A CNN-LSTM FRAMEWORK FOR AUTHORSHIP CLASSIFICATION OF PAINTINGS

Kevin Alfianto Jangtjik[1], *Trang-Thi Ho*[1], *Mei-Chen Yeh*[2], *Kai-Lung Hua*[1]

[1]Dept. of CSIE, National Taiwan University of Science and Technology
[2]Dept. of CSIE, National Taiwan Normal University

ABSTRACT

The authenticity of digital painting image is an urgent demand in the field of art. Yet, determining the authorship of a certain painting is a challenging task due to two reasons: (1) various artists might share similar painting styles; and (2) an artist could create different styles. In this paper, we present a novel method for authorship classification of paintings based on a CNN-LSTM framework. First, a multiscale pyramid is constructed from a painting image. Second, a CNN-LSTM model is learned and it returns possibly multiple labels for one image. To aggregate the final classification result, an adaptive fusion method is employed. Experimental results show that the proposed method has superior classification performance compared with the state-of-the-art techniques.

***Index Terms*—** Digital image classification, multiscale pyramid representation, convolutional neural network, long short-term memory networks.

1. INTRODUCTION

Nowadays, the authorship is increasingly concerned in many fields because the internet facilitates the collection and utilization of data by using just a few clicks. However, determining the authorship of an artwork is not easy for a general user. Unlike paintings in a museum collection, the digital painting images often lack detailed information about the content. Thus, the authenticity of digital paintings typically can only be identified by well-experienced experts. Considering a large number of unattributed digitized art images on the internet, it is in urgent demand to develop an automatic algorithm to classify, analyze, and understand the digital painting images. One of the most essential information about paintings is authorship, by which painting collections could be indexed and unknown paintings could be identified.

In recent years, researchers have started to develop computational methods to study painting authorship. For example, Hanchao and Shannon [1] proposed a visual stylometry method for a binary classification task (van gogh or non van gogh), while this work deals with a more challenging task of 13 artist classes. Besides, Meijun Sun [2] presented a stroke based sparse hybrid convolutional neural networks (CNNs) method for author classification of Chinese ink-wash paintings. As revealed in the title, the study was limited to Chinese paintings. Another work on painting classification is proposed by Zhao et al [3]. They categorized painting images into eight emotional groups (fear, sadness, disgust, anger, amusement, awe, excitement, and contentment). Several features such as balance, harmony, variety, contrast, movement, and gradation were designed. Four classifiers (KNN, softmax regression, global weight, and shared sparse learning) were investigated in [3]. Not long after that, K. Peng and T. Chen reported a cross-layer CNN-based method [4]. Inspired by [4], [5] decomposed an painting image into a hierarchical representation from which several image patches were extracted and used to predict the image class.

Similar to [5], in this work, we view an image a multilayer pyramid containing the whole image and a set of local image patches. Using such a representation provides more training data and considers both globally and locally the information contained in one image. However, unlike [5] that considers image patches independently, we characterize the label correlations among local image patches using long short-term memory (LSTM). We propose a CNN-LSTM model (detailed in section 3) which returns possibly multiple labels for one image. Finally, an adaptive fusion scheme is applied to obtain the final decision result.

2. BACKGROUND

Since we aim to characterize the label correlations among image patches, we employ LSTM neurons as our recurrent neuron, which has been demonstrated to be a powerful model of long-term dependency.

2.1. Recurrent neural network

Recurrent neural network (RNN) is a class of neural network that maintains internal hidden states to model inputs with dependency through directed cyclic connections between its units. For example, given a sentence and the goal is to predict the word following that sentence. It is apparent that the predicted word must have semantic relationship with those occur in the sentence. RNN shares the parameters for every element of a sequence and generates outputs that depend on current and previous inputs. It uses hidden states to hold information

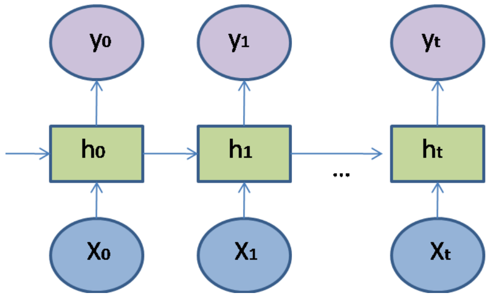


Fig. 1. unrolled RNN

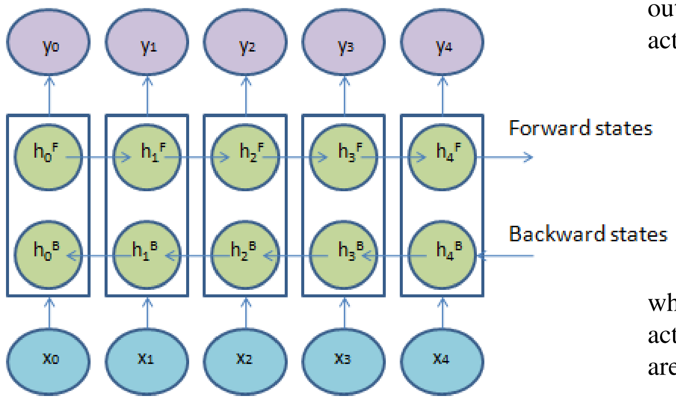


Fig. 2. An unrolled bidirectional neural network.

on previous inputs. Figure 1 shows a RNN architecture unrolled into a full network, in which x_t, y_t, h_t represents the input, output and hidden state at the time t respectively.

As shown in Fig. 1, the hidden state h_t receives information from the previous hidden state as well as current input, acting like the memory of network that keeps information about what previously computed. The parameters involved in a RNN are described as follows:

$$h_t = \tanh\left(W_{xh}x_t + W_{hh}h_{t-1} + b_h\right), \quad (1)$$

$$y_t = W_{hy}h_t + b_y, \quad (2)$$

where W_{xh}, W_{hh}, W_{hy} represent the hidden-to-hidden layer, input-to-hidden layer, and output-to-hidden layer weight matrices. b_h and b_y are the biases of the hidden and output layer.

2.2. Bidirectional RNN

A bidirectional RNN [6] extends RNN in that it computes the ouputs based on not only previous inputs but also future inputs. This is achieved by processing the inputs in two directions—forward and backward. Bidirectional RNNs have shown great performance on various recognition tasks such as handwriting [7] and speech recognition [8]. Figure 2 shows a bidirectional RNN. The outputs of a bidirectional RNN is calculated as follows:

$$h_t^F = \tanh\left(W_{xh}^F x_t + W_{hh}^F h_{t-1}^F + b_h^F\right), \quad (3)$$

$$h_t^B = \tanh\left(W_{xh}^B x_t + W_{hh}^B h_{t-1}^B + b_h^B\right), \quad (4)$$

$$Y_t = W_{hy}^F h_t^F + W_{hy}^B h_t^B + b_y, \quad (5)$$

where h_t^F and h_t^B are forward and backward hidden layers.

2.3. Long short term memory network

LSTM network can learn long-term dependency in a sequence. It extends RNN by adding three gates to a RNN neuron: a forget gate to control whether to forget the current state; an input gate to control if the input should be stored; an output gate to control whether to ouput the state. The LSTM activations are calculated as follows:

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right), \quad (6)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right), \quad (7)$$

$$c_t = f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right), \quad (8)$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right), \quad (9)$$

$$c_t h_t = o_t \tanh\left(c_t\right), \quad (10)$$

where i_t, f_t, c_t and o_t denote the input gate, forget gate, cell activation vectors and output gate at time t. b_i, b_f, b_o, and b_c are the biases of the gates.

3. APPROACH

We view an image a multi-layer pyramid containing the whole image and a set of local image patches. Using such a representation provides more training data and considers both globally and locally the information contained in one image. To characterize the label correlations among local image patches, we propose to use LSTM to train the classification model. The proposed CNN-LSTM model returns possibly multiple labels for one image. Therefore, an adaptive fusion scheme proposed in our previous study [5], is applied to obtain the final decision result. Figure 4 shows the system architecture. Technical details are described in the next subsections.

3.1. Training the CNN models

We construct a multi-scale pyramid for an image. The goal is to augment data to train the deep-architecture-based classification models. As depicted in Fig. 3, we subdivide the image at three different layers of resolution. For each layer, we train a CNN model. Specifically, all the training images in the first layer are fed into the ImageNet-trained CNN model [9] to fine-tune the weights of the pre-trained network by continuing the back-propagation. Similarly, the higher layer models are also fine-tuned with the same technique. Note that for higher layers, instead of one patch for a training image where the first layer applies, we now have 4 and 16 patches for each training image for the second and third layer respectively.

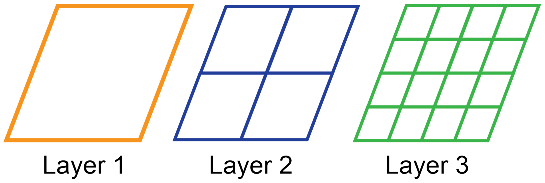


Fig. 3. Multi-scale pyramid representation.

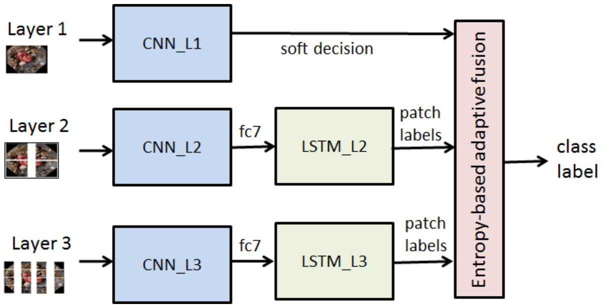


Fig. 4. Architecture of the proposed CNN-LSTM model for artist-based image classification.

3.2. Training the LSTM models

The multi-scale CNNs are effective for classifying painting images of different artists [5]. However, the CNN models for the second and third layer view each image patch indenpendently, without the consideration of the correlations among them. We propose to model the relationships between the image patches using LSTM. As shown in Fig. 4, the layer-2 (or layer-3) CNN model is used to extract discriminant features (fc7) from an image patch, which are consecutively fed into a LSTM model to *jointly* predict the class labels for the image patches of a layer. We use the column-wise raster scan approach to convert an image into a 1-D image patch sequence. In the implementation, the LSTM models are trained by using the Theano library [10].

3.3. Adaptive fusion scheme

Given a test image, the proposed CNN-LSTM model returns a label for each image patch, as illustrated in Fig. 5. Specifically, we obtain 21 labels for each test image (i.e., one for the first layer, 4 for the second layer and 16 for the third layer). Our strategy for aggregating these results is to rely more on the decisions of the layers performing relatively well. We employ an adaptive fusion method [5] that uses class entropy to determine the decision quality of each layer and to combine the results. Let p_i^j denote the probability of being identified as class i at layer j, we compute the corresponding entropy H^j of layer j as follows:

$$H^j = -\sum p_i^j \log\left(p_i^j\right). \tag{11}$$

Next, the weight of a layer W^j is defined as the inverse value of H^j, because a model returning a diverse class distribition behaves less trusty in comparison to the other models.

$$W^j = 1/H^j. \tag{12}$$

Hence the aggregated multi-scale class probability of calss i, denoted by p_i, is:

$$p_i = \frac{\sum_i W^j \times p_i^j}{\sum_i \sum_j W^j \times p_i^j}. \tag{13}$$

The final classification result is obtained by selecting the class with the maximal probability:

$$i^* = \underset{i}{\operatorname{argmax}} = \{p_i\}. \tag{14}$$

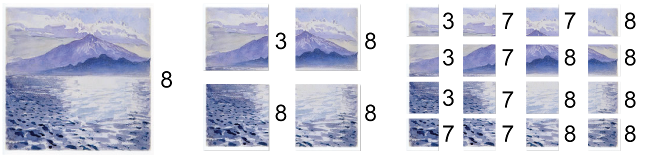


Fig. 5. Classification results (shown at the right-hand side of each sub-region patch) for layer 1 (left), layer 2 (middle), layer 3 (right).

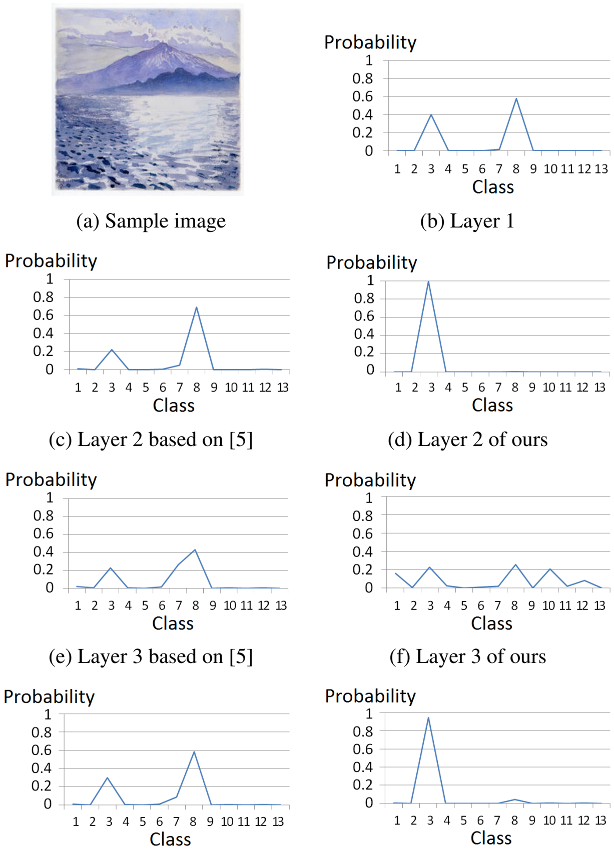
4. EXPERIMENTS

In this section, we evaluate the performance of the proposed classification algorithm compared to two state-of-the-art methods [5][9]. In our experiments, the dataset of [5]—containing 1,300 digital painting images from 13 artists (each artist has 100 images)—is utilized. For each artist, we randomly select 80 images as training data and 20 images as testing data.

4.1. Improvement based on LSTM model

Figure 6 shows the probability distribution of each layer obtained from [5] and from the proposed method. The ground truth label of this input image is class 3.

Figure 6(b) shows the probability distribution for layer 1 for both methods. The proposed method and [5] share the same distribution for this layer since in layer 1, there is only one single image, hence no correlation is considered. Observing from Fig. 6(c)(d), it is obvious the proposed LSTM model is effective and returns a confident result. From Fig. 6(e)(f), it could be observed that both methods do not show a confident result, which potentially imply for this scale, various artists would fit the corresponding painting style. Finally, the entropy-based adaptive algorithm is employed to fuse the final result as shown in Fig. 6(g)(h), we could observe that the proposed method returns a precisely correct result with high confidence while the compared method returns a relatively flat probability distribution which would even select incorrect class (8) as the final result based on the maximal probability criterion.



(a) Sample image

(b) Layer 1

(c) Layer 2 based on [5]

(d) Layer 2 of ours

(e) Layer 3 based on [5]

(f) Layer 3 of ours

(g) Fusion result of [5]

(h) Fusion result of ours

Fig. 6. Example Improvement using RNN

Table 1. Performance evaluation of classification result

Top-1			
	Precision	Recall	F Score
Pre-trained model [9]	65.26%	66.15%	65.7%
Multi-scale pyramid [5]	71.22%	72.30%	71.7%
Ours	73.71%	74.23%	73.97%

4.2. Result and Discussion

Table 1 shows the classification performance of several methods. The proposed CNN-LSTM model achieves the best performance. The value of precision, recall, and F-Score of the proposed method are 73.71%, 74.23%, and 73.97%, respectively. These results are superior to two state-of-the-art methods [5][9].

The confusion matrix of the proposed approach is shown in Fig. 8. The blue boxes show the correct classification result. The confusion matrix shows our performance of class 13 needs to be significantly improved because the testing images of class 13 are likely to be misclassified as class 11. From our investigation, we found interesting connection between artist-11 (Peter Paul Rubens) and artist-13 (Tiziano Vecellio). Though they are artists from different times and places, during 1600 - 1608, Peter Paul Rubens (class 11) studied classical art by coping works of Tiziano Vecellio (class 13) and others. Therefore their paintings show certain similar drawing skills and are difficult to be differentiated. To show the similarity between these two artists, Fig. 7 demonstrates one example pair of the paintings, from class 11 and class 13, that are misclassified.

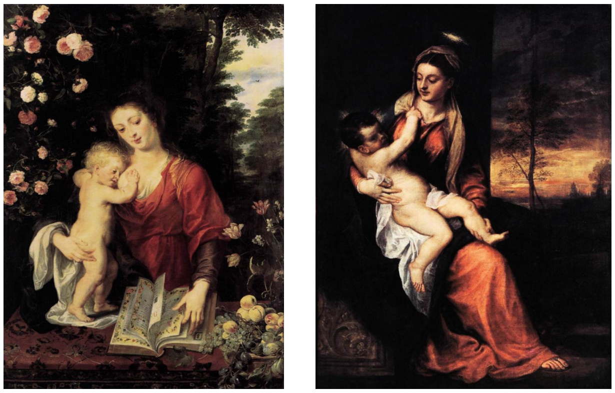
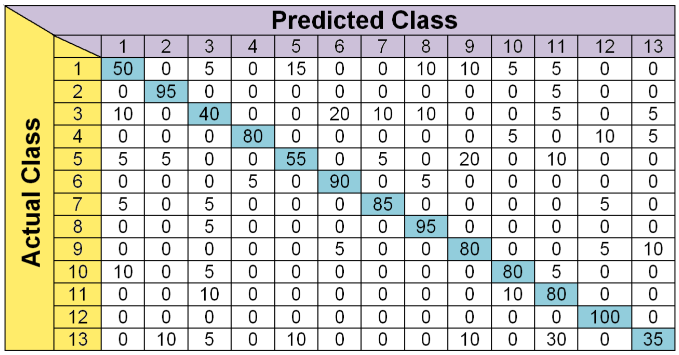


Fig. 7. An example image pair of class 11 (left side) and class 13 (right side).

5. CONCLUSION

In this paper, we presented a novel method for authorship classification of paintings through a CNN-LSTM framework. The proposed method constructed a multi-layer pyramid for the input image. Then a long short-term memory approach was utilized to characterize the label correlations among local image patches while learning our recurrent neural network models. Finally, a weighted fusion scheme, which adaptively combined the models based on their estimated posteriori probability, was employed to compute the final decision result. Experimental results showed that our approach achieved a promising authorship classification performance and outperformed two state-of-the-art techniques.



Actual Class \ Predicted Class	1	2	3	4	5	6	7	8	9	10	11	12	13
1	50	0	5	0	15	0	0	10	10	5	5	0	0
2	0	95	0	0	0	0	0	0	0	0	5	0	0
3	10	0	40	0	0	20	10	10	0	0	5	0	5
4	0	0	0	80	0	0	0	0	0	5	0	10	5
5	5	5	0	0	55	0	5	0	20	0	10	0	0
6	0	0	0	5	0	90	0	5	0	0	0	0	0
7	5	0	5	0	0	0	85	0	0	0	0	5	0
8	0	0	5	0	0	0	0	95	0	0	0	0	0
9	0	0	0	0	0	5	0	0	80	0	0	5	10
10	10	0	5	0	0	0	0	0	0	80	5	0	0
11	0	0	10	0	0	0	0	0	0	10	80	0	0
12	0	0	0	0	0	0	0	0	0	0	0	100	0
13	0	10	5	0	10	0	0	0	10	0	30	0	35

Fig. 8. Confusion matrix of the proposed approach.

6. REFERENCES

[1] H. Qi and S. Hughes, "A new method for visual stylometry on impressionist paintings," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2036–2039, IEEE, 2011.

[2] M. Sun, D. Zhang, J. Ren, Z. Wang, and J. S. Jin, "Brushstroke-based sparse hybrid convolutional neural networks for author classification of chinese ink-wash paintings," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 626–630, IEEE, 2015.

[3] S. Zhao, H. Yao, X. Jiang, and X. Sun, "Predicting discrete probability distribution of image emotions," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 2459–2463, IEEE, 2015.

[4] K.-C. Peng and T. Chen, "Cross-layer features in convolutional neural networks for generic classification tasks," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 3057–3061, IEEE, 2015.

[5] K. A. Jangtjik, M.-C. Yeh, and K.-L. Hua, "Artist-based classification via deep learning with multi-scale weighted pooling," in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 635–639, ACM, 2016.

[6] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[7] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

[8] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 273–278, IEEE, 2013.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[10] P. LamblinP, "Theano: A cpu and gpu math expression compiler," *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 2010.

[11] *PaintingDb fastest growing art gallery in the web*, 2015.