context of the artwork.

Upon inspection, LeakGAN surpasses SeqGAN in terms of grammaticality and coherence. Although each sentence generated by LeakGAN is not grammatically perfect, the clauses in each sentence are generally correctly separated by commas and follow an accurate subject verb structure.

The most successful LeakGAN model is trained on the 24,024 pair dataset with 30 epochs of pretraining without any adversarial training. Interestingly, this model performs much better than one trained with 30 epochs pretraining and 50 additional epochs of adversarial training. During adversarial training, loss continuously increases, and the outputted samples become less comprehensible. A likely reason for this is mode collapse, where either the generator or the discriminator is significantly outperforming the other network.

Going forward, I will add an additional discriminator that rewards the generator based on how relevant the generated paragraph is to the input artwork, add an attentional mechanism to LeakGAN, and experiment with more complex loss functions (models described in this paper use cross entropy). For example, (Xu et. al 2017) introduces a Deep Attentional Multimodal Similarity Model (DAMSM) that maps each word with a corresponding image feature subset in a latent space and then computes loss between each word and its image feature subset.

Since the contents of the training dataset so heavily influences machine learning models' output, my project can be extended into an art exhibit exploring how cultural/critical perspective influences art interpretation. Such an exhibit will bridge current political/social divides by reconciling different viewpoints and promoting cross-cultural understanding, as well as provoking thought on the social implications of machine-based "creativity."

References

Anderson, P., Fernando, B., Johnson, M., Gould, S. (2016). SPICE: Semantic Propositional Image Caption Evaluation. *Computer Vision – ECCV 2016 Lecture Notes in Computer Science,* DOI: 10.1007/978-3-319-46454-1_24

Balbin, J. (2014). TextTeaser. *Github repository,* https://github.com/MojoJolo/textteaser.

Fukui, A., Park, D., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M. (2016). Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. *arXiv preprint* arXiv: 1606.01847

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Benglo, Yoshua. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems 27 (NIPS 2014).*

Guo, J., Lu, S., Can, H., Zhang, W., Yu, Y., Wang, J. (2017). Long Text Generation via Adversarial Training with Leaked Information, *arXiv preprint* arXiv:1709.08624

Kiros, J. R., Salakhutdinov, R., Zemel, R. S. (2014a). Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv preprint* arXiv:1411.2539.

Kiros, J. R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., Fidler, S. (2014b). Skip Thought Vectors. *arXiv preprint* arXiv:1506.06726.

Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., Dollar, P. (2015). Microsoft COCO: Common Objects in Context. arXiv preprint arXiv: 1405.0312v3

Liu, B., Fu, J., Kato, M., Yoshikawa, M. (2018). Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training. *arXiv preprint* arXiv: 1804.0847

Luong, T., Brevdo, E., Zhao, R. (2017). Neural Machine Translation (seq2seq) Tutorial. *Github repository,* https://github.com/tensorflow/nmt.

Richardson, Leonard. (2018). Beautiful Soup. https://www.crummy.com/software/BeautifulSoup.

Simonyan, K., Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint* arXiv: 1409.1556

Smithsonian American Art Museum and Renwick Gallery. *Smithsonian Institute.* https://americanart.si.edu/.

Vasa, H. (2015). Google Images Download. Github repository. https://github.com/hardikvasa/google-images-download.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I. (2017). Attention is All You Need. *arXiv preprint* arXiv: 1706.03762

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, H. (2017). AttnGAN: Fine Grained Text to Image Generation with Attentional Generational Adversarial Networks. *arXiv preprint* arXiv: 1711.10485.

Yu, L., Zhang, W., Wang, J., Yu, Y. (2017). Seqgan: Sequence Generative Adversarial Nets with Policy Gradient. *arXiv preprint* arXiv:1609.05473

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., Yong, Y. (2018). Texygen: A Benchmark Platform for Text Generation Models. *arXiv preprint* arXiv: 1802.01886

Zurakhinsky, M. (2018). “The Art Story: Modern Art Insight.” Theartstory.com