Dependency Exploitation: A Unified CNN-RNN Approach for Visual Emotion Recognition

Xinge Zhu[1], Liang Li[2*], Weigang Zhang[1,3], Tianrong Rao[4], Min Xu[4], Qingming Huang[1,2*], Dong Xu[5]
[1]University of Chinese Academy of Sciences, China
[2]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, China
[3]Harbin Institute of Technology, Weihai, China
[4]University of Technology, Sydney, Australia
[5]University of Sydney, Australia

Abstract

Visual emotion recognition aims to associate images with appropriate emotions. There are different visual stimuli that can affect human emotion from low-level to high-level, such as color, texture, part, object, etc. However, most existing methods treat different levels of features as independent entity without having effective method for feature fusion. In this paper, we propose a unified CNN-RNN model to predict the emotion based on the fused features from different levels by exploiting the dependency among them. Our proposed architecture leverages convolutional neural network (CNN) with multiple layers to extract different levels of features within a multi-task learning framework, in which two related loss functions are introduced to learn the feature representation. Considering the dependencies within the low-level and high-level features, a bidirectional recurrent neural network (RNN) is proposed to integrate the learned features from different layers in the CNN model. Extensive experiments on both Internet images and art photo datasets demonstrate that our method outperforms the state-of-the-art methods with at least 7% performance improvement.

1 Introduction

Emotion recognition is a crucial part of visual understanding as it can benefit a broad range of applications, such as recommendation, advertisement and option mining. Recently, predicting the emotion from visual content has attracted increasing attention [Machajdik and Hanbury, 2010; Sartori *et al.*, 2015; Rao *et al.*, 2016b] . However, unlike object recognition, understanding emotions from images often goes beyond the recognition of individual objects as emotion recognition needs to bridge the “affective gap” between low-level visual features and high-level emotional reactions. Therefore, emotion recognition is more difficult, due to the complexity and subjectivity of emotions [Machajdik and Hanbury, 2010].

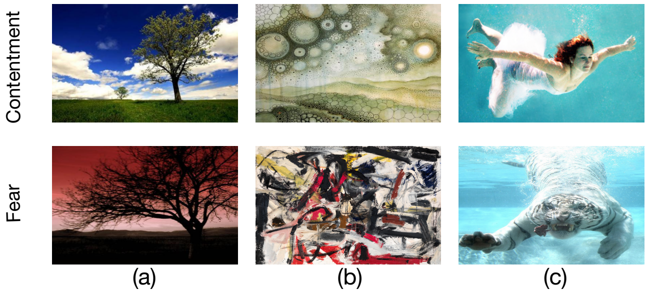


Figure 1: The same emotion can be evoked from different emotion stimuli. The top three images are from “Contentment”, and the bottom images are from “Fear”. We can find image emotion is related to many factors, such as low-level features (e.g. color (a)), middle-level features, (e.g. composition (b)), and high-level features, (e.g. semantic content (c)). Best viewed in color.

Many studies have already proven that human's emotions are related to many factors from low-level to high-level [Lang, 1979]. As shown in Fig. 1, the same emotion can be evoked from different visual stimuli. Low-level visual features, such as color (Fig.1 (a)), and shape, were first used to classify emotions [Wang and He, 2008; Kang, 2003]. Then, some studies utilized middle-level visual features, such as composition (Fig.1 (b)), texture and emphasis, for emotion recognition [Joshi *et al.*, 2011; Sartori *et al.*, 2015]. Moreover, the work [Machajdik and Hanbury, 2010] indicated that high-level semantic content (Fig.1 (c)) of the image, has significant impact for recognizing emotions from pictures. However, most existing methods treat different levels of features as independent cues, which degrades their performances.

As deep convolutional neural networks (CNN) have shown great success in semantic content recognition [Szegedy *et al.*, 2015; He *et al.*, 2015], more and more works started to employ the CNNs for emotion recognition [You *et al.*, 2016]. In [Rao *et al.*, 2016b], they designed three networks to represent different levels of features and integrated all three levels of features to classify the emotions. In fact, a CNN model itself can support recognition at several levels of abstraction (e.g., colors, edges, and objects). Many works have been proposed to understand the representations learned by CNNs and justify the assumption in [Zeiler and Fergus, 2014; Zhou *et al.*, 2014]. In which, they showed that different lay-

[*]corresponding author:{liang.li;qingming.huang}@vipl.ict.ac.cn

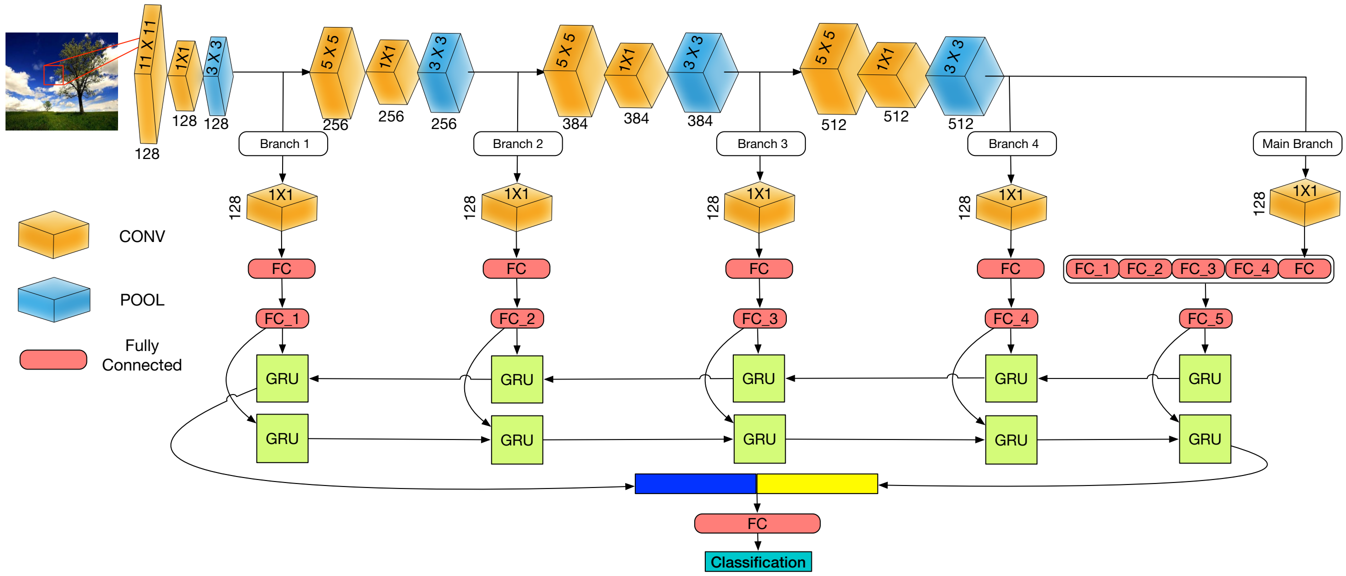


Figure 2: Our unified CNN-RNN framework for visual emotion recognition. We first extract different levels of features from multiple branches in the CNN model, which include low-level features (e.g. color, edge), middle-level features (e.g. texture) and high-level features (e.g. part, object). Then different levels of features flow into our newly proposed Bidirectional GRU model to integrate these features and exploit their dependencies. Two features generated from our Bi-GRU model are concatenated as the final features to predict the emotion from images. (Best viewed in color.)

ers from each network prefer different parts of the image. For instance, the earlier layers prefer low-level features, such as color, line, and shape, while the later layers tend to be attracted by the object parts and semantic content.

How to integrate different levels of features is crucial in some existing methods as these methods aim to exploit multiple levels of representations. Many previous methods performed feature fusion by using dimensionality reduction on the feature vectors [Machajdik and Hanbury, 2010], or using the *max* and *avg* aggregation function [Rao *et al.*, 2016b]. However, existing fusion methods do not consider the dependencies among different levels of features. We assume there is a strong correlation among different levels of features. For example, for middle-level features, such as textures, it is composed of low-level features, such as lines, and meanwhile it leads to high-level features, such as the parts of object. Such dependency among different levels of features benefits visual emotion recognition because we need to consider different types of emotion stimuli in this task. To this end, we propose a new bidirectional model for feature fusion by exploiting the dependency among different levels of features. Experimental results justify and demonstrate the effectiveness of our newly proposed RNN method for feature fusion.

Specifically, we propose a unified CNN-RNN framework for visual emotion recognition, which effectively learns different levels of features and integrates them by exploring the dependencies. As shown in Fig. 2, the total framework consists of two parts, i.e., feature extraction and feature fusion. The image features are extracted from multiple branches in CNN, which can represent different levels of features from the local view to global view. When training the CNN, we introduce the multi-task losses in order to learn more discriminative representation, in which the classification loss aims to classify the emotion while the contrastive loss aims to satisfy the contrastive objective. Considering the dependencies among different levels of features, a bidirectional RNN model consisting of gated recurrent unit is proposed to capture this relationship and integrate different levels of features together. Finally, we concatenate the fused features extracted from our RNN model to predict the emotion.

The main contributions are summarized as follows.

- We propose a CNN model with branches to extract different levels of features, in which we introduce the multi-task losses to satisfy the classification objective and contrastive objective simultaneously.
- To our best knowledge, it is the first time to propose the the bidirectional RNN model for feature fusion by exploring the dependency among different levels of features in this task. Extensive experiments demonstrate the effectiveness of our bidirectional RNN model in combination with the CNN model.
- Experiments conducted on Internet image and Art photo datasets demonstrate that our proposed method achieves much better performance when compared with the state-of-the-art methods.

The rest of this paper is organized as follows. We present the related work in Section 2 and introduce our proposed CNN-RNN model in Section 3. Experimental results are reported in Section 4, followed by conclusion in Section 5.

2 Related Work

Previous researches on visual emotion recognition can be roughly grouped into two categories, i.e., single-level feature-based approaches and multi-level feature-based approaches.

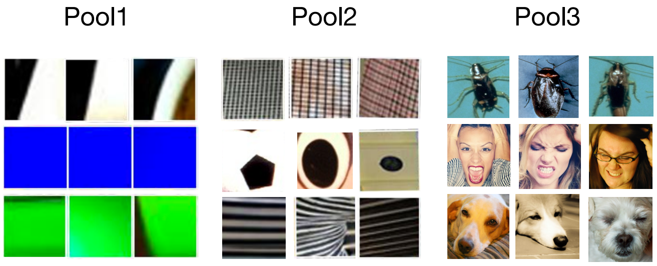


Figure 3: The receptive fields of 3 samples of pool1, pool2, and pool3 respectively. These image patches are the top activation regions inside the receptive fields.

Many previous works utilized single-level emotion-related features including Gabor and Wiccest features [Yanulevskaya *et al.*, 2008], color-based features [Solli and Lenz, 2009], global and local RGB histogram [Siersdorfer *et al.*, 2010] and so on. Due to the complexity of emotions, more and more works were proposed to exploit multiple levels of features. In [Machajdik and Hanbury, 2010], inspired by art and psychology theory, authors designed a series of hand-crafted features including color variance, balance, composition, and semantic content. The work in [Zhao *et al.*, 2014] investigated the concept of principles-of-art and designed robust and invariant visual features according to these principles. CNN based features have also been employed in this task. In [Rao *et al.*, 2016b], authors utilized three different networks to capture different levels of visual features with high expense of parameters. However, the previous works did not explicitly exploit the dependency between low-level features and high-level features and most of these works lack of effective fusion methods for different levels of features.

Recurrent neural networks can effectively model the long-term dependency in sequential data. It has been widely applied in many tasks including machine translation [Sutskever *et al.*, 2014], sequential modeling [Chung *et al.*, 2015], and so on. In this work, we show that the RNN can also exploit the relation between low-level features and high-level features.

In multi-task learning, multiple related tasks are solved at the same time by exploiting commonalities and differences across tasks [Caruana, 1998]. Multi-task learning performs well because what is learned for each task can help other tasks be learned better. It has been successfully applied in machine translation [Luong *et al.*, 2015] and visual recognition [Donahue *et al.*, 2014]. When training the CNN model, we also employ the multi-task learning to predict the emotion and learn the contrastive objective simultaneously.

3 CNN-RNN for Visual Emotion Recognition

As shown in Fig. 2, our proposed method mainly consists of two components, CNN feature extractor and Bidirectional-GRU (Bi-GRU) feature fusion. The input image is first fed to the CNN model to extract multiple levels of features at different branches. These features from different layers represent different parts of images, such as line, color, texture, and object, which characterize different levels of features from the local view to global view. Our Bi-GRU model aims to integrate the different levels of features by exploiting the dependency between low-level and high-level features. The integrated features from our Bi-GRU model are concatenated for visual emotion classification.

Layer	conv11	conv12	pool1	Branch1 conv	conv21	conv22	pool2	Branch2 conv	conv31	conv32	pool3	Branch3 conv	conv41	conv42	pool4	Branch4 conv
Channels	128	128	128	128	256	256	256	128	384	384	384	128	512	512	512	128
Kernel Stride	11x11 4	1x1 1	3x3 2	1x1 1	5x5 1	1x1 1	3x3 2	1x1 1	5x5 1	1x1 1	3x3 2	1x1 1	5x5 1	1x1 1	3x3 2	1x1 1
Receptive Field	11	11	19	19	51	51	67	67	131	131	163	163	291	291	355	355

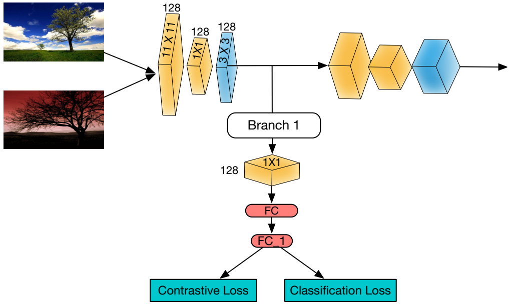


Figure 4: Top table: the parameters of the main architecture of CNN. Bottom panel: the pipeline for training CNN with multi-task learning, the first branch is taken as an example.

3.1 CNN with Multiple Branches

Compared with the top layer, the bottom layer and the intermediate layer can provide complementary information, including low-level features and middle-level features. Based on the additional information, many methods achieved promising performance [Yue-Hei Ng *et al.*, 2015; Li *et al.*, 2012]. As shown in Fig. 2, we design four branches and one main branch to learn different levels of features from the local view to global view. The detailed architecture of our CNN is summarized in the top table of Fig. 4. Motivated by the auxiliary classifier in GoogleNet [Szegedy *et al.*, 2015], we generate different branches from each pooling layer by inserting 1×1 convolution layer and fully connected layer. The 1×1 convolution layer with 128 filters is used for dimension reduction and rectified linear activation. Furthermore, the main branch integrates features from different layers to provide the global representation.

To intuitively illustrate the features extracted from different layers, we visualize the receptive fields for 3 channels from different pooling layers in Fig. 3 according to [Zhou *et al.*, 2014]. We observe that, as the layers go deeper, the receptive fields with top activation are becoming more semantically meaningful and more discriminative. There are five branches from the pooling layers and different levels of features extracted from multiple branches will be fed into our new Bi-GRU model as the inputs of different time steps.

3.2 Multi-Stage Multi-Task Learning for CNN

Inspired by the training method in [Liu *et al.*, 2015], our model is trained in an incremental manner. Given a limited number of training images, the multi-stage training strategy can achieve better performance because it introduces much less parameters at each stage. The details for the training process are described in Section 4.1.

Considering some emotions are with subtle differences and it is thus difficult to use the traditional loss function to distinguish some images from those similar emotion categories. In this work, we propose to use an additional contrastive loss function to enforce the feature vectors extracted from each pair of images from the same category to be close to each other, and enforce the feature vectors extracted from each pair of images from different categories to be far away with each other. With this new contrastive loss function, we can better classify the images from those similar emotion categories.

As shown in Fig. 4, when training the CNN model, a pair of images a and b are fed to the CNN model to extract the visual features $\mathbf{V_a}$ and $\mathbf{V_b}$. We predict the emotion category by using the softmax function for each image, which is defined as follows:

$$P(z = c|\mathbf{V}) = \frac{\exp(\mathbf{W_c V})}{\sum_{\mathbf{k}} \exp(\mathbf{W_k V})} \tag{1}$$

where z is the emotion of the image and $\mathbf{W}$ is the weight matrix with $\mathbf{W_c}$ and $\mathbf{W_k}$ representing its cth and kth column. We employ the negative log-likelihood (NLL) function to define the classification loss:

$$L_{cls}(\mathbf{V}) = -\log(P(z = c|\mathbf{V})) \tag{2}$$

Next, the contrastive objective is learned from the new max-margin loss. If a pair of images are from the same category, we use the L2 norm to penalize this pair of images with large distance.

$$L(\mathbf{V_a}, \mathbf{V_b})^+ = \|\mathbf{V_a} - \mathbf{V_b}\|_2^2 \tag{3}$$

If a pair of images are from two different categories, we set a margin μ to penalize the negative pair which is closer than the margin.

$$L(\mathbf{V_a}, \mathbf{V_b})^- = \max\{0, \mu - \|\mathbf{V_a} - \mathbf{V_b}\|_2^2\} \tag{4}$$

Given a pair of images, the total loss can be defined as the sum of classification loss and contrastive loss.

$$\begin{aligned} L = L_{cls}(\mathbf{V_a}) + L_{cls}(\mathbf{V_b}) + [l = 1]L(\mathbf{V_a}, \mathbf{V_b})^+ \\ + [l = -1]L(\mathbf{V_a}, \mathbf{V_b})^- \end{aligned} \tag{5}$$

where $l = 1$ if a pair of images are from the same category while $l = -1$ if they are from two different categories. The total loss can be optimized in an end-to-end manner using back propagation. After we finish training the CNN model, we discard the total loss function and directly apply the features extracted from the fully connected layer as the inputs to our new Bi-GRU model.

3.3 Bi-GRU

In order to exploit the dependency among different levels of features, we treat the features from the lower level to the higher level and from the higher level to the lower level as two sets of sequential data, and propose a new RNN based approach with bidirectional connections to better model such dependencies. Specifically, we propose a new bidirectional GRU model to integrate the features from different levels as our comprehensive experiments on the benchmark datasets show that the GRU method can achieve better performance than the LSTM method for visual emotion recognition.

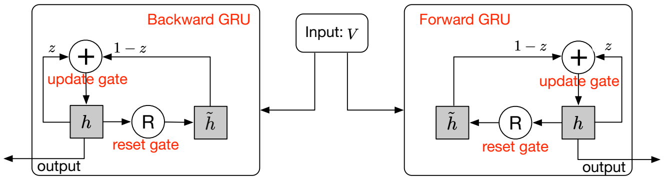


Figure 5: The information flow of bidirectional gate recurrent unit. The bidirectional GRU consists of a forward GRU (right) and a backward GRU (left).

We use $\mathbf{V_t}$ to represent the visual features extracted from the branch at the time step t. Then the total pipeline in a GRU (illustrated in Fig. 5) at the time step t can be presented as follows:

$$\begin{aligned} r_t &= \sigma(\mathrm{W}_{vr}\mathbf{V_t} + \mathrm{W}_{hr}h_{t-1} + b_r) && (6) \\ z_t &= \sigma(\mathrm{W}_{vz}\mathbf{V_t} + \mathrm{W}_{hz}h_{t-1} + b_z) && (7) \\ \tilde{h}_t &= \tanh(\mathrm{W}_{v\tilde{h}}\mathbf{V_t} + \mathrm{W}_{h\tilde{h}}(r_t \odot h_{t-1} + b_{\tilde{h}})) && (8) \\ h_t &= z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t && (9) \end{aligned}$$

where r_t, z_t, $\tilde{h}_t$, h_t are the reset gate, update gate, hidden candidate, and hidden state respectively. $\mathrm{W}_{[\cdot][\cdot]}$ are the weight matrices and $b_{[\cdot]}$ are the bias terms. In addition, σ stands for the sigmoid function in our Bi-GRU and $\odot$ represents the element-wise multiplication. These gate mechanisms allow the GRU to capture information from local to global view and produce the output based on different levels of features.

Since the dependencies among different levels of features can be estimated from both local to global view and global to local view, we utilize the bidirectional GRUs that consist of a forward GRU and a backward GRU (illustrated in Fig. 5) to model the relationships from two different views, which follows the practical intuition. The final hidden states from our bidirectional GRUs model are concatenated to be fed into the softmax classifier.

$$\mathbf{H} = [\overrightarrow{h_T}, \overleftarrow{h_T}] \tag{10}$$

We also use the same loss, i.e. negative log-likelihood(NLL) function, to train our Bi-GRU model.

$$\mathcal{L} = L_{cls}(\mathbf{H}) + \lambda\|\theta\|_2 \tag{11}$$

where λ is the weight factor and θ represents the weight parameters in Bi-GRU.

4 Experiments

4.1 Implementation Details

During the training process, we apply the multi-stage training strategy. Specifically, if the first branch training is finished, the second branch training will begin with the first branch fixed, and so on. When the whole CNN training is finished, the Bi-GRU training process will begin and the CNN part will be fine-tuned with a relatively small learning rate. The detailed architecture of CNN is presented in the Table in Fig. 4, and the size of the input patches is cropped as 375 × 375 from the corners and the center. The three fully connected layers from the top to the bottom in Fig. 2 contain 1,024, 512 and

512 neurons, respectively. The Bi-GRU has 512 hidden units and we apply the dropout on top of the output of Bi-GRU to avoid overfitting. We set $\lambda = 0.5$ in E.q (11) to balance the loss function and the regularization term, and set margin $\mu = 1$. The batch size is set to 64, and the CNN part is optimized by using the SGD with learning rate = 0.001 and Bi-GRU is optimized by using Rmsprop [Tieleman and Hinton, 2012] with the learning rate as 0.0001. In addition, a staircase weight decay is applied after 10 epoches. The parameters in these optimizers are initialized by using the default setting. Our model is implemented by using Torch7 [Collobert *et al.*, 2011] on one Nvidia GTX Titan X. Our model and results are available online[1].

4.2 Compared Methods

We compare our model with the following baseline methods.

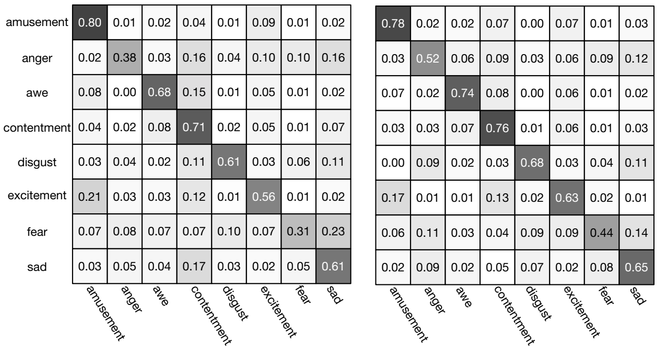
- **Machajdik** [Machajdik and Hanbury, 2010]: It explores the psychology and art theory to extract features that are specific to the domain of artworks.
- **ResNet-101** [He *et al.*, 2015]: It is pre-trained based on the ImageNet and fine-tunes by using emotion dataset.
- **AlexNet+SVM** [You *et al.*, 2016]: It extracts the features from AlexNet and uses SVM to classify emotions.
- **Zhao** [Zhao *et al.*, 2014]: It uses hand-crafted principles-of-art-based features to classify the emotions.
- **Rao** [Rao *et al.*, 2016a]: It uses the hand-crafted multiple-level features extracted from different image patches.
- **MldrNet** [Rao *et al.*, 2016b]: It uses multiple levels of features extracted from three different networks.

To verify the contributions of different components in our model, we design different variants of our model as follows:

- **CNN+Without Branch+Softmax:** It only uses the main architecture of the CNN part to extract features and classifies emotions with the softmax function.
- **CNN+5 Branches+Ensemble:** It uses five branches (i.e. four branches and one main branch) to train five classifiers respectively, then predicts emotions according to the average category scores from five classifiers.
- **CNN+5 Branches+LSTM:** It uses five branches to extract the features at different levels, and utilizes a unidirectional LSTM to integrate these features.
- **CNN+5 Branches+GRU:** It uses five branches to extract the features at different levels, and utilizes a unidirectional GRU model to integrate these features.
- **CNN+4/6 Branches+Bi-GRU:** It uses four or six branches (i.e. 3/5 branches and 1 main branch) to extract the features, and utilizes our bidirectional GRU model to integrate these features.
- **CNN+5 Branches+Bi-GRU:** It uses five branches to extract the features at different levels, and utilizes our bidirectional GRU model to integrate these features.

Methods	Accuracy
Zhao	46.52%
Rao	51.67%
AlexNet+SVM	57.89%
ResNet-101	60.82%
MldrNet	65.23%
CNN+W/O Branch+Softmax	59.61%
CNN+5 Branches+Ensemble	66.78%
CNN+5 Branches+LSTM	70.52%
CNN+5 Branches+Bi-LSTM	72.24%
CNN+5 Branches+GRU	71.33%
CNN+4 Branches+Bi-GRU	69.75%
CNN+6 Branches+Bi-GRU	72.97%
CNN+5 Branches+Bi-GRU	**73.03%**

Table 1: Emotion classification accuracy of different methods on the large scale emotion dataset.



	amusement	anger	awe	contentment	disgust	excitement	fear	sad
amusement	0.80	0.01	0.02	0.04	0.01	0.09	0.01	0.02
anger	0.02	0.38	0.03	0.16	0.04	0.10	0.10	0.16
awe	0.08	0.00	0.68	0.15	0.01	0.05	0.01	0.02
contentment	0.04	0.02	0.08	0.71	0.02	0.05	0.01	0.07
disgust	0.03	0.04	0.02	0.11	0.61	0.03	0.06	0.11
excitement	0.21	0.03	0.03	0.12	0.01	0.56	0.01	0.02
fear	0.07	0.08	0.07	0.07	0.10	0.07	0.31	0.23
sad	0.03	0.05	0.04	0.17	0.03	0.02	0.05	0.61

amusement	anger	awe	contentment	disgust	excitement	fear	sad
0.78	0.02	0.02	0.07	0.00	0.07	0.01	0.03
0.03	0.52	0.06	0.09	0.03	0.06	0.09	0.12
0.07	0.02	0.74	0.08	0.00	0.06	0.01	0.02
0.03	0.03	0.07	0.76	0.01	0.06	0.01	0.03
0.00	0.09	0.02	0.03	0.68	0.03	0.04	0.11
0.17	0.01	0.01	0.13	0.02	0.63	0.02	0.01
0.06	0.11	0.03	0.04	0.09	0.09	0.44	0.14
0.02	0.09	0.02	0.05	0.07	0.02	0.08	0.65

Figure 6: The confusion matrix of MldrNet (left) and our model (right).

4.3 Experiments on Large Scale Emotion Dataset

The large scale emotion dataset is recently published in [You *et al.*, 2016], which contains 8 different emotion categories including positive emotions: *Amusement*, *Awe*, *Contentment* and *Excitement* and negative emotions: *Anger*, *Disgust*, *Fear* and *Sad*. The original dataset consists of 90,000 noisy images collected from Instagram and Flicker by searching the emotion keywords. They further labeled them by using Amazon Mechanical Turk. Finally, there are 23,308 labeled images for emotion recognition [2]. We use the labeled dataset and the same training/testing split as in [Rao *et al.*, 2016b] to evaluate these methods. Specifically, the dataset is randomly split into a training set (80%, 18,532 images), a testing set (15%, 3,474 images) and a validation set (5%, 1,158 images).

We compare our proposed model and its variants with these baseline methods on the large scale emotion dataset. The results are shown in Table 1. We have the following observations. First, the methods using deep representation outperform the methods using the hand-crafted features. Then these models based on different levels of features (MldrNet and ours) outperform the methods using single-level features

[1] https://github.com/WERush/Unified_CNN_RNN

[2] We can only download 23,164 images with accurate labels as some images do not exist in the Internet now.

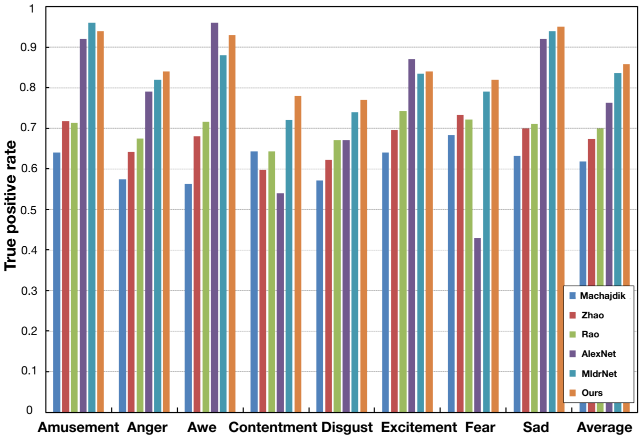


Figure 7: Performance evaluation on the ArtPhoto dataset.

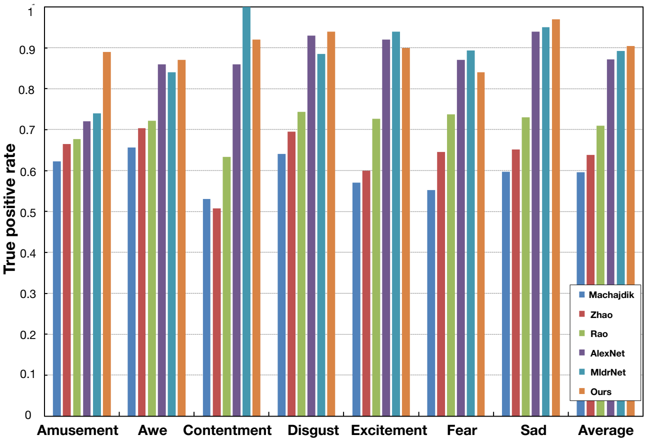


Figure 8: Performance evaluation on the IAPS-Subset dataset.

(ResNet and AlexNet). Moreover, our feature fusion methods using the RNN (LSTM/GRU) significantly outperform all baseline methods. More specifically, our methods with bidirectional GRU achieve better performance than methods using LSTM or unidirectional GRU, which demonstrates the bidirectional GRU approach plays a crucial role in our model. The model with 4 branches performs worse than the five branches due to the lack of high-level features, and the model with 6 branches achieves comparable performance with the five branches but with more parameters.

To further compare our method with MldrNet [Rao *et al.*, 2016b], we report the confusion matrix for each emotion. MldrNet also utilizes the features from different levels to predict the emotions, however, it does not utilize the RNN based feature fusion approach. As shown in Fig. 6, our method achieves a more balanced performance, especially for some negative emotions, such as *anger* and *fear*.

4.4 Experiments on Two Small Scale Datasets

Two small scale datasets including ArtPhoto and IAPS are also widely used in previous works. The ArtPhoto dataset [Machajdik and Hanbury, 2010] consists of 806 photos and all images are collected from art sharing sites. In [Mikels *et al.*, 2005], 395 images are collected from the standard IAPS dataset and labeled with arousal and valence values, which formed the IAPS-Subset dataset.

For fair comparison, we follow the pervious work [Machajdik and Hanbury, 2010] to evaluate different methods on the small scale emotion datasets, in which the same "one against all" strategy is used to train the emotion classifier. In addition, we pre-train our model using the large scale emotion dataset and fine-tune the last fully connected layer by using the small scale emotion datasets. We separate the dataset into the training and testing sets with K-fold cross validation (K=5). The *true positive rate per class* is reported to evaluate different methods. Note that there are only eight images in the category *anger* in the IAPS-Subset dataset, so we are unable to train a classifier for this category, and we only report the results over seven categories for this dataset.

We compare our method with five baseline methods mentioned in section 4.2 on the two small scale datasets. The results on the Artphoto dataset is presented in Fig. 7. We report not only the true positive rate per class but also the average true positive rate over all emotion categories in the last column. According to the results, our model is better on average than other methods, especially for some difficult emotion categories including *Fear* and *Sad*.

Fig. 8 shows the performances on the IAPS-Subset dataset. From the results, we observe that the deep learning methods outperform methods using the hand-crafted features. The methods based on multiple levels of features (MldrNet and ours) generally outperform the methods which only utilize the high-level features (AlexNet). Furthermore, although our method does not achieve the best results in all emotion categories due to the limited number of training images, on average our method still achieves the best performance because we exploit the dependency for feature fusion.

5 Conclusions

In this paper, we have proposed a unified CNN-RNN model for visual emotion recognition. Our model leverages different levels of features from multiple branches in CNN and effectively integrates these features by exploiting the dependencies among them with the bidirectional GRU approach. The proposed method achieves much better performance than the state-of-the-art methods. To the best of our knowledge, our work is the first to utilize Bi-GRU for feature fusion, and the in-depth studies demonstrate that our Bi-GRU model plays an important role for performance improvement.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (NSFC): 61332016, 61402431, 61620106009, U1636214 and 61650202, in part by National Basic Research Program of China (973 Program): 2015CB351800, in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013. We also wish to thank Yuexin Ma for discussion and anonymous reviewers for their valueable comments.

References

[Caruana, 1998] Rich Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.

[Chung *et al.*, 2015] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *NIPS*, pages 2980–2988, 2015.

[Collobert *et al.*, 2011] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, number EPFL-CONF-192376, 2011.

[Donahue *et al.*, 2014] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, pages 647–655, 2014.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015.

[Joshi *et al.*, 2011] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5):94–115, 2011.

[Kang, 2003] Hang-Bong Kang. Affective content detection using hmms. In *ACM MM*, pages 259–262. ACM, 2003.

[Lang, 1979] Peter J Lang. A bio-informational theory of emotional imagery. *Psychophysiology*, 16(6):495–512, 1979.

[Li *et al.*, 2012] Liang. Li, Shuqiang. Jiang, and Qingming. Huang. Learning hierarchical semantic description via mixed-norm regularization for image understanding. *IEEE Trans. on Multimedia*, 14(5), 2012.

[Liu *et al.*, 2015] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In *ICCV*, pages 1377–1385, 2015.

[Luong *et al.*, 2015] Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. *arXiv:1511.06114*, 2015.

[Machajdik and Hanbury, 2010] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, pages 83–92. ACM, 2010.

[Mikels *et al.*, 2005] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior research methods*, 37(4):626–630, 2005.

[Rao *et al.*, 2016a] Tianrong Rao, Min Xu, Huiying Liu, Jinqiao Wang, and Ian Burnett. Multi-scale blocks based image emotion classification using multiple instance learning. In *ICIP 2016*, pages 634–638. IEEE, 2016.

[Rao *et al.*, 2016b] Tianrong Rao, Min Xu, and Dong Xu. Learning multi-level deep representations for image emotion classification. *arXiv:1611.07145*, 2016.

[Sartori *et al.*, 2015] Andreza Sartori, Dubravko Culibrk, Yan Yan, and Nicu Sebe. Who's afraid of itten: Using the art theory of color combination to analyze emotions in abstract paintings. In *ACM MM*, pages 311–320. ACM, 2015.

[Siersdorfer *et al.*, 2010] Stefan Siersdorfer, Enrico Minack, Fan Deng, and Jonathon Hare. Analyzing and predicting sentiment of images on the social web. In *ACM MM*, pages 715–718. ACM, 2010.

[Solli and Lenz, 2009] Martin Solli and Reiner Lenz. Color based bags-of-emotions. In *International Conference on Computer Analysis of Images and Patterns*, pages 573–580. Springer, 2009.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[Tieleman and Hinton, 2012] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2), 2012.

[Wang and He, 2008] Weining Wang and Qianhua He. A survey on emotional semantic image retrieval. In *ICIP*, pages 117–120, 2008.

[Yanulevskaya *et al.*, 2008] Victoria Yanulevskaya, JC Van Gemert, Katharina Roth, Ann-Katrin Herbold, Nicu Sebe, and Jan-Mark Geusebroek. Emotional valence categorization using holistic image features. In *ICIP*, pages 101–104. IEEE, 2008.

[You *et al.*, 2016] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. *arXiv:1605.02677*, 2016.

[Yue-Hei Ng *et al.*, 2015] Joe Yue-Hei Ng, Fan Yang, and Larry S Davis. Exploiting local features from deep networks for image retrieval. In *CVPR Workshops*, pages 53–61, 2015.

[Zeiler and Fergus, 2014] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.

[Zhao *et al.*, 2014] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *ACM MM*, pages 47–56. ACM, 2014.

[Zhou *et al.*, 2014] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv:1412.6856*, 2014.