PriMera Scientific Engineering
Volume 3 Issue 1 July 2023
ISSN: 2834-2550

PriMera Scientific Publications

Image Æsthetic Score Prediction using Image Captioning

Type: Review Article
Received: May 02, 2023
Published: June 29, 2023

Citation:
Aakash Pandit., et al. "Image Aesthetic Score Prediction using Image Captioning". PriMera Scientific Engineering 3.1 (2023): 61-68.

Aakash Pandit*, Animesh, Bhuvesh Kumar Gautam and Ritu Agarwal

Information Technology, Delhi Technological University, Delhi, India

***Corresponding Author:** Aakash Pandit, Information Technology, Delhi Technological University, Delhi, India.

Abstract

Different kinds of images induce different kinds of stimuli in humans. Certain types of images tend to activate specific parts of our brain. Professional photographers use methods and techniques like rule of thirds, exposure, etc, to click an appealing photograph. Image aesthetic is a partially subjective topic as there are some aspects of the image that are more appealing to the person's eye than the others, and the paper presents a novel technique to generate a typical score of the quality of an image by using the image captioning technique. The model for Image Captioning model has been trained using Convolutional Neural Network, Long Short Term Memory, Recurrent Neural Networks and Attention Layer. After the Image caption generation we made, a textual analysis is done using RNN- LSTM, embedding layer, LSTM layer, and sigmoid function and then the score of the image is predicted for its aesthetic quality.

Keywords: Image Aesthetic; Convolutional Neural Network; Long Short Term Memory; Recurrent Neural Networks; Attention Layer; Embedding Layer; Image Captioning

Introduction

The aesthetic quality of an art piece, like a photograph, concludes in psychological reactions in people. There are various angles that empower a high aesthetic nature of a photograph. As opposed to individuals liking the genuine characteristics of a photograph, they often like different abstract characteristics of a photograph also, for example, regardless of whether the composition is balanced and how the color is dispersed.

Low level image elements like sharpness, clarity and saliency that intently identify with human comprehension are named as picture image aesthetic features. It is difficult to plan a metric to evaluate these aesthetic characteristics, however, different investigations show that algorithms can be planned and tuned to anticipate metrics identified with color and composition.

In photography, it typically implies that a picture appeals to the eye. There is something about its composition, subject, and shading that makes us notice it. Similar to beauty, aesthetics are not easy to be characterized in basic words. Everything relies upon the viewer's photographic knowledge, experiences, and preferences. Fig. 1. shows some aesthetic photographs which are appealing to the eyes, whereas Fig. 2. shows some unaesthetic photographs.

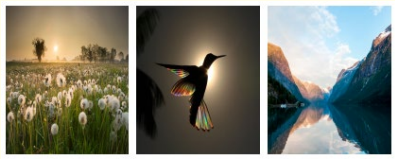

Figure 1: Aesthetic Photographs.

Figure 2: Unaesthetic Photographs.

Photographic artists or specialists, purposefully join such properties to frame a bunch of visual guidelines, to catch high-fidelity and appealing pictures to satisfy the audience or want the emotional impact for a huge audience.

There are a few aesthetic principles of photography:

Composition

- Leading Lines - It is the procedure of outlining lines inside a picture to attract the viewer's eye towards that specific point. Since leading lines catch the viewer's concentration, it attracts them to see the magnificence of image.
- Rule of Thirds - It is the method involved with dividing a picture into one-thirds, utilizing two vertical and two horizontal lines. This imaginary framework yields nine sections with four points of intersection. Utilizing the rule of thirds makes adjusted creations that are normally satisfying to the eye.
- Symmetry - It is the conventional equilibrium of weight inside a picture. Similar to leading line and rule of thirds, it is likewise satisfying to the human eye since symmetry handles the whole image in a balanced way.

Lighting

- Using available lighting - Trying different things with light doesn't generally mean utilizing artificial light sources and staging them to make delightful lighting styles. This implies utilizing whatever light is accessible, such that it catches and turns into a textural part of the picture.
- Using shadows creatively - An incredible method for utilizing light to make fascinating pieces is by making shadows. Shadows can make intriguing surfaces and subjects that become a piece of the photograph's composition when it is utilized imaginatively.
- Soft or Harsh Light - Figuring out which nature of light we like more, will likewise assist with developing one's own aesthetic feature. Soft light accomplished falsely with diffusers or normally at Golden Hour can contribute apparently to the pictures whereas harsh light can assist with making more emotional pictures.

Color schemes

- Use colors with intention - Both" color photography" and "black and white photography", have their advantages and disadvantages. It is important to utilize the color intentionally. Purposeful utilization of explicit coloring plans is an incredible method for making an image that is aesthetically pleasing.
- Color Theory - Color theory means that the different combinations of colors can have different effects on the viewer, psychologically. So for shooting aesthetical images, the photographer needs to understand the color theory. Stated in section IA, IB and IC are only a few of the many parameters which are there to check the quality of image aesthetic.

Related Work

In this section, we discuss some of the state-of-the-art techniques from the literature such as: 1) Visual Attention mechanism 2) Multi-modal analysis 3) Computational approach for Image Aesthetics score Prediction.

Visual Attention mechanism

Attention mechanisms started from the examination of the vision of humans. In cognitive science, of the bottlenecks of data processing in the brain, just a small portion of all apparent data is seen by humans. Roused by this visual attention mechanism, scientists have attempted to track down the model of visual selective attention to reproduce the visual perception process of humans, and model the distribution of attention of humans while observing photos and videos. Taking into consideration its widespread applications, a great number models of visual attention has been put forward in the literature, as in [1, 2]. Only recently, the potential success of incorporating visual attention mechanism in image aesthetic prediction has been given more attention by researchers. These methods used recently under- mine the process of human perception though they achieve great performance by using saliency to sampling, although these methods achieve impressive performance. While we are observing the images, we give attention to various portions of visual space sequentially in order to gain any important information, and to make an inter- nal representation of the scene, we join the portions of data from multiple fixations. The trial results exhibit that it can accomplish better performance than customary techniques.

Multi-Modal Analysis

With the fast development of multimedia information, various types of modalities that depict same content can be effortlessly acquired, like sound, pictures, and text. These modalities are interrelated to each other and can give corresponding data to one another. Single-modal methodologies have been outweighed by Multi-modal methodologies in many works, for example, He. et al. in [3] and Bai et al. in [4], for Image classification, the authors have joined language and vision to boost up the performance. Multimodal methods have not been used as much in image aesthetic prediction despite having success in many tasks, with a few exceptions like [5, 6]. The only difficult work in multi-modal methodology is to combine the multi-modal information optimally. Test results exhibit it can accomplish critical improvement on the image aesthetic score prediction tasks.

Computational approach for Image Aesthetic score Prediction

The purpose of Computational image aesthetics is to design algorithms to perform aesthetic predictions, in a similar way as human beings. In the past two decades, there has been a lot of development in computational image aesthetics prediction, the credit goes to deep learning algorithms and huge annotated datasets. It has influenced image retrieval, image enhancement and recommendations in many ways. Many researchers have attempted to tackle the problem of image aesthetic prediction [7-13]. The earlier approaches were based on handcrafted features. In the features here, colorfulness, hue, saturation, and exposure of light [7, 8] are global features that can be used for all types of photos. Local features such as Composition, clarity contrast, Dark channel and geometry should be planned as per the assortment of photograph content [14, 7]. Deep learning networks are often being used nowadays by researchers for image aesthetic quality assessment. Few fundamental works in present day computational aesthetics prediction were put forward by Datta et al. [8] and Ke et al. [15] in 2006. Kucer et al. [12] with the deep learned features consolidate the hand-designed features to lift up the performance. Kao et al. [11], for the extra supervision, used tags on images to predict the image aesthetics and put forward a deep multi-task network. The methods pose attention on encoding the composition (global) and finer details(local). The local view is addressed by cropped patch and global view is addressed with a distorted image.

Proposed Methodology

To check the aesthetic quality of an image, we have used deep learning models and have obtained a decent quality checker.

CNN

Image processing and identification which mainly deals with the pixel information, is done by convolutional neural networks (CNN), which are a type of Artificial Neural Network (ANN). A CNN uses a framework that decreases the processing necessities like a multilayer perceptron. The layers of a CNN consist of 3 layers which are input, output, and a hidden layer. They help in incorporating different pooling layers, convolutional layers, normalization, and fully connected layers.

Caption Generator

To increase the quality of our prediction, we check the quality of images using the content inside the image, and for obtaining that, we have extracted an image caption for every image and then from that generated caption we have predicted the aesthetic quality of the image. For image caption generator we have used [16] which follows the model structure as stated below.

- Encoder: CNN - The model takes a raw picture and creates an encoded subtitle. We have utilized a CNN for extracting the feature vectors. We have taken features from a lower convolutional layer which permits the decoder to specifically focus on specific pieces of a picture by choosing a part of all the feature vectors.
- Decoder: LSTM - Long Short-Term Memory (LSTM) networks are a kind of Recurrent Neural Networks (RNN) used in prediction problems of sequencing for learning order dependence. We utilize an LSTM network that creates a caption by producing a single word at each time step based on a context vector, the recently generated words, and the previously hidden states. Fig. 3. shows an LSTM cell.
- RNN - RNNs are an incredible and vigorous sort of neural networks, and consists of a very promising algorithm because they have internal memory. Due to their internal memory, RNN's can recollect significant things about the information they obtain, which permits them to have a precise prediction of what's coming in the future.
- Deterministic Soft Attention Layer - Attention is a procedure that imitates intellectual attention. The impact upgrades the significant pieces of the input information and fades the rest. Learning stochastic attention requires testing the attention area st each time, rather we can make the assumption for the context vector zt and plan a model which is deterministically attentive.

$$E_{p(s_t\|a)}[\hat{z}_t] = \sum_{i=1}^{L} \alpha_{t,i}\mathbf{a}_i \qquad (1)$$

The entire model is smooth and differentiable under deterministic attention, so using backpropagation to learn end-to-end is trivial. So, the model is prepared end-to-end by limiting the negative log-probability:

$$L_d = -log(P(\mathbf{y}\|\mathbf{x})) + \lambda \sum_{i}^{L}(1 - \sum_{i}^{C} \alpha_{ti}) \qquad (2)$$

Caption Analysis

We analyse the caption of image using RNN-LSTM model for the generating the score of the image. We use three layers of network as shown in Fig. 4.:

- Embedding Layer - We have too many words in our vocabulary, so it is computationally very expensive to do one-hot encoding of these many classes, so we add an embedding layer as the initial layer. We utilize this layer as a lookup table instead of one-hot encoding.
- LSTM Layer - Long Short Term Memory layer takes care of results obtained from the previous layers as it stores two types of memory, long term memory and short term memory within itself. It has four gates, learn gate, forgot gate, remember gate, use gate; An LSTM cell is shown in Fig. 3. The Input data goes in this layer, and the layer with the help of previous and current information, predicts the result.
- Sigmoid function - A single sigmoid output gives a probability between 0 and 1, which we use in the Multi Cross-Entropy loss as the loss function.

$$S(x) = \frac{1}{1 + e^{-x}} \qquad (3)$$

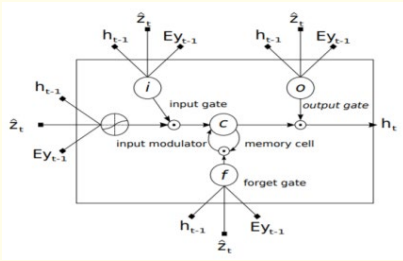

Figure 3: An LSTM cell [16].

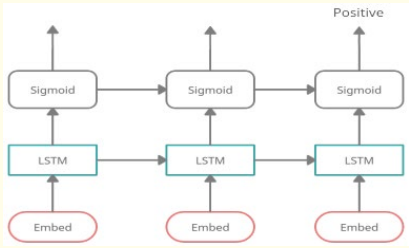

Figure 4: Layers of our model.

Experimental Analysis

Training dataset

We have used Aesthetics and Attribute database (AADB) given by Kong et al. [17]. It consists of 8,000 images with 8 attributes (vivid color, rule of thirds, object emphasis, color harmony, good lighting, shallow depth of field, interesting content, balancing element) having overall scores of aesthetic quality and also having binary labels of them which have been rated by five Amazon Mechanical Turk (AMT) workers. The AADB and AVA give a bigger scope, more extensive score circulation, more extravagant semantic and style characteristic comments than Photo.Net. AADB contains a significantly more adjusted distribution of visual symbolism of genuine scenes down- loaded from Flickr. However, the number of images are less compared to other datsets and the attribute tag is binary in the AADB dataset (high or low aesthetic). Table 1. shows the comparison of AADB dataset with AVA and Photo.Net dataset.

This dataset contains rater identities and informative attributes along with the score distributions of images. These explanations empower us to concentrate on the utilization of people's rating on the quality of image aesthetic for training our model and investigate how the trained model performs contrasted with individual human raters.

	AADB	AVA	Photo.Net
Rater's ID	Yes	No	No
Real photos (All of them)	Yes	No	Yes
Attribute Label	Yes	Yes	No
Score Distribution	Yes	Yes	Yes

Table 1: Comparision of AADB dataset with AVA and Photo.Net dataset.

Score Prediction with Caption Generation

- Training Dataset for Caption Generation - Here we have used the MS- COCO dataset to train our Caption Generation model. The MS COCO dataset is a large-scale captioning, object detection and segmentation dataset published by Microsoft [18]. It is a popular dataset among the computer vision pioneers in the machine learning community. As the main task of computer vision is to understand the scene and conclude what's going on in a scene and then coming up with a semantic description. The state of the art MS-COCO data is suitable for this task of image captioning. There are more than 82,000 images in this dataset, with each image having no less than five captions with different annotations. The dataset has about 91 categories of objects. COCO has less categories and more instances per category. We are using 20,000 images with 30,000 captions where each image is having multiple captions from the MS-COCO dataset. We create 80-20 split randomly of training and validation sets. So there are 16,000 images for training and 4,000 for testing. As MS-COCO dataset is very large and reliable, we used it to train our model.
- Caption Generation - Images are resized to 299px*299px and passed to Inception V3 model which classifies the images and features are extracted using the last convolutional layer of the Inception V3 model. For processing the captions, we first tokenize

them, which gives us the vocabulary of all the words which are unique in the dataset and then we limit the size of the vocabulary upto 5,000 words only to save the memory and then replace other words with "UNK"(unknown) token. After that we create mappings of index-to-word and word-to-index and then pad zeroes at the end of each sequence to make all the sequences of same length, where the length would be of the longest sequence available. Attention layer allows our model to focus on the parts of the image which are of prominence in generating the caption and gives better results for our caption generator model. After extracting the vector from lower convolutional layer of the Inception V3 model, we pass this vector through CNN Encoder which has a single fully connected layer. The output of the encoder, hidden state (initialized to zero), and start token are passed to the decoder. Decoder returns its hidden state and the predictions of caption. Then the hidden state is returned back and the predictions are utilized for loss calculations. The RNN is then used to analyze the image to predict the next word. Fig. 5. shows the caption generation model and Fig. 6. shows an example of a caption generated by our model.

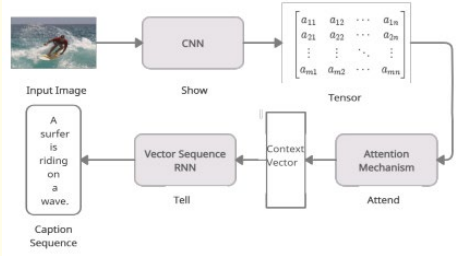

Figure 5: Caption Generation Model.

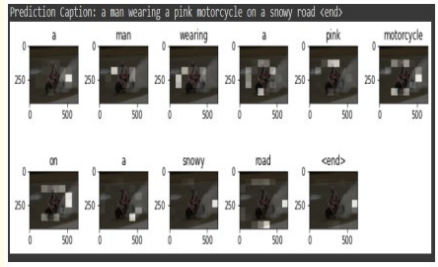

Figure 6: Image Caption Generated through attention mechanism.

- Generating aesthetic score from captions - The file of captions generated by our caption generator model is used as an input for our next model which generates an aesthetic score from the captions. Steps for doing the same are:
  1. The first step of this model encodes each word with a number using embedding layer as the first layer. This step converts our data suitable for feeding into a neural network. We remove all punctuations and emojis and split the text into the words in a single line.
  2. We utilize the spacy and regular expression to take out only the textual data for the process of lemmatization and tokenization of each word of our caption.
  3. We count all the distinct words in our vocabulary, and sort the number according to their occurrence and map them starting from one.

Our labels are also integers, from one to five. So, we have used one-hot encoding here. We convert the captions to the same length for accurate results and less computation. We fix a specific input length for our captions, and then we convert all other captions to that specific length only. To do this task, we followed two steps:

(a) We delete extremely small and extremely large captions.
(b) For the small captions, we pad zeroes at the end, and for large captions, we truncate it from the end.

For the above step, we choose the fixed length to be twenty in this project. So, whenever a caption size is less than twenty, we pad zero at the end, and if its size is greater than twenty, we truncate it from the end. Now our data is of fixed length, so now we divide it into the testing, training, and validation data. While training our model, for changing the attributes/weights of RNN we use Adam Optimizer, as it is robust, and also it takes momentum into consideration of the updated gradients. We use the LSTM layer for predicting the result with the help of previous information and the current information, and use a sigmoid function for entropy loss calculation. Fig. 7. shows the flow of our textual analysis model which generates a score from the captions.

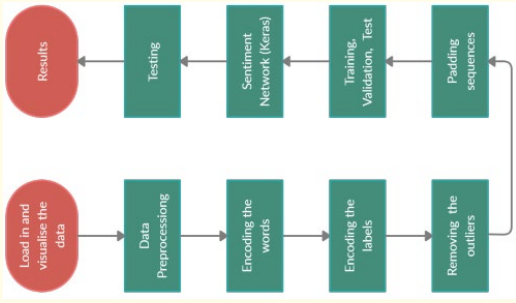

Figure 7: Flowchart for Textual Analysis model.

General Model

So, in our general model, the image is taken as an input for the Caption Generator Model. This model generates a caption from the image, which gives a description of the image. Then the generated caption of the image is provided as an input for the Text Analysis model, which then predicts the Aesthetic quality of the image by generating a score on the scale of 1-5. Fig. 8. shows the flowchart of our general model.

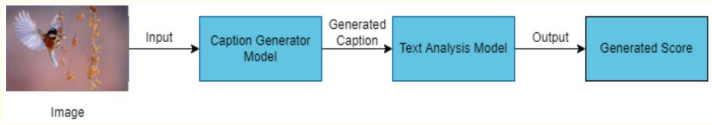

Figure 8: General Model.

Results

First we used the CNN model inside the images to test the accuracy of prediction of the quality of the images. After testing we got a low accuracy of 28.72%. Because of this low accuracy, we tried to make different type CNN models and combined the results of all the CNN-models to generate the score of the image. With this the accuracy improved to around 30-35% accuracy.

In our text analysis model we made a vocabulary of words and it consisted of a total of 1687 words and after which we got after this process was of around 40%. We tested the model using dataset of 800 images, and the model gave correct accuracy as shown in Table 2.

Score	1	2	3	4	5
Predicted count	3	10	276	24	6
Groundtruth count	40	127	357	188	88

Table 2: Summary of the scores generated by our model.

Table 2. shows the number of predicted images which have same scores as of ground truth images and are distributed using scores ranging from 1 to 5. The extreme low and high scores, like 1 and 5, are predicted less accurately because of less sample size of these scores.

Conclusion

In this paper, we calculate Image Aesthetic Score by first generating the captions for the image and then use it to predict the aesthetic score. We predict the aesthetic score of an image on a scale of 1-5, and while taking into consideration the subjectivity of the task, our model shows promising results. Future work maybe about using photo critique captions to enhance the performance. Image aesthetics is a partially subjective topic as there are some aspects of the image that are more appealing to the person's eye than the others, and the paper presents a novel technique to generate a typical score about the quality of an image by using the image captioning technique. The Image Captioning model has been trained using Convolutional Neural Network, Long Short Term Memory, Recurrent Neural Networks, and Attention Layer and we achieved the accuracy of ~40%.

References

1. AM Obeso., et al. "Forward- backward visual saliency propagation in Deep NNs vs internal attentional mechanisms". 2019 9th International Conference on Image Processing Theory, Tools and Applications, IPTA (2019).
2. V Mnih., et al. "Recurrent models of visual attention". Advances in Neural Information Processing Systems 3 (2014): 2204-2212.
3. X He and Y Peng. "Fine-grained image classification via combining vision and language". Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2017): 7332-7340.
4. X Bai., et al. "Integrating scene text and visual appearance for fine-grained image classification". IEEE Access 6 (2018): 66322-66335.
5. Z Yu., et al. "Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering". IEEE Transactions on Neural Networks and Learning Systems 29.12 (2018): 5947-5959.
6. Y Zhou., et al. "Joint image and text representation for aesthetics analysis". MM 2016 - Proceedings of the 2016 ACM Multimedia Conference (2016): 262-266.
7. X Tang, W Luo and X Wang. "Content-based photo quality assessment". IEEE Transactions on Multimedia 15.8 (2013): 1930-1943.
8. R Datta., et al. "Studying aesthetics in photographic images using a computational approach". Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 3953 (2006): 288-301.
9. L Guo., et al. "Image esthetic assessment using both hand-crafting and semantic features". Neurocomputing 143 (2014): 14-26.
10. M Nishiyama., et al. "Aesthetic quality classification of photographs based on color harmony". Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2011): 33-40.
11. DY Kao, R He and K Huang. "Deep Aesthetic Quality Assessment with Semantic Information". IEEE Transactions on Image Processing 26.3 (2017): 1482-1495.
12. M Kucer., et al. "Predicting Image Aesthetics". 27.10 (2018): 5100-5112.
13. Y Chen., et al. "Engineering deep representations for modeling aesthetic perception". IEEE Transactions on Cybernetics 48.11 (2018): 3092-3104.
14. Y Luo and X Tang. "Photo and Video Quality Evaluation". Quality 8.08 (2008): 386-399.
15. Y Ke, X Tang and F Jing. "The design of high-level features for photo quality assessment". Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1 (2006): 419-426.
16. K Xu., et al. "Show, attend and tell: Neural image caption generation with visual attention". 32nd International Conference on Machine Learning, ICML 3 (2015): 2048-2057.
17. S Kong., et al. "Photo aesthetics ranking network with attributes and content adaptation". Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9905 (2016): 662-679.
18. TY Lin., et al. "Microsoft COCO: Common objects in context". Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5 (2014): 740-755.