Towards Generating Accurate Art Descriptions: Small WGA Captions Dataset and its Baseline

Hyunwook Seo[1] and Dahee Kim[2]

[1]Senior student, Korean Minjok Leadership Academy
[2]Arts teacher, Korean Minjok Leadership Academy

***Abstract*—The image captioning field of machine learning deals with generating captions for input images. While the field itself prospers due to the existence of massive, reliable datasets such as MSCOCO or Flickr, captioning artworks suffers from an unavoidable lack of datasets available for training. In this paper, we propose a novel Small WGA Captions Dataset of 2891 pictures and caption sets created by gathering descriptions of images contained within the artworks provided by the WGA dataset and individually discerning the usability of each caption. This dataset functions as an artwork captioning dataset able to discern mainly European paintings, sculptures, and architecture, and identify the specifics of these artworks such as what it depicts. Testing this dataset on a few models, we provide the baseline model to best utilize this dataset, which is the VGG16-LSTM. We also compare the captions generated by an InceptionV3 – GRU vision-language model trained on MSCOCO to the Small WGA Captions Dataset.**

***Keywords*—** Vision-language model; Image captioning; Visual Art; Fine-tuning; European Art

I. Introduction

Caption generation is a task that creates captions of given images. Unlike when given real-life pictures, caption generation in artworks are a challenging task due to the different art styles, themes, and materials used for the creation of the artwork being different in every individual painting. Therefore, datasets with an iconographic description able to be generated just from given artworks are a necessity.

In addition, artwork caption generation must be done with datasets that accurately depicts the figures in the photo or themes of the artwork, and does not provide any other noisy captions on the artist, the time period or anything else. Because of this, a "good" dataset is very difficult for art.

In this paper, we talk about a new Small WGA Captions dataset that provides 2891 pictures containing many common scenes and motifs of european art with iconographic captions, generated by applying two steps, algorithmic and manual, to the given captions by the Web Gallery of Art, and make it more fit for machine learning purposes. Testing the data on multiple models, we also give a conclusion on which model was best to utilize this datset, and also provide a comparison between MSCOCO and our Small WGA captions, and conclude that our dataset is much more fit in terms of generating accurate iconographic captions.

II. Related Works

A. Fine Art Image Captioning

The field of caption generation of artworks is a rising field, currently.However, the foundations for the subject has not been laid enough, as the field still suffers from a shortage of available datasets to train their models on. [1] [2] [3] talk about different ways to achieve the goal of captioning fine art. [1] depicts a dataset, the Iconographic captions dataset, aimed towards generating iconclass captions for fine art paintings. Iconographic captions are captions that can accurately describe what is going on in the artwork, and beyond that try and understand the symbolism and the many meanings behind the artworks. The iconclass captions dataset provided by the paper, however, are keywords, and not full sentences. [2] This work is similar to the work done in this paper. A dataset consisting of 4000 photo-captions of 9 iconographies that include *annunciation, adoration, baptism* and more is evaluated using various state of the art models. [3] talks about a style-transfer method of generating the datset. Using MSCOCO and applying an artistic style transfer based on the Wikiart dataset onto it, the paper trains an image captioning model and a feature extractor on these photos, and then evaluates them using images of real artworks. [4] talks about a dataset created using Semart, then captioning the dataset using crowd-workers, creating art-history relevant captions; but the caption itself cannot be generated just using the picture itself, and will take extra effort to utilize correctly.

B. Fine Art and Artificial Intelligence

Fine art in artificial intelligence, not just image captioning, is a topic of interest in many papers in recent years. Due to the difficulty of evaluating art styles of thousands of different artists all in different art movements, nationalities and more categories, it is of no surprise that the overall accuracy of even simple classification tasks are low. [5] evaluates many state-of-the-art fine-tuned CNN model being able to classify genre, artist, style, time period, as well as performing classification based on a specific national artistic context. [6] talks about figure detection in artworks, utilizing various state of the art models to evaluate performances on people detection in fine art. On a people-art dataset, it is mentioned that VGG16 shows the best performance. Our paper today also uses VGG16 to extract features based on the work of this paper. The accuracy of this paper also does not go beyond a 60 percent mark, however. The works of [7] show a dataset that is similar to the one depicted here, but contains a lot more noise such as background knowledge about the author, that is impossible to discern just by looking at the picture.

III. Methodology

This section depicts the creation of the dataset. The foundations of the dataset comes from the descriptions in the Web Gallery of Art website, [8]. Since the main purpose of these descriptions are to give artwork-related information to the viewers, they range from explaining the painting itself to the life of the artist, the symbolic meaning of various details of the painting, historical contexts to the events transpiring in the artwork, and other graphic-unrelated descriptions. Most descriptions contained some these elements rolled together. Since we need information that can be figured out directly from the artwork, these were not ideal. Therefore, through the many steps of processing we applied onto the dataset, the priorities were to make it short and concise, as well as removing as much noise from the descriptions as possible.

This relief came from the castle of Counts Spinelli in Rignano, Tuscany.

(a) This caption tells us where the painting came from, which is not an information that can be found in the picture.

The portico of the Scheveningen Pavilion from 1826 is still indebted to the severe Neoclassical style.

(b) This tells us the name and date of the architecture, which is not an information that can be found in the picture

The painting is signed and dated 'C F Aagaard 1888' at lower left.

(c) This tells us that the picture is signed and dated. Usually this information is tiny, making it hard to glean from the photo. Also, sometimes the signed/dated information isn't shown in the photo but is present somewhere else, such as the back of the painting.

Fig. 1: The removed captions.

A. Creating the Dataset

1) First Filter: Algorithmic Pruning: The first step was to scrape the WGA website, since the captions aren't provided through any downloadable files. The Web Gallery of Art, however, provided users with a massive 52,867 image catalogue with hyperlinks to each artwork, which can then be used to grab captions for each of these artworks that are provided in the webpage for each artwork. Some photos did not contain descriptions, so after the pruning we were left with a total of 45,004 photos with descriptions. Since shorter captions have smaller amounts of tokens and leads to a shorter training time, captions with lengths over 150 words were removed.

2) Second Filter: Manual Pruning: Then came the manual pruning. Out of the 7153 remaining image-captions, we manually combed through each data and removed those deemed unusable. Examples of these unusable captions are included in figure 1 below.

Then, out of the remaining 3784 data, we went through the same process again, but this time adding another shortening process. Much unnecessary information was cut off, making most captions readily discernable from the picture itself, while making the captions shorter and therefore more learning-friendly.

B. Finding the Baseline

1) The Baseline: We selected the VGG16-LSTM vision-language model as the baseline for this dataset. Two vision models (InceptionV3, VGG16) and two NLP models (GRU, LSTM) were considered for the baseline, but the comparison between the success rates of generated captions yielded the best results in the VGG16-LSTM model. Success rates were manually calculated using the same 30 photos and evaluating the generated captions into either successful and unsuccessful, then calculating the successful counts. The sentences was determined to be "successful" if the sentence was understandable and if it contained an accurate description of either the photo or at least one of the subjects. 4 crowd-workers with basic knowledge of european art history evaluated the captions.

TABLE I: GRU

	InceptionV3-GRU	VGG16-GRU
Success	22	26

TABLE II: LSTM

	InceptionV3-LSTM	VGG16-LSTM
Success	27	**33**

TABLE III: Success rate

	Small WGA Captions	MSCOCO
Success	**22**	2

2) Implementation Details: The training was done for 50 epochs for every model, using the Adam optimizer. The CNN model was not trained but was taken from Tensorflow Hub with weights pretrained on Imagenet for feature extraction.

IV. Experiment

This section talks about the experiments conducted with the model. Since models in paintings were what we wanted the models to be able to discern, we opted to use an attention-based vision-language model, close to the one described in [9]. Using attention, we hoped the model would be able to discern models by attributing them to the symbols they usually carry around, such as Mary Magdalene with oil or Jesus with a cross.

A. Finding the Baseline

To find the baseline model for this dataset, we used two CNN models and two RNN models. Transformers were considered, but the actual experiment yielded terrible results due to the small WGA dataset being small, and led to the model overfitting/underfitting itself more times than it could provide an actual description of the picture. All four instances of the model showed okay performance. The VGG16-LSTM model showed the best performance, then InceptionV3-LSTM, VGG16-GRU, and finally InceptionV3-GRU. The success rate is depicted in the table. The VGG16-LSTM showed a better performance than the other models. Some examples are depicted in figure 2 below. However, the VGG-LSTM model seemed to do better in generating acceptable captions across a wide variety of images; the inceptionV3-GRU model seemed to be better in generating great captions for a small amount of pictures. A few captions by the InceptionV3-GRU is depicted in figure 3.

B. Comparisons with Other Datasets

The small WGA captions dataset was compared with captions generated by the same model trained on MSCOCO. Figure 4 shows some captions generated by the MSCOCO model. As shown here, the MSCOCO had difficulty understanding and generating captions on artworks as oppossed to close to real-life pictures.

V. Conclusion & Discussion

In this paper, we proposed a novel dataset, the Small WGA Captions dataset, that provides iconographic captions of 2891 artworks gathered from the WGA dataset that is able to discern what is happening in artworks and was also able to tell who and what scene, even commonly depicted scenes from the bible, is being depicted.

this detail shows a semi-circular face, side of the angel.

(a) Acceptable description of a head of a saint.

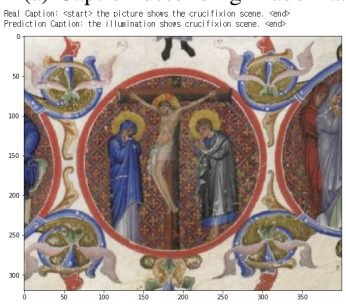
the painting shows the baptist.

(b) Acceptable description of the picture depciting saints including St John the Baptist.

Fig. 2: Captions by VGG16-LSTM.

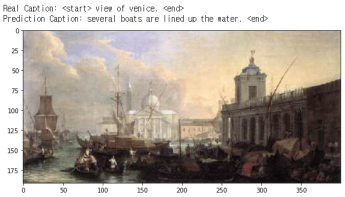
Comparing it to 30000 MSCOCO photo-caption sets, we successfully discerned that our data performs much better in creating art-relevant photos. We also proposed a baseline that the data functions best on, which was the VGG16-LSTM vision-language model. This dataset does have its limitations, however, despite the many layers of processing we went through, the data still remains noisy. Also, since most captions are fragments of the original, some captions may be grammatically incorrect, resulting in generation of linguistaclly unstable captions. More works must be done on stabilizing and clarifying the captions so that the captions from photos will be accurately depicted. Also, out of the leftover captions consisting of lengths longer than 150 will also need to be evaluated and shortened to create even more data on artworks, as around 3k data may be small

(a) Caption describing Madonna.



(b) Caption describing the scene of Crucifixion.

Fig. 3: Captions by InceptionV3-GRU.



(a) Successful caption by MSCOCO model.

(b) Unsuccessful caption by MSCOCO model.

Fig. 4: MSCOCO captions.

for some tasks.

REFERENCES

[1] E. Cetinic, "Iconographic image captioning for artworks," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 502–516.

[2] J. Gupta, P. Madhu, R. Kosti, P. Bell, A. Maier, and V. Christlein, "Towards image caption generation for art historical data," in *Proceedings of the AI Methods for Digital Heritage, Workshop at KI2020 43rd German Conference on Artificial Intelligence, Bamberg, Germany*, 2020, pp. 21–25.

[3] Y. Lu, C. Guo, X. Dai, and F.-Y. Wang, "Data-efficient image captioning of fine art paintings via virtual-real semantic alignment training," *Neurocomputing*, vol. 490, pp. 163–180, 2022.

[4] Z. Bai, Y. Nakashima, and N. Garcia, "Explain me the painting: Multi-topic knowledgeable art description generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5422–5432.

[5] E. Cetinic, T. Lipic, and S. Grgic, "Fine-tuning convolutional neural networks for fine art classification," *Expert Systems with Applications*, vol. 114, pp. 107–118, 2018.

[6] N. Westlake, H. Cai, and P. Hall, "Detecting people in artwork with cnns," in *Computer Vision – ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds. Cham: Springer International Publishing, 2016, pp. 825–841.

[7] B. Saleh and A. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," *arXiv preprint arXiv:1505.00855*, 2015.

[8] E. Kren and D. Marx, "Web gallery of art," 1996, [Online; accessed July 6, 2022]. [Online]. Available: https://www.wga.hu/index.html

[9] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.