CONVOLUTIONAL NEURAL NETWORKS FOR IMAGE EMOTION RECOGNITION BY FUSING DIFFERENTIAL AND SUPPLEMENTARY INFORMATION

Shashank Rapolu, Aman Singh, Ajaz Hussain, Ayush Dhingra, Tushar Sandhan

Indian Institute of Technology Kanpur

ABSTRACT

Images are the most accurate depictions an individual can experience and one of the most effective ways to communicate emotions is through an image. Emotion recognition is used in many applications, such as advertising, social networking and cinema. In this paper, we propose a deep convolutional neural network (CNN) fusion technique made up of two parts: a differential-CNN system to extract emotional features from an image and combine them using an idea of convex-combination, and a supplementary-CNN to extract central object details of an image and combine it with differential-CNN features. We combine and amplify the difference between minute latent representations of an image via the convex combination of differential-CNN features. It also ensures the compactness of the latent feature space. The type of emotion an image elicits is highly influenced by its central object, hence supplementary-CNN is helpful in supplementing the differential-CNN features to improve emotion recognition. In comparison to the contemporary state-of-the-art methods, our proposed method showed an absolute gain of 6% on the primary dataset and outperformed them on the other secondary datasets.

Index Terms— Deep learning, image emotion, latent information, convolutional neural networks

1. INTRODUCTION

Research on emotion analysis has grown significantly in the past two decades. Several studies indicated that spatial organization, colorfulness, composition and other factors might affect emotional response to an image. Previous attempts to solve this challenge used handcrafted methods based on psychological ideas [1] and image perception. Color, edges, lines, texture, composition and picture descriptors like Histogram of Oriented Gradients (HOG) were handcrafted. Balance, emphasis and harmony were linked to the emotions elicited from an image. Lu et al. [2] attempted to retrieve shape features and classify the emotions of an image. Yanulevskaya et al. [3] retrieved Gabor and Wiccest surface texture features from images and mapped them to emotions. The type of line influences the emotional response as well. In the image, oblique, horizontal and vertical lines had distinct effects on the emotional response. Chang et al. [4] extracted edges, ridges and lines to classify images. Zhao et al. [5] extracted features using image descriptors (such as Histogram of Oriented Gradients) and combined them with other handcrafted features. Both Mehrabian and Valdez [6] attempted to map color primitives to emotions and classify images.

Following the introduction of convolutional neural networks (CNNs) [7] and their rising prevalence in ubiquitous applications, the emphasis switched from handcrafted methods to the CNN usage. Due to their ability to learn, CNNs proved to be significantly superior to handcrafted features. Based on the emotion categories inspired by Machadjik and Hanbury [8], You et al. [9] created a large-scale dataset for visual emotion identification. Most recent works have evaluated the performance of their methods using this dataset and the benchmark results were produced with AlexNet [10]. Xu et al. [11] suggested a CNN-based framework for visual sentiment prediction. Considering the emotional subjectivity of an image, Yang et al. [12] provided a framework for retrieving and classifying affective images. Campos et al. and Jou et al. [13] employed CNNs to predict visual sentiment by visualizing the image's local patterns. Chen et al. [14] presented a method for classifying images based on visual sentiment concepts using deep CNN architectures.

In our work, we employed a differential-CNN system to extract emotional elements from images and combine them with supplementary information to predict emotions. We also developed a novel idea of convex combination of features, which uses emotional data extracted from the differential-CNN system and predicts using the combined features. For training and evaluation, we used a large-scale dataset classified into eight emotions as well as additional small-scale datasets only for evaluation. As previously demonstrated [15], the emotion evoked by an image depends on the central object present in that image. For example, the presence of snakes and insects in an image elicits predominantly fear (or) disgust; whereas the presence of mountains and lakes (water) elicits awe (or) contentment. As a result, we had provided the central object information using supplementary-CNN along with emotional aspects (differential-CNN features) for emotion prediction. Not only does the central object has an influence, but our experiments demonstrate that amplification of it abruptly alters emotion detection from an image.

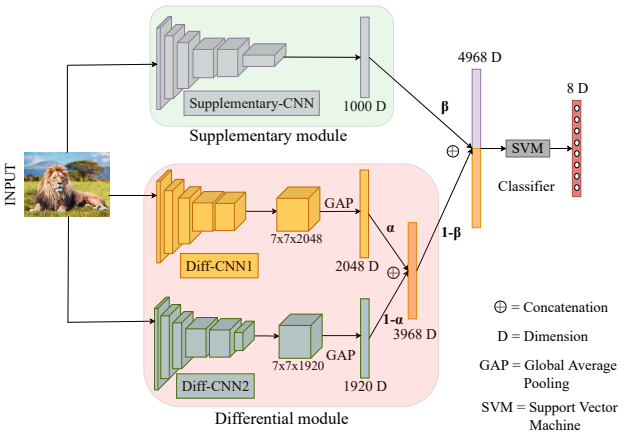

Fig. 1: Our final proposed network consists of a differential module containing a dual-CNN network for emotional-feature extraction and a supplementary module for image's central object details. Parameters α and β are differential and amplification factors, respectively. Convex combination of differential features and supplementary-feature amplification are performed separately and are further fine-tuned using SVM classifier with the CNN layers frozen.

2. OUR METHOD

Different CNNs extract and highlight different features from the same image. The intuition behind employing a dual CNN network in differential-CNN structure for feature extraction is that by merging them, all of the emotional elements from an image will be included because the features extracted from one differential CNN may not exist or may be repressed by the other. As finding the right combination is challenging, we have tested with several already available CNN models. It is observed from the experiments conducted by us that deep CNNs are superior at emotion recognition than shallow CNNs and hence we have used deep CNNs for differential feature extraction and their combination.

2.1. Differential CNNs for amplifying incremental latent information

First, we train Diff-CNN1 and Diff-CNN2 separately on the Image Emotion dataset. The feature matrix extracted from Diff-CNN1 is passed through GlobalAveragePooling (GAP) layer obtaining a vector representing the corresponding emotional features (differential-CNN features). These features extracted from the GAP layer are then passed to a dense layer containing eight nodes based on eight different emotions. The output from this last layer is transformed into a probability distribution using the softmax function. Similar training is implemented for Diff-CNN2, where a different feature matrix is obtained, eliciting a different vector when passed through the GAP layer. These two differential feature vectors are concatenated before passing them through the 8-node classifier.

Table 1: Number of images present in each emotion category of various datasets evaluated in this paper. The abbreviations IE, E-ROI, and AP stand for Image Emotion, EmotionROI and AbstractPaintings, respectively and '-' denotes absence of that specific emotion category.

Emotion	IE [9]	E-ROI [12]	Artphoto [8]	AP	Ours
Amusement	4724	-	101	33	-
Anger	1176	330	77	9	-
Awe	2883	330	102	28	-
Contentment	5131	-	70	72	407
Disgust	1591	330	70	29	471
Excitement	2727	-	105	36	430
Fear	969	330	115	41	418
Sadness	2635	330	166	32	420
Total	**21836**	1650	806	280	2146

Not only does the combination of features extracted from differential-CNN module has an effect, but the weightage of this combination also significantly affects the performance. With this idea, we have introduced a concept of convex-combination of features extracted from the differential module, governed by a parameter α called differential factor: (See Fig. 1 for representation)

$$X = \alpha X_1 \oplus (1-\alpha) X_2 \tag{1}$$

where X_1 and X_2 are feature vectors of Diff-CNN1 and Diff-CNN2 respectively, X is a concatenated feature vector with a differential factor α and $\oplus$ represents concatenation.

$0 \leq \alpha < 0.5$	X_1 is suppressed, X_2 is highlighted
$\alpha = 0.5$	X_1 and X_2 are equally considered
$0.5 < \alpha \leq 1$	X_1 is highlighted, X_2 is suppressed

2.2. Supplementary feature fusion and amplification

Furthermore, the differential-CNN system is supplemented with additional information containing central object details. As elucidated in Section 1, a 1000-D vector is concatenated with the differential-CNN feature vector, acting as supplementary information. Another pre-trained CNN initialized with imagenet weights producing a probability distribution of 1000 different objects is used, called as Supplementary-CNN. Similar to convex-combination, amplifying the image's central object information might affect how emotions are perceived. The supplementary-feature amplification, representing the central object w.r.t the combined emotional aspects (from differential module), is controlled by an amplification factor, denoted by the parameter β. It should be noted that supplementary-feature amplification is performed for direct concatenation of differential features i.e. without their convex combination. A higher value of β denotes higher prominence of central object present in that image, whereas lower β denotes its lesser prominence.

Table 2: Performance of our methods on all the datasets used for evaluation. The parameter β=1 is irrelevant for the supplementary-feature amplification model since emotional qualities are ignored, and '-' reflects it. Abbreviations IE, E-ROI, and AP stand for Image Emotion, EmotionROI, and AbstractPaintings and the best results obtained are represented in **bold**.

Parameter	Image emotion recognition accuracy (%)									
	Differential convex combination model (α)					Supplementary-feature amplification model (β)				
	IE	AP	ArtPhoto	E-ROI	Ours	IE	AP	ArtPhoto	E-ROI	Ours
0	64.85	19	44	21	64	66.71	26	**45**	23	**71**
0.25	65.76	**35**	14	21	68	65.27	26	44	21	**71**
0.5	66.62	32	43	**62**	67	66.65	26	**45**	21	**71**
0.75	**66.80**	34	44	61	70	65.76	26	**45**	22	70
1	66.28	33	43	60	66	-	-	-	-	-

2.3. Fine-tuning with SVM classifier

To further enhance emotion recognition of our models, fine-tuning with a Support Vector Machine (SVM) classifier is implemented. The features extracted from the differential-CNN and supplementary-CNN modules are contained in the concatenated vector, which is extracted and passed through the SVM classifier with rbf kernel. Only the SVM classifier is trained at this stage after all preceding CNN layers are frozen.

3. EXPERIMENTS

Experiments were conducted on various datasets, implemented using specific training parameters and compared with the baselines as mentioned below.

Datasets: In our work, both large-scale and small-scale datasets are taken into account. The primary dataset is the Image Emotion dataset, which consists of 21,836 images unevenly distributed into 8 emotion categories. Like prior studies, we randomly and uniformly split the entire dataset into 85% for training and 15% for testing. Abstract Paintings, ArtPhoto and EmotionROI are small-scale datasets and are only used for evaluation. We also constructed a dataset from scratch to further test our approach on a medium-scaled dataset. Initially, each of us collected images from the internet totaling up to 2500 images. The gathered images were then examined together and classified into five emotion categories. In order to scrape images from the web, we employed several keywords. As an illustration, we used cold beverages, games, and romance to create excitement and trash, filth, and diseases to create disgust. However, we did not include the emotion anger because it is hard to locate images that elicit anger. We omitted awe and amusement as image visuals overlap with contentment and excitement. To test our final model, we gathered 2146 images based on the remaining categories in this manner. Details of all the datasets are given in Table 1.

Implementation details: ResNet101 and DenseNet201 are chosen for Diff-CNN1 and Diff-CNN2, respectively and ResNet101 is used as supplementary-CNN. A 224x224x3 pixel image is passed through the convolutional layers. Images were rescaled by a factor of 1/255 and a batch size of 64 was used for training and evaluation. Before passing the input data to the convolutional layers, we preprocessed it with CNN-specific preprocessors. The optimizer used was SGD with a learning rate of 0.001 and categorical cross-entropy as a loss function. All models were trained until the accuracy of the test dataset increased while test loss fell dramatically and then increased marginally towards the end.

Table 3: Comparison of our two fine-tuned CNN variants (in **bold**) with various state-of-the-art methods.

Type	Methods	Accuracy (%)
Handcrafted	SIFT [16]	37.56
	HOG [12]	44.67
	SentiBank [14]	49.09
Deep Learning	DeepSentiBank [12]	54.10
	AlexNet (fine-tuned) [9]	56.55
	Binary Assisted [17]	61.31
	LiteEmote [18]	61.67
	Supplementary-feature amplification (β=0)	**67.41**
	Differential convex-combination (α=0.75)	**67.75**

Baselines: We contrasted our model against both handcrafted and deep learning approaches. In handcrafted methods we compared with SIFT [16], HOG [12] and SentiBank [14]. Low-level features like HOG and mid-level features like SentiBank were retrieved from images using handmade algorithms. For deep learning methods, we compared with DeepSentiBank [12], which classifies emotions using adjective-noun pairs, a fine-tuned Alexnet model [9] trained on the Image Emotion dataset, the Binary Assisted Classification Network [17], which recognizes emotions using a single CNN and binary classification of emotions, and the LiteEmote model [18], which combines object- and category-specific features with emotional features.

3.1. Results on the Image Emotion (IE) dataset

Experiments were carried out on the test dataset comprised of 15% of the images from the Image Emotion dataset. Diff-CNN1 and Diff-CNN2, trained individually on Image Emotion dataset, scored 64.14% and 65.06%, respectively. Our fine-tuned differential convex-combination model (α=0.75) achieves the best accuracy of 67.75%. The differential convex combination in Table 2 demonstrates how performance is significantly impacted by the differential factor (α). Ac-

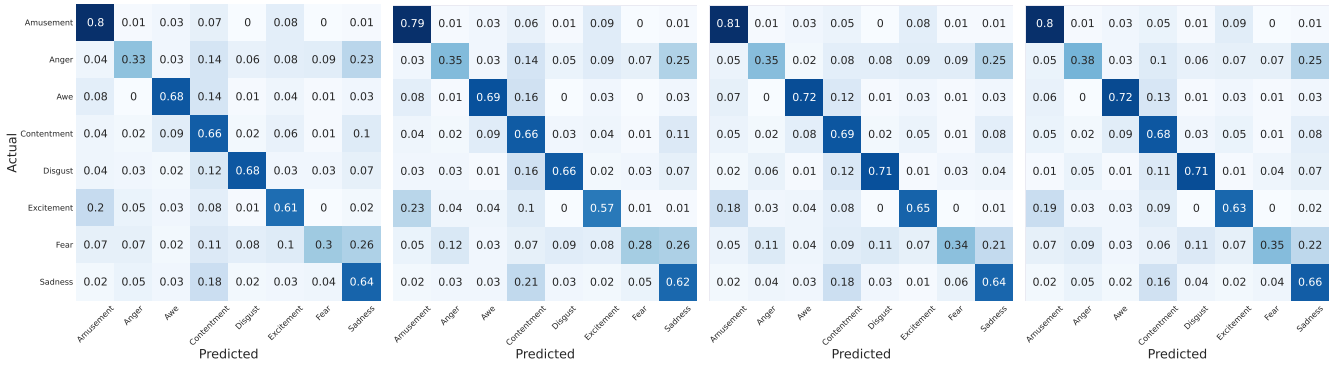

Fig. 2: Confusion matrices to evaluate performance on each emotion category of the Image Emotion dataset. The two confusion matrices (from left) are for two Diff-CNN models trained individually, respectively. The third confusion matrix is for the best (α=0.75) differential convex-combination model and the final is for the best (β=0) supplementary-feature amplification model.

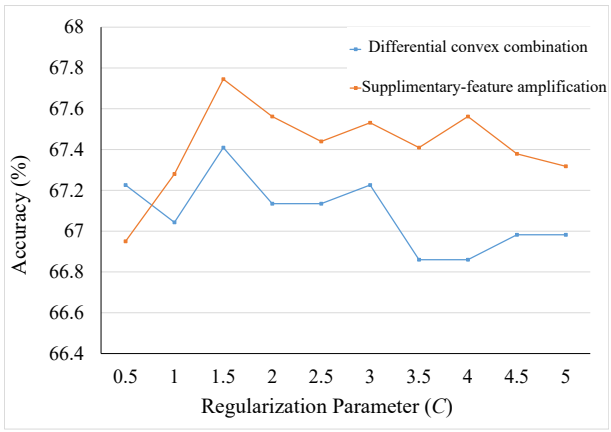

Fig. 3: Accuracy (vs) regularization parameter C of our two best-performing final models on the Image Emotion dataset.

curacy increased from 65.06% to 66.28% for α = 1, which only includes Diff-CNN1 with the supplementary-CNN. Accuracy increased from 64.14% to 64.85% when α = 0, which only applies to Diff-CNN2 with the supplementary data. Furthermore, accuracy increased to 66.62% when α = 0.5, denoting that features from the differential-CNN system are given equal importance. This proves that combining two different CNN characteristics will mask any CNN features that are missing from one CNN. The supplementary-feature amplification model's performance increased by 1.38% when β value increased by 25% from β=0.25 to β=0.5, demonstrating that amplification of the image's central object has a substantial impact on the model's performance. Performance was further improved by fine-tuning the two aforementioned models with the SVM classifier. Accuracy increased from 66.71% to 67.41% for supplementary-feature amplification model with β=0 and from 66.80% to 67.75% for the differential convex-combination model with α=0.75.

3.2. Effect of regularization parameter (C)

We have also tested the effect of SVM classifier's regularization parameter C. Fig. 3 demonstrates how changes in the parameter have a sudden impact on the fine-tuning of our two models. These two models, with α=0.75 and β=0, respectively had the best results for C=1.5.

3.3. Results on small-scale datasets

We also evaluated our models on publicly available small-scale datasets. With α values of 0.25 and 0.5, our differential convex-combination model performed best on AbstractPaintings and EmotionROI datasets, respectively. Surprisingly, the supplementary-feature amplification model had the best accuracy on Artphoto and our proposed dataset, compared to the aforementioned model. On the Artphoto dataset, our model outperformed LiteEmote model [18] by **2%**. Despite having lesser images, the model's performance on our proposed dataset, obtaining 71%, is unexpectedly better than on the Image Emotion dataset. (Results in Table 2)

4. CONCLUSION

In this study, we suggested a unique method for recognizing emotions from images. Based on the idea that features obtained by one CNN might not include characteristics extracted by the second CNN, a differential-CNN system using a dual-CNN network improves emotion recognition over a single-CNN network. We also proposed the convex combination of differential features and supplementary feature amplification using differential and amplification factors. Adding more CNNs to the differential module without sacrificing computing intensity might result in even better outcomes. Other emotion-dependent elements, such as low-level or background features, can be investigated further to enhance the detection of an image's emotion.

5. REFERENCES

[1] 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*, vol. 28, no. 5, pp. 94–115, 2011.

[2] Xin Lu, Poonam Suryanarayan, Reginald B Adams Jr, Jia Li, Michelle G Newman, and James Z Wang, "On shape and the computability of emotions," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 229–238.

[3] Victoria Yanulevskaya, Jan C van Gemert, Katharina Roth, Ann-Katrin Herbold, Nicu Sebe, and Jan-Mark Geusebroek, "Emotional valence categorization using holistic image features," in *2008 15th IEEE international conference on Image Processing*. IEEE, 2008, pp. 101–104.

[4] Le Chang, Yufeng Chen, Fengxia Li, Meiling Sun, and Chenguang Yang, "Affective image classification using multi-scale emotion factorization features," in *2016 International Conference on Virtual Reality and Visualization (ICVRV)*. IEEE, 2016, pp. 170–174.

[5] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun, "Exploring principles-of-art features for image emotion recognition," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 47–56.

[6] Patricia Valdez and Albert Mehrabian, "Effects of color on emotions.," *Journal of experimental psychology: General*, vol. 123, no. 4, pp. 394, 1994.

[7] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[8] Jana Machajdik and Allan Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 83–92.

[9] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *Proceedings of the AAAI conference on artificial intelligence*, 2016, vol. 30.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[11] C Xu, S Cetintas, KC Lee, and LJ Li, "Visual sentiment prediction with deep convolutional neural networks (2014)," *arXiv preprint arXiv:1411.5731*.

[12] Jufeng Yang, Dongyu She, Yu-Kun Lai, and Ming-Hsuan Yang, "Retrieving and classifying affective images via deep metric learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32.

[13] Victor Campos, Brendan Jou, and Xavier Giro-i Nieto, "From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction," *Image and Vision Computing*, vol. 65, pp. 15–22, 2017.

[14] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," *arXiv preprint arXiv:1410.8586*, 2014.

[15] Hye-Rin Kim, Yeong-Seok Kim, Seon Joo Kim, and In-Kwon Lee, "Building emotional machines: Recognizing image emotions through deep neural networks," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 2980–2992, 2018.

[16] Tianrong Rao, Min Xu, Huiying Liu, Jinqiao Wang, and Ian Burnett, "Multi-scale blocks based image emotion classification using multiple instance learning," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 634–638.

[17] Xuanyu He and Wei Zhang, "Emotion recognition by assisted learning with convolutional neural networks," *Neurocomputing*, vol. 291, pp. 187–194, 2018.

[18] Yan-Han Chew, Lai-Kuan Wong, John See, Huai-Qian Khor, and Balasubramanian Abivishaq, "Liteemo: lightweight deep neural networks for image emotion recognition," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–6.