GENERATING AESTHETIC BASED CRITIQUE FOR PHOTOGRAPHS

Yong-Yaw Yeo†  John See‡ ⋆ Lai-Kuan Wong† Hui-Ngo Goh†

† Faculty of Computing and Informatics, Multimedia University, Malaysia
‡ School of Mathematical and Computer Sciences, Heriot-Watt University Malaysia
(⋆ Corresponding author)

ABSTRACT

The recent surge in deep learning methods across multiple modalities has resulted in an increased interest in image captioning. Most advances in image captioning are still focused on the generation of factual-centric captions, which mainly describe the contents of an image. However, generating captions to provide a meaningful and opinionated critique of photographs is less studied. This paper presents a framework for leveraging aesthetic features encoded from an image aesthetic scorer, to synthesize human-like textual critique via a sequence decoder. Experiments on a large-scale dataset show that the proposed method is capable of producing promising results on relevant metrics relating to semantic diversity and synonymity, with qualitative observations demonstrating likewise. We also suggest the use of Word Mover’s Distance as a semantically intuitive and informative metric for this task.

Index Terms— Image captioning, aesthetic quality assessment, text synthesis, encoder-decoder network, Word Mover’s Distance

1. INTRODUCTION

Humans are capable of perceiving many facets of information from an image, such as the contents, styles, qualities, themes and aesthetics of an image. Lately, image captioning has been increasingly studied in the field of computer vision and artificial intelligence, with the aim to generate dynamic yet human-like captions from photographs. Straightforward captions that describe the contents, i.e. objects, scene and interactions between them, are typically done in most earlier works [1, 2, 3, 4, 5, 6], where the main focus is on generating factual-based captions. A number of recent works [7, 8, 9] diversified the task of image captioning to different styles – some based on sentiments (positive, negative) and themes (romantic, humorous). However, generating aesthetic-based captions remains a relatively new and challenging task.

In this paper, we focus on the task of generating aesthetic-based image captions, which has many useful end-user applications such as automatic photo review/critique, image retrieval and even as a personal guidance to improve photography skills. It is also natural for human beings to focus on the aesthetic aspect of the given image – how visually appealing the image is when told to comment or critique on an image. While factual-based image captioning approaches rely on a typical encoder-decoder architecture [3, 4, 5, 6], this motivates our work to further encode aesthetic features which are then decoded to generate aesthetic-based critique of photos.

Lately, the availability of large-scale data has also advanced the task of image aesthetic assessment [10, 11, 12, 13], where machine learning models in particular convolutional neural networks (CNNs), can be trained to predict the aesthetic appeal of images, either as a binary classifier or a regression scorer. We hypothesize that these trained aesthetic features can provide valuable information to guide the decoder towards synthesizing image captions that contain nuances of aesthetic detail. At the same time, not all of the current set of captioning metrics [14] (e.g. BLEU, METEOR, CIDEr) are intuitive enough to capture the diversity of captions based on a semantic embedding space (see Figure 1).

The main contributions of this paper are as follows:

1. We propose a technique of leveraging aesthetic features learned via image aesthetic assessment as the encoder, which then feeds the decoder with salient information for synthesizing aesthetic-based image captions.
2. We demonstrate the advantages of the proposed method, which outperformed existing methods on metrics relating to semantic diversity and synonymity.

3. We propose to adopt the usage of Word Mover’s Distance (WMD) which can intuitively measure the semantic similarity between the generated caption and the ground-truth set.

## 2. RELATED WORK

In this section, we will provide a brief review of related works in image aesthetic assessment and image captioning.

### Image Aesthetic Assessment

[15] introduced a large-scale dataset, namely AVA, for aesthetic visual analysis, which brought about significant advances to this task. Researchers then began to utilize deep CNNs [16, 17, 12, 18] to provide stronger generalization by learning robust hierarchical features that worked well. In these works, image aesthetic assessment can be performed as a binary classification problem [16, 18], a regression problem [17] or a distribution matching problem [12]. Recently, [13] successfully showed that with similar CNN architecture as previously mentioned works, the use of Earth’s Mover Distance (EMD) [19] as the training loss function boosted the performance by a significant margin.

### Image Captioning

In the early stages of image captioning, traditional approaches using graphs or trees [1] and ML approaches [2] were generally used. However, these approaches have limited flexibility and are heavily dependent on hand-crafted features. [3] proved that the use of deep learning framework alongside a large dataset, e.g. Flickr30k, MSCOCO etc., can generate content-relevant image descriptions. Most works [3, 4] applied the encoder-decoder architecture for the image captioning task, whereby the encoder is a CNN which is pre-trained on the image classification task, while the decoder that projects an output sequence is a Recurrent Neural Network (RNN), which is fundamental in NLP tasks such as machine translation. Subsequent ideas of using visual attention [5] and region of interests [6] made big strides in improving factual-based captioning. Several works explored the possibility of diversifying the type of captions generated by machines. [7] looked into how sentiment oriented captions (positive or negative) can be generated, with two LSTM streams to handle factual and sentiment contents. Stylized captions such as romantic and humorous captions have been proven to be achievable in [9] while a recent work by [8] introduced an end-to-end trainable model, MSCap, without relying on style-specific paired data.

### Aesthetic Image Captioning

Aesthetic-based image captioning remains a challenging task as opinions on the aesthetic aspect of images are highly subjective among different people. [20] proposed an approach to solve this task by introducing multiple LSTMs specialized in several common aesthetic aspects, i.e. Composition, Color, Lighting and Subject of Photo. Each LSTM takes in features from a CNN encoder to generate aspect-centric aesthetic features, which are passed to a general LSTM decoder for caption generation. The obvious drawback here is the use of multiple LSTMs, which is computationally very expensive. [21] proposed a model that can rate the aesthetic level of images as well as produce comments in an end-to-end manner via a shared semantic layer. A recent work [22] focused on learning the aesthetic topics from image reviews using Latent Dirichlet Allocation (LDA), with its features incorporated as a weakly-supervised label. While this idea is promising, it is heavily reliant on the effectiveness of the topic modeling step.

## 3. METHODS

In this section, we briefly describe the training method of our proposed aesthetic image captioning model. Model training is divided into two parts: the first part focuses on training an image aesthetic regressor, while the second part will train the entire pipeline (shown in Figure 2) to produce an aesthetic-based image captioning model.

### 3.1. Aesthetic Features

Motivated by [13], we train an image aesthetic regressor that predicts the score distribution given an image. Using the AVA dataset [15] where each image is labelled with its corresponding D-dimensional vote counts, we perform L1-normalization on the discrete votes across the ordered bins to obtain a normalized probability distribution, resulting in \( \sum_{d=1}^{D} P(d, I) = 1 \). For instance, given an image \( I \), \( P(d = 3, I) \) refers to the probability of votes distributed to bin 3. Since this distribution is being L1-normalized, the mean opinion score (expected value) can be calculated by summing the product of \( d \)-th bin and its distribution, i.e.

\[
\text{Scores}(I) = \sum_{d=1}^{D} d \times P(d, I).
\]

In contrast to [13], which only reported the CNN backbone on VGG16, MobileNet and Inception-v2, we extended this by utilizing ResNet-50 [23]. We managed a slightly better results measured in Spearman Rank Correlation Coefficient (SRCC) and Linear Correlation Coefficient (LCC) between ground-truth and predicted score distribution (SRCC: 0.6648, LCC: 0.6196). Formally, our CNN backbone comprises of a pretrained ResNet-50 on ImageNet [24], but only allowing gradients to flow through the top two residual blocks. The last block is flattened and connected to a FC-layer (256 nodes) followed by the output layer of \( D \)-dimension. We adopt the Earth’s Mover Distance [19] as our training loss function:

\[
EMD(P, \hat{P}) = \left( \frac{1}{D} \sum_{d=1}^{D} (CDF_P - CDF_{\hat{P}}) \right)^{\frac{1}{2}}
\]

where \( P \) and \( \hat{P} \) are the ground-truth and predicted score distributions respectively. Here, a regressor is trained to perform image aesthetic assessment on a given image \( I \), predicting the aesthetic score distribution. It is worth noting that the CNN backbone of this regressor is then used as the feature encoder in Section 3.2 to extract aesthetic features from images.
Generates aesthetic captions that are relevant to the content present in images, together with related aesthetic aspects is challenging. To this end, we designed our aesthetic-based image captioning model based on an encoder-decoder architecture since this architecture has been proven to work quite well in most image captioning models mentioned earlier. However, alterations to the encoder is necessary to further incorporate aesthetic features extracted from images. In order not to lose part of the factual contents present in the images, we propose a multi-encoder to perform the encoding task.

**Encoder.** In our proposed model, an image $I$ will be encoded twice: first by a pre-trained CNN on ImageNet [24], denoted as $Enc_{fact}$ and secondly, by a CNN encoder trained to perform image aesthetic assessment (as in Section 3.1), denoted as $Enc_{aest}$. Intuitively, by encoding image $I$ with $Enc_{fact}$, factual information (i.e. objects, scenes) from $I$ will be extracted since these features are commonly used in computer vision tasks such as image classification and object detection. Meanwhile, by encoding image $I$ with $Enc_{aest}$, aesthetically relevant information will be extracted since $Enc_{aest}$ has been trained to rate images based how visually appealing they are. As both sets of features represent contrasting aspects (factual and aesthetic) of the given image, we propose to introduce learnable weights in a linear layer $(\alpha, \beta)$ to integrate these features seamlessly into the hidden states and cell states of the decoder LSTM. Formally, the encoder output is given as:

$$F = \alpha \cdot Enc_{fact}(I) + \beta \cdot Enc_{aest}(I) \quad (2)$$

**Decoder.** The decoder network consists of a single LSTM layer with hidden state of size 512, connected to a final output layer that corresponds to the size of the vocabulary. Since the dataset used is large, we found that initializing the word embedding randomly within the range of $[-0.1, +0.1]$ for encoder-decoder training is sufficiently good. We incorporate soft visual attention of [5] on both encoded features from image $I$ and concatenate these features with their corresponding word vector at timestep $t$ to allow the decoder to “attend” to relevant features. Gradient of LSTM is clipped at 5.0 to prevent exploding gradients which commonly occurs during LSTM training. Weighted dropout, or DropConnect [25] is added to the LSTM, applying a uniform dropout mask to the weights of the recurrent connections. An additional dropout layer is added before the final vocabulary prediction layer with dropout rate of 0.5. The entire model is trained with Adam optimizer with decoder learning rate set at $10^{-3}$ and encoder learning rate at $10^{-4}$. Both learning rates are decayed with a one tenth factor every eight epochs without improvement, measured based on BLEU-4 validation scores.

**4. EXPERIMENTAL RESULTS**

**4.1. Datasets**

The large-scale AVA dataset [15] is used for training both encoder and decoder models. For the aesthetic assessment task, the score labels of the images ($D = 10$) are obtained directly from the dataset, thus only pre-processing steps such as the probability distribution normalization (in Sec. 3.1) are necessary. The training and validation splits follow the original settings in [15]. Meanwhile, for the aesthetic-based image captioning task, we split the images according to the method proposed by [22], resulting in 230,695 training images and 9,362 validation images. The original AVA dataset by [15] does not come with captions. We adopted the AVA-Captions dataset by [22] where the original comments of AVA photos [26] have been pre-processed and filtered through an informativeness score, which chooses nouns for unigram vocabularies and “descriptor-object” patterns (i.e. noun, adjective/adverb in first term, and noun/adjective in second term) for bigram vocabularies. The resulting dataset comprised of a total of $\sim$1.318 million captions. From this, we built our own vocabulary with 33,380 unique tokens where each word has a minimum frequency occurrence of 5 across the entire corpus. The word embedding dimension is set to 256, initialized randomly between $-0.1$ to 0.1 and kept learnable during training.
Table 1. Comparison between methods proposed in [22] and our approach. The captions generated by our approach tends to be more opinionated and human-like. It is also particularly accurate in identifying the waterfall in the second photo.

<table>
<thead>
<tr>
<th>Methods</th>
<th>FF [22]</th>
<th>TF [22]</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual Features + LSTM (FF)</td>
<td>0.500</td>
<td>0.535</td>
<td>0.464</td>
</tr>
<tr>
<td>Topic Features + LSTM (TF) [22]</td>
<td>0.280</td>
<td>0.282</td>
<td>0.238</td>
</tr>
<tr>
<td>Aesthetic + Factual Features + LSTM (Ours, AF + FF)</td>
<td>0.149</td>
<td>0.150</td>
<td>0.122</td>
</tr>
<tr>
<td></td>
<td>0.073</td>
<td>0.074</td>
<td>0.063</td>
</tr>
<tr>
<td></td>
<td>0.105</td>
<td>0.107</td>
<td>0.109</td>
</tr>
<tr>
<td></td>
<td>0.253</td>
<td>0.254</td>
<td>0.262</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td>0.060</td>
<td>0.059</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>0.062</td>
<td>0.061</td>
<td>0.055</td>
</tr>
</tbody>
</table>

Table 2. Performance comparison on Aesthetic Image Captioning using AVA-Comments.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>FF [22]</th>
<th>TF [22]</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU-1</td>
<td>0.500</td>
<td>0.535</td>
<td>0.464</td>
</tr>
<tr>
<td>BLEU-2</td>
<td>0.280</td>
<td>0.282</td>
<td>0.238</td>
</tr>
<tr>
<td>BLEU-3</td>
<td>0.149</td>
<td>0.150</td>
<td>0.122</td>
</tr>
<tr>
<td>BLEU-4</td>
<td>0.073</td>
<td>0.074</td>
<td>0.063</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.105</td>
<td>0.107</td>
<td>0.109</td>
</tr>
<tr>
<td>ROUGE</td>
<td>0.253</td>
<td>0.254</td>
<td>0.262</td>
</tr>
<tr>
<td>WMD</td>
<td>N/A</td>
<td>N/A</td>
<td>0.077</td>
</tr>
<tr>
<td>CIDEr</td>
<td>0.060</td>
<td>0.059</td>
<td>0.051</td>
</tr>
<tr>
<td>SPICE</td>
<td>0.062</td>
<td>0.061</td>
<td>0.055</td>
</tr>
</tbody>
</table>

4.2. Results and Analysis

Table 2 compares the performances on the recently proposed AVA-Captions [22] dataset on a number of standard captioning metrics [14]. In summary: BLEU-n measures the n-grams similarities between reference and hypothesis sentences; METEOR uses synonyms and paraphrases matching, measuring the harmonic mean of precision and recall of unigram matches in both sentences; ROUGE compares overlapping n-grams, word sequences and word pairs; CIDEr evaluates image captioning purely on the linguistic level, with tf-idf weighting over n-grams; SPICE parses the hypothesis sentences into a scene graph, and calculates $F_1$-scores based on the agreement between reference and hypothesis tokens.

In terms of metrics performance, our proposed method outperformed methods proposed by [22] on METEOR and ROUGE but not in other metrics such as BLEU, SPICE and CIDEr. With this observation, we can conclude that our model is unable to predict the exact same tokens as with the ground-truth (which is important for factual captioning), but its superiority on the METEOR and ROUGE indicates that our model is able to generate relevant synonyms of phrases of the ground-truth captions. This is evidential that our approach can create more diverse and human-like comments.

The majority of evaluation metrics reported involved measuring similarities in n-grams or word phrases, between reference and hypothesis captions. As language itself consists of many variations, sentences with totally different tokens can be semantically similar. Since aesthetic captions are subjective and heavily dependent on personal opinions, we adapted Word Mover’s Distance (WMD) [27] to measure the semantic differences between ground-truth and generated captions. In WMD, the captions first have their stopwords removed, before computing their respective normalized bag-of-word (nBOW) vectors, $v$ and $v'$. The document distance between the two captions are calculated in word2vec embedding space (i.e. embedding $x_i \in \mathbb{R}^d$) as $d_{WMD}(i, j) = \sum_{i,j} T_{ij}||v_i - v_j||_2$ where the distance between word $i \in v$ and $j \in v'$ is given by $c(i, j) = ||x_i - x_j||_2$ and $T_{ij}$ is a flow matrix denoting the minimum cost to move each word in $v$ to $v'$. Similar to findings in [14], we observe WMD to be a feasible metric that can capture the semantic relevancy between the generated caption and the ground-truth. In future, we surmise that a human-in-the-loop procedure can provide substantial feedback to refine the learning process of this framework.

5. CONCLUSION

In this paper, we propose a new approach towards generating aesthetic-based image captions by directly leveraging on aesthetic features encoded by an image aesthetic scorer, alongside factual features trained from object-based classifier. Our method is able to outperform existing works on several relevant metrics of semantic diversity and synonymity.

Acknowledgement: This work is supported by Malaysia Ministry of Education FRGS Research Grant (Project No: FRGS/1/2018/ICT02/MMU/02/2).
6. REFERENCES


[22] Koustav Ghosal, Aakanksha Rana, and Aljosa Smolic, “Aesthetic image captioning from weakly-labelled photographs,” in ICCV Workshops, 2019, pp. 0–0.


