Generating Captions for Images of Ancient Artworks

Shurong Sheng
Department of Computer Science, KU Leuven
shurong.sheng@cs.kuleuven.be

Marie-Francine Moens
Department of Computer Science, KU Leuven
sien.moens@cs.kuleuven.be

ABSTRACT

The neural encoder-decoder framework is widely adopted for image captioning of natural images. However, few works have contributed to generating captions for cultural images using this scheme. In this paper, we propose an artwork type enriched image captioning model where the encoder represents an input artwork image as a 512-dimensional vector and the decoder generates a corresponding caption based on the input image vector. The artwork type is first predicted by a convolutional neural network classifier and then merged into the decoder. We investigate multiple approaches to integrate the artwork type into the captioning model among which is one that applies a step-wise weighted sum of the artwork type vector and the hidden representation vector of the decoder. This model outperforms three baseline image captioning models for a Chinese art image captioning dataset on all evaluation metrics. One of the baselines is a state-of-the-art approach fusing textual image attributes into the captioning model for natural images. The proposed model also obtains promising results for another Egyptian art image captioning dataset.

CCS CONCEPTS

• **Information systems → Multimedia content creation**; • **Computing methodologies → Scene understanding**.

KEYWORDS

Image captioning, Artwork type, Neural encoder-decoder

ACM Reference Format:
Shurong Sheng and Marie-Francine Moens. 2019. Generating Captions for Images of Ancient Artworks. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350972

1 INTRODUCTION

The application of artificial intelligence techniques to the cultural heritage field has attracted increasing attention in recent years [8, 21, 22, 28–30, 39]. Most of these work focus on automatic metadata annotation such as predicting the author, material, and date of an artwork. In this work, we target automated image caption generation for cultural heritage images employing a deep neural network. This captioning of images would allow a visitor of a museum or cultural heritage site to obtain a detailed description of an artwork on his or her mobile device by just taking the picture of the artwork. Such a service would facilitate the personalized interaction between artworks and art lovers. Besides, such a captioning system could be used for automatically generating explanations in a catalogue of artworks to be searched online, hereby saving time and labor of manual annotation. We show the feasibility of captioning images of artworks with two collections of images, one containing ancient Egyptian (8000 B.C. - 1000 A.D.) and the other ancient Chinese (1368 A.D. - 1912 A.D.) artwork images. Figure 1 shows an exemplary artwork image and its ground-truth caption.

Captions: (1) Lion-Shaped Furniture Leg. (2) In both Egypt and Nubia the lion was associated with the sun god and symbolized royalty.

Figure 1: An Egyptian artwork image with its caption.

Generating captions for ancient artworks faces three key domain-specific challenges compared with image captioning for natural images. First, the captions for ancient artworks often contain high-level semantic information beyond the image content such as the background of a historical person, human judgment and uncertain illustrations based on expert speculation about artworks which are indicated by the caption word 'probably'. For example, the Egyptian artwork caption in Figure 1 contains the explanation of the lion's symbolic meaning in history. This information is obviously difficult to be derived from the artwork image. In [37], the authors also demonstrate that a cultural image is a narrative image with stories behind it. In such a setting, it is challenging to generate good descriptions leveraging only on artwork images. Second, professional knowledge is needed to annotate ancient artwork images. This makes it unrealistic to train with a dataset of a size similar to the one of datasets with natural images such as MSCOCO [20]. The datasets employed in previous research on automated cultural image annotation [35, 37, 38] are composed only of a few hundred images. Lastly, the textual descriptions for ancient artifacts used for

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10.
https://doi.org/10.1145/3343031.3350972

training a model often contain many special symbols and incomplete sentences [18], posing additional challenges when generating well-formed captions.

To address the issues mentioned above, we introduce an artwork type enriched image captioning model in the encoder-decoder paradigm. The assumption is that the type of an artwork has a latent influence on the caption content no matter whether it actually occurs in the caption or not. We select this meta-data also because we empirically believe that the artwork type will be more accurately predicted than other meta-data information e.g., the period of an artwork [30]. We have explored several approaches to incorporate the artwork type into the captioning model, for example, by concatenating the input image vector and the artwork type vector or by computing a step-wise weighted sum of the artwork type vector and the hidden representation that produces the output of the decoder. Overall, the main contributions of this paper are:

(1) We collect two image captioning datasets referring to ancient Egyptian art and ancient Chinese art. They contain respectively 17,940 and 7,607 artwork images and corresponding captions.
(2) We propose an artwork type enriched encoder-decoder image captioning model for ancient works. This model explicitly encodes the relative importance of the artwork type vector and the hidden representation that produces the output of the decoder before they are fed into the fully connected layer. It outperforms three baseline image captioning models for the Chinese art image captioning dataset on all evaluation metrics. One of the baselines is a state-of-the-art approach among the works integrating textual image attributes into the captioning model for natural images. The proposed model also achieves encouraging results for the Egyptian art image captioning dataset.
(3) We investigate multiple approaches to incorporate the artwork types into the captioning model and test the efficiency of three existing image captioning models built for natural images when they are applied to annotate cultural images.
(4) A comprehensive quantitative and qualitative analysis on the results of all the models introduced in this work is made to guide future research.

The remainder of this paper is organized as follows. Section 2 reviews related research. Next, Section 3 describes our model architecture. Section 4 illustrates the experiments and evaluation metrics and then Section 5 discusses the results obtained by different models. Finally, Section 6 concludes this paper and provides directions for future research.

2 RELATED WORK

2.1 Multi-class and Multi-label Classification

Single-label multi-class image classification [6] regards the task of classifying the image with one type of artwork chosen from a set of artwork types. In the multi-label classification of images [7], an image can be labeled with multiple compatible artwork types from a set of artwork types, for example, an artwork can be labeled as both statue and stone. The artwork types in the single-label multi-class classification task are fine-grained, e.g., an artwork is labeled as metalwork relief instead of a coarse-grained category relief. In contrast, coarse-grained artwork types are used in the multi-label image classification task, e.g., relief, sculpture, and stone. We will experiment with both tasks and corresponding classification taxonomies when testing whether the artwork type granularity influences captioning performance.

Many works have been devoted to multi-class cultural image classification. In [25], the authors classify images of Mexican buildings into three different architectural styles with GoogleNet [32] and AlexNet [16]. In [41], a small dataset comprising 432 images corresponding to 4 regions of cultural interest is collected from Flickr and then the image labels are predicted by the fine-tuned AlexNet. This work demonstrates that art type prediction is more accurate using the fine-tuned AlexNet convolutional neural network (CNN) architecture than when implementing a support vector machine (SVM) with SIFT features [24]. Yang and Min [39] perform classification on multiple art datasets using various CNN architectures and confirm that DenseNet [12] achieves the best performance. Unlike the dataset used in [39] which contains mostly paintings, our dataset consists of various artwork types. Therefore, in this paper we have chosen ResNet [9] as backbone neural architecture which is proved to obtain good results on multiple computer vision tasks, e.g., object detection and image classification. No efforts have been put into the multi-label classification for cultural images as far as the authors know.

2.2 Image Caption Generation

Several studies have contributed to generating annotations or descriptions for cultural images [35, 37, 38]. In [35], an automatic linguistic indexing system of pictures is built to learn the expertise of a human annotator based on a small-scale annotated EMPEROR Collection. More intuitively, the authors first train a 2-D multiresolution hidden Markov model (2-D MHMM) to find the correspondence between a cultural image and its descriptive concepts. Then, when a new unannotated image comes in, this system describes the new image with the learned concepts of a similar image indexed in the system. This approach is called retrieval-based image captioning [11], but it cannot generate image-specific captions. Xu and Wang [38] and Xu et al. [37] introduce respectively an ontology and hierarchical model to perform image description creation of Dunhuang frescoes which constitute a special area in the field of Chinese cultural heritage. The two latter works leverage low-level image features e.g., the image texture and meta-data of cultural images but encode them in different-structured models. All of the above works heavily rely on feature engineering.

Current state-of-the-art encoder-decoder image captioning models for natural images are data-driven methods using powerful deep neural networks. Vinyals et al. [34] introduce the first image captioning model using a classical encoder-decoder schema. In contrast to Vinyals et al. [34] who consider the image as the input for the first decoding step, the models proposed in [36] dynamically attend to the input image vector at each step in the decoder module. Although these works obtain good results for natural images, it is not known yet whether they can achieve equally nice performance for cultural images due to the domain-specific challenges mentioned in Section 1. In this work, we test the effectiveness of the above models for generating captions for images of ancient Egyptian and Chinese

artworks. Recently, Anderson et al. [2] have proposed to represent the whole image with salient regions and achieve state-of-the-art performance when this image representation method is employed for image captioning of natural images. Due to the lack of region information for the images of the artworks, in the research that we report we only use representations of the whole image. Different from Vinyals et al. [34] and Xu et al. [36] who pursue performance improvement by manipulating images, in [40], the authors treat both the textual image attributes and the previous word in a caption as independent inputs of the decoder module. Our artwork type enriched captioning model is mainly inspired by this work. But instead of applying an element-wise addition to the attribute vector and the word vector obtained from the previous step, we merge the artwork type and hidden representation output of the decoder in the very top layer of the decoder. We also explicitly model the relative importance of these two elements at each decoding step. This way, we assume that the artwork type can guide the decoding network more clearly.

3 METHODOLOGY

Figure 2 gives an overview of our model. We have implemented an encoder-decoder framework for image captioning where the encoder is a CNN and the decoder is a long short-term memory (LSTM) network [31]. The encoder in our model extracts not only the input image vector but also the artwork type representation of an artwork. These two vectors are then inputted into different gates of the decoder. Formally, given an input image I_i, a CNN encoder is applied to extract its feature vector I_{if} and the artwork type vector T_i. Then, the LSTM decoder is adopted to generate the caption $S_i = \{S_{i0}, S_{i1}, ..S_{iN_s}\}$ depending on the two elements I_{if} and T_i. N_s equals the caption length.

In this section, we first describe the encoder structure to extract the image feature vector and the artwork type vector. Next, we elaborate on how we incorporate the artwork type vector into the decoder module. Finally, we introduce alternative models that we have implemented.

3.1 Encoder

The encoder is employed to extract the image feature vector and artwork type vector. As mentioned in Section 2.1, we perform multi-class classification and multi-label classification to check the granularity effects of the artwork types. For both multi-class classification and multi-label classification, the encoder uses the 18-layer ResNet18 structure followed by an adaptive pooling operation on the sub-image information as shown in the first row of Figure 2. The ResNet18 structure is eventually selected because it extracts a lower-dimensional input image feature vector and can thus reduce the dimensional gap between the input image vector and the artwork type vector. This will eliminate the dimensional influence when modulating their relative importance in the captioning task. The CNN model is initialized from the ResNet18 pre-trained on ImageNet [16]. This shared CNN is then fine-tuned on the multi-class or multi-label dataset.

In the *multi-class* classification task, given an input image I_i, the output of the last fully-connected layer is fed into a c-way softmax over the C class labels. Suppose that there are N training examples, $y_{si} = [y_{si1}, y_{si2}, ..., y_{siC}]$ is the ground-truth one-hot encoded vector of the $i-th$ image where $y_{sij} = 1$ if the image is annotated with artwork type j, and $y_{sij} = 0$ otherwise. If the predictive probability vector is $p_{si} = [p_{si1}, p_{si2}, ..., p_{siC}]$, the cost function to be minimized for this task is:

$$C_s = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{sij} log(p_{sij}) \quad (1)$$

In contrast, the output of the last fully-connected layer in the *multi-label* classification task is inputted into a sigmoid function where given an input image I_i, the sum of all the elements in the output vector does not equal to one. Suppose the label vocabulary size is E in this task, the ground-truth multi-label vector is $y_{mi} = [y_{mi1}, y_{mi2}, ..., y_{miE}]$, and $p_{mi} = [p_{mi1}, p_{mi2}, ..., p_{miE}]$ is the respective predicted vector. The goal is to minimize the following binary cross entropy loss:

$$C_m = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{E} \left(y_{mij} log(p_{mij}) + (1 - y_{mij}) log(1 - p_{mij})\right) \quad (2)$$

Training. Mini-batch stochastic gradient descent is used to optimize the fine-tuning process with a mini-batch size of 256. The learning rate and the momentum are set to 0.001 and 0.9 respectively, no dropout operation is involved. All the layers in ResNet18 except the last fully connected layer are fixed during training. We executed 100 epochs in total. The 512-dimensional vector obtained from the adaptive pooling operation serves as the input image feature vector I_{if} and the network output is considered as the artwork type representation T_i. We will introduce how we use these vectors in the following section. Note that training the encoder is performed beforehand instead of end-to-end with the decoder.

This encoder structure can be replaced by a region proposal network [27] if partial-image information is available in the future.

3.2 Decoder

The decoder is an image caption generator, i.e., it is trained to predict each word of the caption sentence for an input image. But instead of treating only the input image as an element in the decoding process, we also consider the artwork type vector obtained from the encoder module as an independent input. In such a case, given an input image I_i, the goal is to maximize the probability of the correct description given the image and its artwork type using the following formulation:

$$\theta^* = \arg\max_{\theta} \sum_{(I_{if}, T_i)} log\, p(S_i | I_{if}, T_i) \quad (3)$$

where θ are the parameters of our model, and I_{if} and T_i are respectively the image feature vector and artwork type vector. $S_i = \{S_{i0}, S_{i1}, ..S_{iN_s}\}$ represents the correct transcription, and the probability to generate this sentence can be computed as follows according to the language model [4]:

$$log\, p(S_i | I_{if}, T_i) = \sum_{t=1}^{N_s} log\, p(S_{it} | (I_{if}, T_i), S_{i0}, ..., S_{i(t-1)}) \quad (4)$$

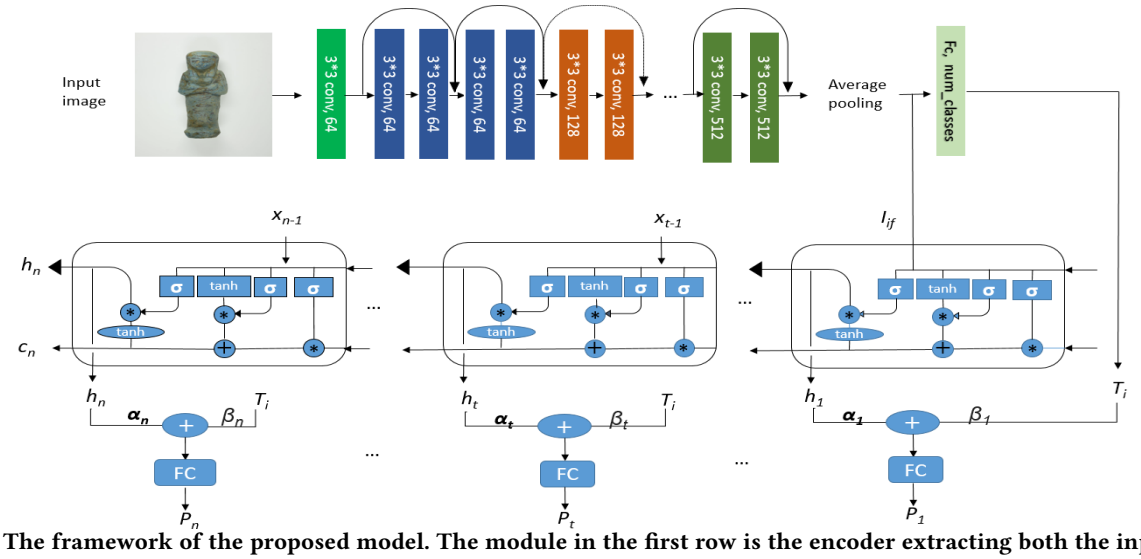

Figure 2: The framework of the proposed model. The module in the first row is the encoder extracting both the input image vector and the artwork type vector. The decoder in the second row contains a merge-gate where the inputs are the hidden representation of an LSTM network and the artwork type vector. This merge gate is then used to predict the probabilities that each word in the dictionary will be generated.

where we dropped the dependency on θ for convenience.

It is natural to model equation 4 with an LSTM network. The core of the LSTM network is a memory cell c that encodes at every time step the inputs having been observed up to this step. In our model, the LSTM network produces a caption by generating one word at every time step conditioned on the artwork type vector T_i, the hidden state created in the last time step h_{t-1} and the previously generated word x_t. We implement it as follows for time step t:

$$f_t = sigmoid(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \quad (5)$$

$$i_t = sigmoid(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \quad (6)$$

$$o_t = sigmoid(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \quad (7)$$

$$g_t = tanh(W_{gx}x_t + W_{gh}h_{t-1} + b_g) \quad (8)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (9)$$

$$h_t = o_t \odot tanh(c_t) \quad (10)$$

$$p_t = FC(\alpha_t * h_t + \beta_t * T_i) \quad (11)$$

Here, f_t, i_t, o_t and c_t are respectively the input, forget, output and memory state of the LSTM. The various W matrices and b vector are trainable parameters. The image, the words, and the artwork type are mapped to the same space, where the image is represented by its CNN embedding, the words by their word embeddings and the artwork type by using a CNN followed by a linear projection layer. Symbol $*$ in equation 11 refers to a scalar multiplying a vector.

In this decoder module, we propose to set up two trainable weight vectors $\alpha = [\alpha_1, \alpha_2, ..., \alpha_{N_{max}}]$ and $\beta = [\beta_1, \beta_2, ..., \beta_{N_{max}}]$ to explicitly encode the relative importance of respectively the hidden representation and artwork type before they are used to predict the caption words probability. N_{max} is the maximum length of the training captions. More specifically, at each time step t, a weighted sum of the hidden representation h_t and the artwork type vector T_i serves as the input to a fully connected (FC) layer and the probability of each word in the training vocabulary will be generated. This approach is represented by equation 11. The weights in α and β are automatically learned by the model during training. In such a way, we let the two elements h_t and T_i guide the learning of the LSTM network more clearly. Moreover, by treating the artwork type vector T_i as an independent input to the fully connected layer, its guiding process is not affected by the previously generated word in the decoder, the artwork type vector can therefore guide the LSTM in a right way even if the previously generated word is wrongly predicted.

Training. We represent each word as a one-hot vector $S_{it} \in R^V$ where V equals to the size of the vocabulary. The word embedding of a word is obtained by applying a projection layer to the one-hot vector. We denote by S_{i0} a special start word and by S_{iN_s} a special delimiter word which designate the start and end of the sentence. In particular, by emitting this delimiter word, the LSTM signals that a complete sentence has been generated. The image feature vector I_{if} is only inputted once, at $t = 0$, to inform the LSTM about the image contents. Our loss is the sum of the negative log likelihood of the correct word at each step as follows:

$$L(I_{if}, T_i, S_i) = -\sum_{t=1}^{N_s} log\, p_t(S_{it}) \quad (12)$$

The above loss is minimized with regard to all model parameters in the decoder. The weight parameters in equation 5-11 are initialized from a uniform distribution with a range between -0.1 and 0.1. The weight parameters in α and β are initialized as random numbers between 0 and 1 from a normal distribution.

We refer to this model as LSTM-MC-$OUT_{dynamic}$ when using the multi-class classifier as the encoder and LSTM-ML-$OUT_{dynamic}$ when using the multi-label classifier as the encoder. The two models are differentiated by 'MC' (multi-class) and 'ML' (multi-label) in their model names. 'dynamic' in the model names is an indication of the multiple different modulating weights in α and β and 'OUT'

suggests that the artwork type is treated similarly to the hidden representation as the output of the LSTM network.

3.3 Model variants

To better explore how to incorporate the artwork type into the captioning model, we investigate several variants of the LSTM-MC-$OUT_{dynamic}$ model that might bring performance improvements. The settings of these models are exactly the same as the one used in the LSTM-MC-$OUT_{dynamic}$ model except for the following parts:

(1) **LSTM-MC-OUT** This model also encodes the artwork type vector T_i together with the hidden representation output h_t as model LSTM-MC-$OUT_{dynamic}$. But instead of considering their relative importance to conduct the final prediction, this model applies a direct element-wise addition of the two vectors with this formulation:

$$p_t = FC(h_t + T_i) \tag{13}$$

(2) **LSTM-MC-$IN_{dynamic}$** Instead of putting the artwork type vector obtained by the multi-class classifier right before the fully connected layer to predict the caption word in the LSTM decoder (equation 11), this model feeds the artwork type vector to the decoder along with the word generated from the previous step, i.e., the x_t in equation 5-8 are replaced with the following equation:

$$x_t = \alpha_t * x_{w(t-1)} + \beta_t * T_i \tag{14}$$

where $x_{w(t-1)}$ is the word generated at time step $t-1$ and T_i is the artwork type vector. Correspondingly, the equation to predict the word at time step t becomes:

$$p_t = FC(h_t) \tag{15}$$

'IN' in the model name suggests that the artwork type is treated similarly to the previously generated word as the input of the LSTM network.

(3) **LSTM-MC-OUT_{static}** This model incorporates the artwork type vector to the LSTM in the same way as the LSTM-MC-$OUT_{dynamic}$ model. But instead of setting up two trainable vectors α and β which can modulate the artwork type vector and the hidden representation in a step-wise fashion, this model sets up two trainable scalars in the decoder which are the same for all the time steps in the decoding process. We use 'static' in the model name to indicate the modulating behavior in this model.

(4) **LSTM-MC-IN_{static}** This model is the same as the LSTM-MC-$IN_{dynamic}$ model but sets up two trainable scalars to modulate the previously generated word $x_{w(t-1)}$ and the artwork type vector T_i as model LSTM-MC-OUT_{static} does.

(5) **LSTM-MC-global controller** This model concatenates the image feature vector and the artwork type vector, then feeds the concatenated vector into the first step of an LSTM network to generate caption words of the respective image.

The multi-class classifier encoder in the above models can be replaced by the multi-label classifier for the *ancient Egyptian art image captioning dataset* introduced in Section 4.1, but we leave this out to keep the proposed model variants employed for the two captioning datasets consistent.

Inference We use beam search to generate words in the test phase. It iteratively considers the set of the k best sentences up to time t as candidates to generate sentences at time step $t+1$, and keep only the resulting best k of them.

4 EXPERIMENTS AND EVALUATION

4.1 Datasets

The datasets involved in this work are collected from the following online sources: the Metropolitan Museum [1], the Brooklyn Museum [2], and the British Museum [3]. Based on these sources, we have created two image captioning datasets: the *ancient Egyptian art image captioning dataset* and the *ancient Chinese art image captioning dataset*. The two datasets are collected based on the geographical location of the origin of the artworks because caption words may differ much depending on the cultural background of the location. Detailed statistics of the two datasets are shown in Table 1. We also build three artwork type classifier datasets to train the encoder: the *Egyptian art multi-class classifier dataset*, the *Egyptian art multi-label classifier dataset*, and the *Chinese art multi-class classifier dataset*. The artworks types in the classifier datasets are standard artwork type categories defined in the Metropolitan Museum or the British Museum. We have not managed to find a suitable multi-label classifier dataset for Chinese art. These classifier datasets are created separately from corresponding captioning dataset, i.e., the overlaps between a classifier dataset and respective image captioning dataset are not counted. Statistics of the three classifier datasets are given in Table 2.

Dataset	Num. Artworks	Aver. Len	Num. Tokens
Egyptian	17940	9	10722
Chinese	7607	10	5902

Table 1: Statistics of the captioning datasets.

Dataset	Num. Artworks	Num. types
Multi-class Egyptian	5300	237
Multi-label Egyptian	11303	100
Multi-class Chinese	7433	519

Table 2: Statistics of the classifier datasets.

For the classifier datasets, we conduct offline data augmentation e.g., image flipping, cropping or transformation to improve the performance and robustness of the classification models. For the captioning datasets, the paragraph-level descriptions are split into multiple sentences and a maximum of 5 sentences are retained for each artwork to reduce data imbalance. In addition, we remove noisy texts from the captions following a specific pattern, e.g, "See 13.26.59". The number is the accession number of an artifact that obviously cannot be derived from the input image or artwork type. We also remove duplicate images in the captioning datasets based on their hash code. Tokens occurring less than 2 times are removed from the training vocabulary. The datasets are all split into an 80%, 10%, and 10% partition for respectively training, validation, and test.

[1] https://www.metmuseum.org/art/collection
[2] https://www.brooklynmuseum.org/opencollection/collections
[3] https://www.britishmuseum.org/research/collection_online/search.aspx

4.2 Experimental Setup and Evaluation

Besides the models introduced in Section 3.2 and Section 3.3, we also experiment with three existing baseline image captioning models:

(1) Model **NIC** [34] puts the image into the first step of the LSTM decoder with no visual attention and extra textual input involved.
(2) Model **SA** is the soft-attention model introduced in [36] which puts the image vector dynamically into every step of the LSTM decoder.
(3) Model **LSTM-A_5** is the best-working model among all the models proposed in [40] and it is originally one of the state-of-the-art image captioning approaches integrating textual image attributes for natural images. Although this model also treats the textual attributes as independent input, it operates on the attribute-inputs in the lower layer of the LSTM network, that is, an element-wise addition with the previously generated word is exploited. We split the LSTM-A_5 model into two versions LSTM-A_5-MC and LSTM-A_5-ML where the first version utilizes the multi-class classifier as the encoder and the latter one adopts a multi-label classifier as the encoder.

We implement all the image captioning models in Pytorch [14]. The dimensions of the image feature vector, the word embeddings, the artwork type vector, and the hidden layer of the LSTM decoder are set to 512. The Adam optimizer [15] is adopted in the back-propagation process with a learning rate of 0.0001 and a dropout rate of 0.5. A beam size of 5 is empirically selected for all the models. We use accuracy [10] to grade the classification achievement of the encoder. To evaluate the captioning performance, we adopt five types of metrics: BLEU@N [26], ROUGE-L [19], METEOR [3], CIDEr [33] and SPICE [1]. All the metrics are computed by using the code released by the COCO Evaluation Server [5].

5 RESULTS AND DISCUSSIONS

Table 3 shows the results comparison for the captioning models. Overall, the captioning performance for cultural heritage images is lower compared to the performance when captioning natural images. The best BLEU-1 score has been up to 0.817 for natural image captioning as communicated in Microsoft COCO image captioning challenge leader board[4], while the best BLEU-1 score for annotating cultural images is only 0.54 obtained by the proposed model LSTM-MC-$OUT_{dynamic}$ on the *ancient Chinese art image captioning dataset*. For this captioning dataset, model LSTM-MC-$OUT_{dynamic}$ outperforms all other models and holds a maximum improvement of 4% in BLEU-1 score compared with the baseline captioning model LSTM-A_5-MC. For the *ancient Egyptian art image captioning dataset*, model LSTM-MC-$OUT_{dynamic}$, LSTM-MC-OUT, and LSTM-ML-$OUT_{dynamic}$ are competing with the best-working baseline captioning model LSTM-A_5-MC when evaluated by ROUGE_L which shows better correlation with human judgments than BLEU scores.

[4]https://competitions.codalab.org/competitions/3221results

5.1 Quantitative Analysis

(1) Benefits of adding artwork types. We have tested multiple models to assess the effect of adding artwork type information. The proposed model LSTM-MC-$OUT_{dynamic}$ and the baseline model LSTM-A_5-MC achieve better results than both model NIC and model SA for respectively the *ancient Chinese art image captioning dataset* and the *ancient Egyptian art image captioning dataset* as shown in Table 3, wherein LSTM-MC-$OUT_{dynamic}$ on the *ancient Chinese art image captioning dataset* obtains 4% higher BLEU-1 score than that of model NIC, this score is 8% higher compared with model SA. The baseline LSTM-A_5-MC model on the *ancient Egyptian art image captioning dataset* obtains 3% higher BLEU-1 score than model NIC and 4% higher BLEU-1 score than model SA. Model LSTM-MC-$OUT_{dynamic}$ and the baseline model LSTM-A_5-MC attend both the input image feature vector and the artwork type vector while model NIC and model SA rely on only the input image in the captioning process. This proves the advantage of adding artwork type information into the captioning model. But most other models that integrated the artwork type information failed to boost the captioning performance for the *ancient Egyptian art image captioning dataset*, confirming the challenge of effectively adding artwork type into the captioning model. In addition, the proposed model LSTM-MC-$OUT_{dynamic}$ obtains better performance with a large improvement over the baseline LSTM-A_5-MC on the *ancient Chinese art image captioning dataset* and also competing results with LSTM-A_5-MC evaluated by ROUGE_L on the *ancient Egyptian art image captioning dataset*. But the advance of the baseline LSTM-A_5-MC for the *ancient Chinese art image captioning dataset* is much less than our proposed model LSTM-MC-$OUT_{dynamic}$. We therefore conclude that our proposed model LSTM-MC-$OUT_{dynamic}$ is more robust than the baseline LSTM-A_5-MC model.

(2) Effects of visual attention. We also explicitly evaluate the effects of visual attention in the decoding process by comparing model NIC and SA, where NIC feeds the image input into only the first decoding step and SA instead enforces visual attention at each time step in the decoder. Model SA performs worse than model NIC on both image captioning datasets as indicated in Table 3. This is different from the case when visual attention mechanisms are applied to models for captioning natural images, probably because image captions for artworks contain many expressions which are beyond what is happening in the image and in which case visual attentions can mislead the decoder. For example, "*Thus all mummies of humans and animals imitated the mummification process and form followed to reanimate Osiris in the next world*" is a sentence referring to the image of a mummy, and it is actually a story behind the mummy instead of a description of the mummy itself.

(3) Effects of the artwork type granularity. It is interesting to investigate whether the captioning performance improves when we provide multiple coarse-grained compatible types of an artwork into the model. For this purpose, we build model LSTM-A_5-MC and LSTM-A_5-ML which refer to using respectively the fine-grained single-label artwork type and coarse-grained multi-label artwork types in the captioning model. It turns out that integrating a single label model wins on most metrics as shown in Table 3. This is probably due to the lower multi-label classification performance. The accuracy achieved by the multi-class image classification model in

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE_L	METEOR	CIDEr	SPICE
Ancient Egyptian Art Image Captioning Dataset								
NIC [34]	0.44	0.34	0.30	0.27	0.43	0.19	1.82	0.28
SA [36]	0.43	0.34	0.29	0.26	0.43	0.19	1.82	0.27
LSTM-A_5-ML [40]	0.46	0.36	0.31	0.28	**0.44**	0.19	**1.87**	0.28
LSTM-A_5-MC [40]	**0.47**	**0.38**	**0.33**	**0.30**	**0.44**	**0.20**	1.81	**0.29**
LSTM-MC-global controller	0.38	0.27	0.19	0.15	0.41	0.14	1.59	-
LSTM-MC-IN_{static}	0.38	0.27	0.19	0.15	0.41	0.14	1.59	-
LSTM-MC-OUT_{static}	0.42	0.33	0.28	0.25	0.42	0.19	1.74	-
LSTM-MC-$IN_{dynamic}$	0.32	0.19	0.11	0.06	0.37	0.12	1.18	0.20
LSTM-MC-OUT	0.42	0.32	0.28	0.25	**0.44**	0.19	1.84	-
LSTM-MC-$OUT_{dynamic}$	0.42	0.33	0.28	0.25	**0.44**	0.19	1.82	-
LSTM-ML-$OUT_{dynamic}$	0.42	0.33	0.28	0.26	0.43	**0.20**	1.83	0.27
Ancient Chinese Art Image Captioning Dataset								
NIC [34]	0.50	0.42	0.36	0.32	0.53	0.20	0.94	0.18
SA [36]	0.46	0.38	0.33	0.29	0.50	0.19	0.80	0.16
LSTM-A_5-MC [40]	0.50	0.42	0.37	0.33	0.53	0.20	0.91	0.18
LSTM-MC-global controller	**0.54**	**0.45**	0.38	**0.35**	0.54	**0.22**	0.96	**0.19**
LSTM-MC-IN_{static}	0.42	0.32	0.25	0.20	0.49	0.16	0.69	0.15
LSTM-MC-OUT_{static}	**0.54**	**0.45**	0.37	0.32	**0.55**	0.21	0.94	**0.19**
LSTM-MC-$IN_{dynamic}$	0.34	0.20	0.12	0.07	0.43	0.13	0.51	0.12
LSTM-MC-OUT	0.53	0.44	0.37	0.33	0.53	0.20	0.93	0.18
LSTM-MC-$OUT_{dynamic}$	**0.54**	**0.45**	**0.39**	**0.35**	**0.55**	**0.22**	**0.97**	**0.19**

Table 3: Performance comparison of different models evaluated on five standard metrics for image captioning. '-' means evaluation not conducted.

the encoder is 0.93 for the *Egyptian art multi-class classifier dataset* while the mean average accuracy for the multi-label classifier experimented on the *Egyptian art multi-label classifier dataset* is 0.59.

(4) Effects of the artwork type location in the decoder and encoding the relative importance of the artwork type and the other input coming along with it. We test the location influence on the captioning performance using models differentiated by 'IN' and 'OUT' in Table 3. Experiments demonstrate that the artwork type is helpful to improve the captioning performance as an input of either the LSTM or the fully connected layer predicting the caption words. We compare model LSTM-A_5-MC with model NIC to assess the improvement when the artwork type serves as input to the LSTM network. The comparison of model LSTM-MC-$OUT_{dynamic}$ and model NIC evidence an increase in performance when the artwork type serves as the input to the fully connected layer. Please refer to Table 3 to check the differences in results. The other interesting observation worth noting is that the effects of encoding the relative importance of the artwork type and the other input coming together with it are closely related to the artwork type location in the decoder. More specifically, when the artwork type vector serves as the input of the LSTM network together with the previously generated word in the decoder in the lower layer, ignoring the relative importance of the two inputs provides better guidance to train the model. This is indicated by the performance difference between model LSTM-A_5-MC that ignores the relative importance and model LSTM-MC-$IN_{dynamic}$ that encodes the relative importance. In contrast, encoding the relative importance when the artwork type vector is provided into the fully connected layer together with the hidden representation output of the LSTM network, yields slightly better results evidenced by the comparison of model LSTM-MC-$OUT_{dynamic}$ and model LSTM-MC-OUT. We model the relative importance of the artwork type and the other parallel input by using a weighted sum function. The weights to modulate the relative importance are automatically learned by the model during training. β_t and α_t are the learned weight values for the object type vector and the LSTM output vector at time step t in model LSTM-MC-*OUT_dynamic*. Their relative weight values in the first few steps are very small as we studied, indicating that the object type vector contributes little at the beginning of the caption generation process. But this value fluctuates along the generation process. The object type vector has a stronger contribution to the *ancient Chinese art image captioning dataset* than to the *ancient Egyptian art image captioning dataset.*

5.2 Qualitative Analysis

We further study the results qualitatively to assess how the artwork type improves captioning performance. Figure 3, 4, and 5 display four artwork images, their captions generated by different models, and their ground truth descriptions. The captions generated for Figure 3(a) and Figure 3(c) demonstrate that model NIC is not able to discriminate bowl and jar while the other three models managed to do so. Therefore, both visual attention and the textual artwork type can help to separate two alike images. Also, if we compare the captions generated for Figure 3(a) by model LSTM-A_5-MC and our model LSTM-MC-$OUT_{dynamic}$, it is easy to see that the latter

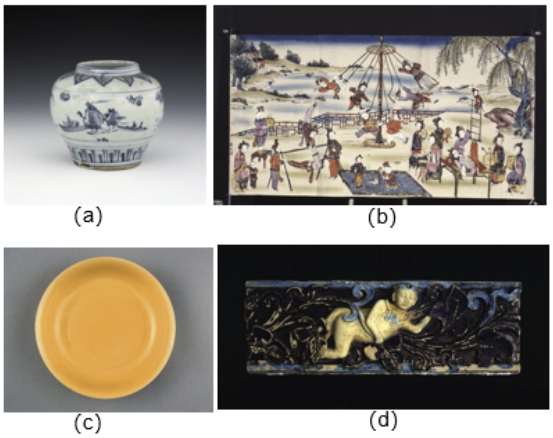

Figure 3: Four ancient Chinese artwork images.

Images	(a)	(b)	(c)	(d)
Artwork type	Jar	Print	Dish	Wall-tile
Model NIC	Porcelain bowl with underglaze blue decoration	Woodcut	There is a mark in underglaze blue on the base	There is an inscription on the base
Model SA	Porcelain jar with underglaze blue decoration	Woodcut	Porcelain dish with rounded sides	Made of blue glazed porcelain
Model LSTM-A_5-MC	This ovoid jar has a short neck with a thickened rim and a recessed base	Woodcut	Porcelain dish with rounded sides	Made of blue glazed porcelain
Model LSTM-MC-$OUT_{dynamic}$	Porcelain jar with underglaze blue decoration	Ink and colours on paper	Porcelain dish with rounded sides	Earthenware wall tile with moulded decoration and green glaze

Figure 4: Generated captions for the images in Figure 4 by model NIC, SA, LSTM-A_5-MC and LSTM-MC-$OUT_{dynamic}$.

(a)	Porcelain jar with underglaze blue decoration. This jar has a narrow raised neck, rounded body and tapering foot. On the unglazed base the spiral marks resulting from turning the jar on the potter's wheel are clearly visible. On the outside an old man is shown in a landscape on one side and another, walking with a servant carrying a 'qin' wrapped in a textile cloth, on the other side; they are separated by plants and rocks. Around the neck is a double ring with banana leaves forming a collar; stylized lappets, like vertical long and short stripes, surround the foot. The cobalt blue appears black where it has not been covered by the blue-green glaze. Originally the jar would have had a domed cover with lotus-bud finial
(b)	Woodcut. Recreation. Acrobats and jugglers and women playing musical instruments. Printed in ink, colour on paper
(c)	Porcelain dish has yellow enamel over thin plain felspathic glaze. There is an inscription on the base, which is glazed
(d)	Wall tile with relief moulded decoration and 'fahua'-palette glazes. This 'fahua'-palette tile is moulded and carved in relief with a celestial boy flying in a contorted pose through scrolling foliage. He is naked except for a torque necklace with three pierced beads. He holds the stem of lotus foliage in his right hand and raises the other hand. The tiny curl in the centre of his forehead indicates his youth. The turquoise glaze is fugitive

Figure 5: Ground-truth captions for the images in Figure 3.

generates a more professional sentence with terminology words 'Porcelain' and 'underglaze'. The captions generated for Figure 3(b) and Figure 3(d) are two instances where our model LSTM-MC-$OUT_{dynamic}$ performs better than all other models by putting the artwork type vector together with the hidden representation output in the decoder. In addition, the artwork type of the artifact in Figure 3(b) 'Print' does not occur in the generated caption 'Ink and colours on paper', indicating that the artwork type affects the caption content in a latent way. We also notice that for both the Egyptian art and the Chinese art datasets, the models generate completely unrelated captions for some images and these captions are mostly popular sentences in the training set. This is probably caused by the lack of enough training data. To tackle this problem, we already employed transfer learning from the visual side by using ResNet18 pre-trained on ImageNet, but textual transfer learning [23] leveraging the knowledge obtained from a model pre-trained on a large textual corpus is not yet explored. This might be an interesting research direction for future work. We also observe that the captions in the *Egyptian art image captioning* dataset are not as stereotypical as the captions in the *ancient Chinese art image captioning dataset* which mainly infer the artwork name, the material indicated by cue phrases (e.g., 'made of') and some fine-grained image details. The Egyptian dataset has large intra-class variance in terms of image patterns, and its number of unique descriptive words is twice the number of the Chinese art set, making the task of image captioning of Egyptian artworks more difficult. These factors prohibit the clear guidance of the decoder in the proposed model LSTM-MC-$OUT_{dynamic}$ and is probably the reason why the proposed model behaves differently on the two captioning datasets.

6 CONCLUSION AND FUTURE WORK

In this paper, we have introduced an artwork type enriched image captioning model for ancient artworks and have implemented several variants of it. The best model explicitly models the relative importance of the artwork type vector and the hidden representation in the LSTM decoding process where the modulating parameters encoding the relative importance in this model are automatically learned by the model during training. This model achieves promising results on two ancient artwork image captioning datasets. We have also adapted three existing captioning models originally built for the captioning of natural images in order to generate descriptions of cultural heritage images. Finally, we have compared the performance of all the models and give a comprehensive quantitative and qualitative analysis of this task. In future research, we will explore fine-grained cultural image/subimage [13, 17] annotation.

7 ACKNOWLEDGMENTS

This work is funded by the KU Leuven BOF/IF/RUN/2015. We additionally thank Katrien Laenen and the anonymous reviewers for their useful comments.

REFERENCES

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic propositional image caption evaluation. In *Proceedings of European Conference on Computer Vision*. Springer, 382–398.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6077–6086.

[3] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 65–72.

[4] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research* 3, Feb (2003), 1137–1155.

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[6] Qi Dong, Xiatian Zhu, and Shaogang Gong. 2019. Single-label multi-class image classification by deep logistic regression. *CoRR* abs/1811.08400 (2019).

[7] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. 2013. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894* (2013).

[8] Nicolas Gonthier, Yann Gousseau, Said Ladjal, and Olivier Bonfait. 2018. Weakly supervised object detection in artworks. In *Proceedings of European Conference on Computer Vision*. Springer, 692–709.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[10] Nicholas J. Higham. 1996. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

[11] Md Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)* 51, 6 (2019), 118.

[12] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. 2014. DenseNet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869* (2014).

[13] Andrej Karpathy, Armand Joulin, and Fei-Fei Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1889–1897.

[14] Nikhil Ketkar. 2017. Introduction to Pytorch. In *Deep learning with python*. Springer, 195–208.

[15] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

[17] Katrien Laenen, Susana Zoghbi, and Marie-Francine Moens. 2018. Web search of fashion items with multimodal querying. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, 342–350.

[18] Séamus Lawless, Maristella Agosti, Paul Clough, and Owen Conlan. 2013. Exploration, navigation and retrieval of information in cultural heritage: ENRICH 2013. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1136–1136.

[19] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of Text Summarization Branches Out*. 74–81.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of European Conference on Computer Vision*. Springer, 740–755.

[21] Hui Mao, Ming Cheung, and James She. 2017. DeepArt: Learning joint representations of visual arts. In *Proceedings of the 25th ACM International Conference on Multimedia*. ACM, 1183–1191.

[22] Thomas Mensink and Jan Van Gemert. 2014. The Rijksmuseum challenge: Museum-centered visual recognition. In *Proceedings of International Conference on Multimedia Retrieval*. ACM, 451.

[23] Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. 2017. Question answering through transfer learning from large fine-grained supervision data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 510–517.

[24] Pauline C Ng and Steven Henikoff. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research* 31, 13 (2003), 3812–3814.

[25] Abraham Montoya Obeso, Mireya S García Vázquez, Alejandro A Ramírez Acosta, and Jenny Benois-Pineau. 2017. Connoisseur: Classification of styles of mexican architectural heritage with deep learning and visual attention prediction. In *Proceedings of the 15th International Workshop on Content-based Multimedia Indexing*. ACM, 16.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 311–318.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. 91–99.

[28] Shurong Sheng, Katrien Laenen, and Marie-Francine Moens. 2019. Can image captioning help passage retrieval in multimodal question answering? In *Proceedings of European Conference on Information Retrieval*. Springer, 94–101.

[29] Shurong Sheng, Aparna Nurani Venkitasubramanian, and Marie-Francine Moens. 2018. A Markov network based passage retrieval method for multimodal question answering in the cultural heritage domain. In *Proceedings of International Conference on Multimedia Modeling*. Springer, 3–15.

[30] Gjorgji Strezoski and Marcel Worring. 2017. OmniArt: Multi-task deep learning for artistic data analysis. *arXiv preprint arXiv:1708.00684* (2017).

[31] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. In *Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association*. 194–197.

[32] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[33] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4566–4575.

[34] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[35] James Z Wang, Kurt Grieb, Ya Zhang, Ching-chih Chen, Yixin Chen, and Jia Li. 2006. Machine annotation and retrieval for digital imagery of historical materials. *International Journal on Digital Libraries* 6, 1 (2006), 18–29.

[36] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of International Conference on Machine Learning*. 2048–2057.

[37] Lei Xu, Albert Merono-Penuela, Zhisheng Huang, and Frank van Harmelen. 2017. An ontology model for narrative image annotation in the field of cultural heritage. In *Proceedings of the 2nd Workshop on Humanities in the Semantic web (WHiSe)*. 15–26.

[38] Lei Xu and Xiaoguang Wang. 2015. Semantic description of cultural digital images: Using a hierarchical model and controlled vocabulary. *D-Lib magazine* 21, 5/6 (2015).

[39] Heekyung Yang and Kyungha Min. 2019. Classification of basic artistic media based on a deep convolutional approach. *The Visual Computer* (2019), 1–20.

[40] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*. 4894–4902.

[41] AM Yasser, Kathy Clawson, and Chris Bowerman. 2017. Saving cultural heritage with digital make-believe: Machine learning and digital techniques to the rescue. In *Proceedings of the 31st British Computer Society Human Computer Interaction Conference*. BCS Learning & Development Ltd., 97.