Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Special Track on AI, the Arts and Creativity (Exhibits)

Captioning Bosch: A Twitter Bot

Cornelia Ferner
Salzburg University of Applied Sciences
cornelia.ferner@fh-salzburg.ac.at

Abstract

The artworks by Dutch painter Hieronymus Bosch are well known for their incredible wealth of details. The popular BoschBot regularly posts small segments of the digitized paintings on Twitter, thus relieving their density and making them more accessible. CaptioningBoschBot, the Twitter bot presented in this demo, reverses the creative process of the artist: It uses the out-of-context painting segments as input for an encoder-decoder model to generate captions that interpret the painted objects. As the model was only trained on realistic, photographic images, curious interpretations of the otherworldly details can be observed. The generated captions are again posted on Twitter to encourage discussions about Bosch's masterpieces and the AI technology in general.

1 Introduction

Twitter is a social media platform where users can post short text messages and/or embedded media content. Users can interact by following, replying, retweeting or liking each other's content. Additionally, Twitter provides an API that allows to automate these tasks. This led to the appearance of so-called bot profiles which post content automatically [Chu *et al.*, 2012]. In 2017, it was estimated that between 9% and 15% of Twitter accounts were bots [Varol *et al.*, 2017].

Some Twitter bots are especially dedicated to the visual arts: A number of bots uses Twitter to post artworks out of different museums' collections [Fitzpatrick, 2017], while others are dedicated to a single artist and post random segments of their original artwork [Benson, 2019]. A more generative approach is taken by BotMondrian[1] and TheTinyGallery[2] [Benson, 2019]. BotMondrian generates compositions in Mondrian's style with the generation process influenced by the weather in Minnesota. TinyGallery presents emojis as exhibition objects in a perceived gallery.

With currently over 90,000 followers, BoschBot[3] is one of the most popular Twitter art bots [McMullan, 2018]. Its purpose is to hourly post a small segment from Hieronymus Bosch's triptych "The Garden of Earthly Delights" (created 1490-1510)[4]. The posted image segments allow for more focused insights on the wealth of details present in Bosch's paintings which manifests in an active discussion of followers by commenting, captioning and retweeting the posts.

But why not have another bot enter the discussion and have an objective AI interpret the images to fuel the discourse? This approach is diametrically opposed to a human interpretation as an AI lacks the cultural background and natural affinity for arts. It delivers sterile descriptions that often convey unintended absurdity. In technical terms, this task is known as image captioning and is usually carried out with newspaper or social media images with photo-realistic scenes and objective descriptions. State-of-the-art models for image captioning such as the M2 transformer [Cornia *et al.*, 2020] rely on visual attention [Xu *et al.*, 2015] in an encoder-decoder network, i.e. encoded image features are used as attention source for the decoder that generates the image descriptions.

The aim of the presented CaptioningBoschBot is to generate captions for the posted image segments of BoschBot using a pretrained model that has never seen artwork before. It is also implemented as Twitter bot to encourage discussions and rise interest and curiosity for the art(ist), the AI technology and its limitations and potential.

2 How Does It Work?

CaptioningBoschBot[5] went online in January, 2022. After a short testing period, it now posts a few Tweets per week. The bot is live and can be followed on Twitter:
https://twitter.com/CaptionBoschBot.

It consists of three main steps as illustrated in Figure 1: (1) Accessing a Tweet from BoschBot and extracting the image segment, (2) generating a caption for the given input image, and (3) constructing and posting a Retweet containing the caption and the original post. The core of CaptioningBoschBot is a pretrained vision-to-text transformer model. The bot is implemented in Python and uses transformers from

[1] https://twitter.com/BotMondrian
[2] https://twitter.com/thetinygallery
[3] https://twitter.com/boschbot

[4] The original artwork is on display at the Museo Del Prado in Madrid, Spain. Another of Bosch's masterpieces, the triptych "The Last Judgement" is on display at the Academy of Fine Arts in Vienna, Austria.

[5] https://github.com/cornelia-lm/captioning-bosch.git

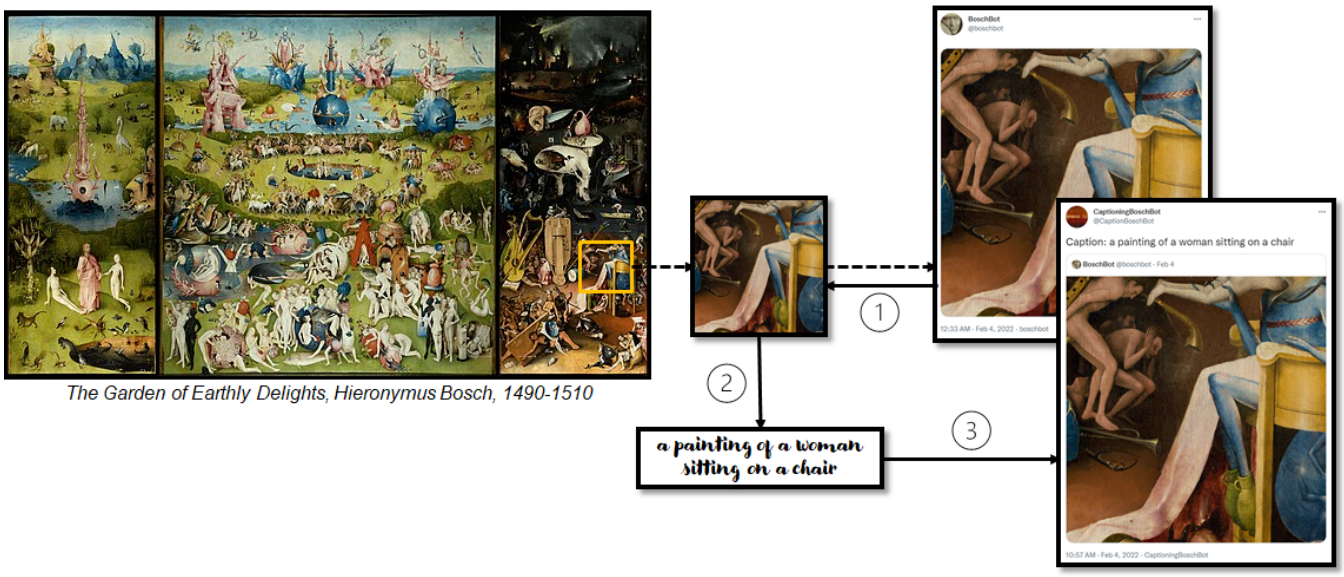


Figure 1: From the (digitized) Bosch painting to the generated caption for an image detail on Twitter. Dashed arrows indicate the behaviour of the existing BoschBot. Solid arrows indicate the three steps performed by the presented CaptioningBoschBot: (1) Accessing Tweets, (2) Generating Captions and (3) Posting Retweets.

huggingface[6] for the image captioning model and tweepy[7] as library to access the Twitter API.

2.1 Accessing Tweets

Retrieving Tweets over the Twitter API[8] requires an authentication step with credentials from a registered Twitter developer account. Given these credentials, an API object can be instantiated that handles the access to the Twitter stream.

Next, Tweets from a user are retrieved by specifying the ID or screen name which is done to access the latest Tweets from BoschBot. The query can be parameterized to exclude replies and retweets from the resulting list of Tweets. The final result list is parsed for posts containing media to extract the corresponding image URL. From this URL, the image is loaded and preprocessed.

2.2 Generating Captions

The vision-to-text transformer model for image captioning consists of an encoder that transforms a given image into a latent representation and a language model as decoder for transforming the latent representation into a textual description. The already pretrained model[9] that was used for CaptioningBoschBot consists of a Vision Transformer (ViT) [Dosovitskiy *et al.*, 2021] neural network as encoder and a GPT-2 [Radford *et al.*, 2019] neural network as decoder. The underlying ViT was pretrained on ImageNet [Deng *et al.*, 2009] and requires images to be rescaled or resized to a resolution of 224x224.

The encoder-decoder model pipeline for image captioning was pretrained on the MS COCO captions [Chen *et al.*, 2015] dataset that consists of over 330,000 images aligned with captions. Different to the image segments from Bosch's painting, images from the MS COCO captions dataset show real-world scenes with the object(s) of interest mostly centered in the picture or prominently in the foreground and comprises target captions such as "*a large bus sitting next to a very tall building*". The encoder-decoder model generates captions for a given input image based on a beam search with 4 beams and a maximum sequence length of 16 tokens over a vocabulary of size 50,257.

2.3 Posting Retweets

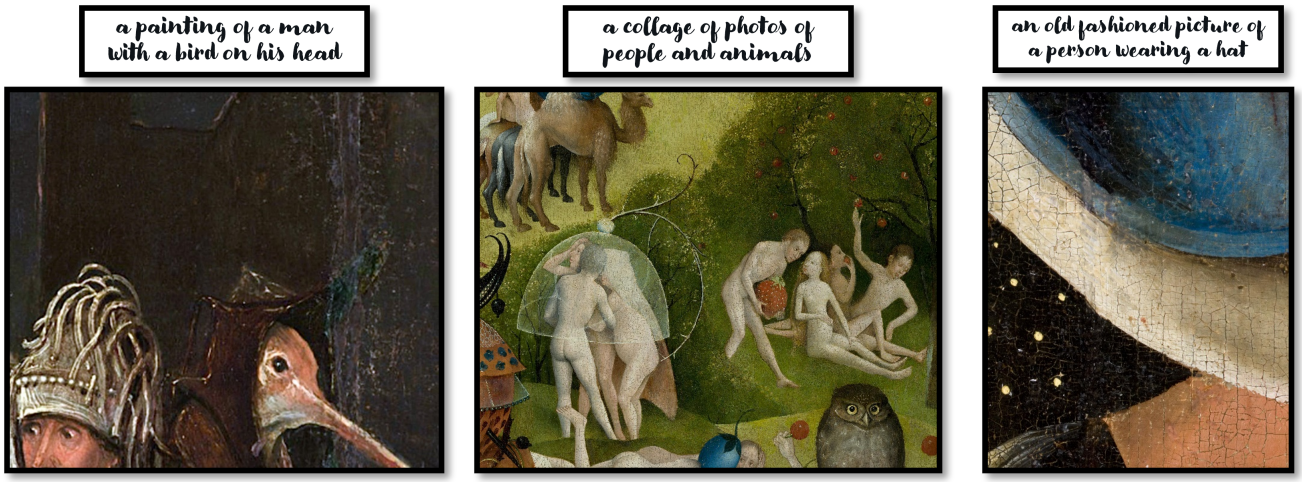
The caption generated by the encoder-decoder model is passed to the API object to update the status, i.e. post a Tweet. In order to retweet the original post, its Tweet ID has to be extracted and passed as attachment URL.

3 Results

Results are best monitored on CaptioningBoschBot's Twitter profile, but this section highlights some examples and discusses pitfalls and subtleties. There are different aspects that influence the quality and appearance of the generated captions. The set of training images has the most influence on the captions that can be generated. As mentioned above, images in the MS COCO dataset show a single scene only which is different from the Bosch image segments. At inference time, the model's expressiveness can be influenced to some extent by adapting the maximum token length of the generated sequence and the number of beams used for beam search.

Besides a number of accurate image descriptions (see Figure 2a, for instance), two main issues can be identified:

[6]https://huggingface.co/docs/transformers/index
[7]https://www.tweepy.org/
[8]https://developer.twitter.com/en/docs/twitter-api
[9]https://huggingface.co/nlpconnect/vit-gpt2-image-captioning



(a) (b) (c)

Figure 2: Image and caption examples taken from the Twitter profile of CaptioningBoschBot.

Fixation on Cracking. The digitized reproduction of Bosch's triptych captures the craquelure of the painting well. This leads to many captions starting with "*a painting of*" (see Figures 1 and 2a). In fact, in some cases, the captioning model even overemphasizes on this aspect, generating the description "*a painting of a painting of a painting of a cow*" posted on 7 February 2022.

Lost Focus. Given the already discussed objective nature of the image-caption pairs in the training data, the model doesn't accurately capture objects in ill-centered image segments or additional objects that are cut off at the borders or appear in the background. This can be seen in Figure 1 where only the human-like figure in the foreground is considered in the caption "*a painting of a woman sitting on a chair*". Another example from 3 February 2022 with the caption "*a cat sitting on top of a wall next to a plant*" ignores the animals (e.g. beetle, snake) in the lower right corner. Besides, the center object is much smaller than in training images, which could be an explanation for the confusion of a rabbit with a cat.

The example in Figure 2b, however, depicts a crowded scene where the model is not able to focus on any single aspect any more. Instead of putting people and animals in context ("next to", "behind"), the caption summarizes the content as "*a collage of photos of people and animals*". With the maximum sequence length and number of beams used for generating the caption increased, the model outputs "*a painting of a woman sitting on top of a lush green field*", exhibiting the same behavior again.

For image segments with a high zoom factor as in Figure 2c, the generated captions often exhibit the same level of detail resulting in realistic descriptions of abstract forms, as in "*an old fashioned picture of a person wearing a hat*". A similar example was posted on 28 April 2022 with the caption "*a painting of a woman sitting on a bench*".

Comprehension Gap. CaptioningBoschBot is only able to recognize familiar objects and scenes in the image segments, i.e. things it has encountered during training. This leads to curious misinterpretations that ignore the (historical) context of the painting: The captions include terms such as "surfboard" ("*a painting of a woman holding a surfboard*" posted on 1 May 2022), "graffiti" ("*an old photo of a building with graffiti on it*" posted on 25 March 2022), "frisbee" ("*a man and a woman are playing frisbee on the grass*" posted on 28 February 2022) or "band" ("*a man playing in a band playing drums while others look on*" posted on 1 May 2022). Moreover, it completely ignores the concept of nudity and only relates the presence of nude skin to bathing and beach scenes (as in "*a painting of a man in a bathing suit*" posted on 1 April 2022).

4 Discussion

CaptioningBoschBot showcases how an AI system that has only ever seen images of clean real-world scenes interprets artistic creativity. Given its training data and setting, it lacks the means to recognize art and the vocabulary to describe it. It illustrates that creativity is not inherent in an AI model per default but also highlights how such lack of imagination manifests in intriguing interpretations. On the one hand, it often reveals the discrepancy between what is actually depicted and what is projected into an image by human interpretations that consider the painting's history and context. On the other hand, it stimulates human imagination through trying to find an explanation for what characteristics of the image served to trigger the generated caption.

Some aspects of the generated captions could be easily improved. Captions that contain "loops" in the text, e.g. "*a painting of a painting*" could be omitted. Training the encoder-decoder model on more diverse image datasets could minimize the tendency to focus on the object in the image center only. Another workaround for this problem would be to generate separate captions for image patches.

Acknowledgments

This project was conducted at the Applied Data Science Lab of the Salzburg University of Applied Sciences under its doctoral support programme and is partially funded by the Science and Innovation Strategy Salzburg (WISS 2025) project "IDA-Lab Salzburg", grant number 20102-F1901166-KZP.

References

[Benson, 2019] Louise Benson. Pointless Profundity? How Twitter Art Bots Are Shaking Up Visual Culture. Elephant, 2019. https://elephant.art/pointless-profundity-twitter-art-bots-shaking-visual-culture/, Last accessed on 2022-04-27.

[Chen *et al.*, 2015] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*, 2015.

[Chu *et al.*, 2012] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6):811–824, 2012.

[Cornia *et al.*, 2020] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020.

[Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*. IEEE, 2009.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*. ICLR, 2021.

[Fitzpatrick, 2017] L. Kelly Fitzpatrick. Anatomy of a Museum Twitter Bot. Medium, 2017. https://medium.com/berkman-klein-center/anatomy-of-a-museum-twitter-bot-2311d81de243, Last accessed on 2022-04-27.

[McMullan, 2018] Thomas McMullan. This Bonkers Bot is the Only Twitter Account Worth Following Right Now. Wired, 2018. https://www.wired.co.uk/article/bosch-twitter-bot-digital-art, Last accessed on 2022-04-27.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9, 2019.

[Varol *et al.*, 2017] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. Online Human-Bot Interactions: Detection, Estimation, and Characterization. In *Eleventh International AAAI Conference on Web and Social Media*, pages 280–289. AAAI, 2017.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2048–2057. PMLR, 2015.