Exploring Visual Features of Emotional Images on Social Media

Satoshi Sanjo[1] and Marie Katsurai[2]
[1]Graduate School of Engineering, Doshisha University
[2]Faculty of Science and Engineering, Doshisha University
1-3, Tatara Miyakodani, Kyotanabe, Kyoto 610-0394, Japan
E-mail : {sanjo, katsurai}@mm.doshisha.ac.jp

Abstract: With the rapid growth of smartphones and social media platforms, images have been used to express users' opinions, attitudes, and emotions in online communication. Detecting the emotions from images is crucial to develop affective retrieval and multimedia data mining techniques. Conventional methods collect a large number of images from social media, which are tagged with emotional words, and then train a Convolutional Neural Network (CNN) to model the relationship between emotions and images. However, how the CNN learns the emotions from social media has not been visually shown. In this paper, we explore visual characteristics of the emotional images on social media through automatic image generation using a class of CNNs called deep convolutional generative adversarial networks (DCGANs). Results of experiments present the visual appearance of each emotion, from which we can observe its specific characteristics. Furthermore, based on the trained DCGANs, we investigate the similarity between public image datasets. Our findings are also consistent with sentiment classification across different datasets.

1. Introduction

With the rapid growth of smartphones and social media platforms, images have been used to express users' opinions, attitudes, and emotions in online communication. For example, Instagram, which allows users to post images with text, has grown to have more than 400 million monthly active users [1]. Detecting emotions from user-generated images is crucial to develop several applications such as affective retrieval and multimedia data mining, and thus modeling of the relationship between images and the emotions has attracted much research attentions in recent years [2–5].

Many conventional methods use low-level features (e.g., color histograms) [6] or multimodal features (e.g., images and their captions [2, 3]) to train a sentiment classifier. On the other hand, recent methods use a Convolutional Neural Network (CNN), which has established new state-of-the-art results in several object recognition tasks, for image sentiment analysis [4, 5]. These methods first collect a large number of images from social media, which are tagged with emotional words, and then train a CNN based on the image collection. Specifically, You et al. [4] presented a new CNN architecture for sentiment polarity classification, while Campos et al. [5] fine-tuned a CNN that was pre-trained using ImageNet [7] for the polarity classification task. However, it is widely known that emotional tags on social media are noisy and depend on users' subjectivity. How the CNN can actually learn the emotions from such a noisy corpus has not been visually shown.

In this paper, as a first step to investigate the characteristic of emotional images collected from social media, we explore their visual features through visualization of model training. Specifically, we collect a set of images tagged with emotional words in similar to the conventional methods [4, 5], and then train a class of CNNs, called deep convolutional generative adversarial networks (DCGANs) [8]. DCGAN presents unsupervised representation learning of a given image dataset. Results of experiments show the visual characteristics memorized by DCGANs for each emotion category, from which we find that the results can depend on how the dataset was collected. To support our findings, we also show sentiment classification performances across different datasets.

2. DCGAN Training Using Emotional Images on Social Media

This section presents DCGAN training using emotional images on social media and calculating similarity matrix. First, we explain details of public datasets collected from social media (see 2.1). Next, we describe an overview of DCGAN training using the datasets (see 2.2). Finally, to analyze the difference of visual characteristics extracted from different emotion categories or datasets, we propose to calculate a similarity matrix between emotional images generated from DCGAN (see 2.3).

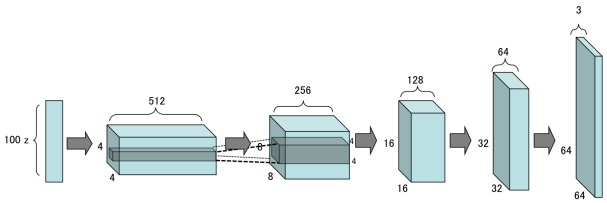
2.1 Dataset construction

Similar to the conventional studies [2, 3, 5], this paper analyzes two emotion categories, i.e., *positive emotion* and *negative emotion*, using two datasets provided by the authors in [2] and [3]. Details of each dataset are described below.

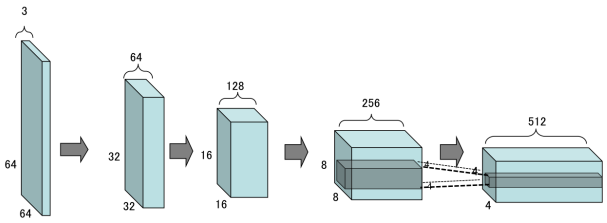
Dataset 1: SentiBank [2]

A dataset provided by [2], called SentiBank[1], includes 1,030,303 Flickr images that are tagged with adjective and noun pairs (ANPs) such as "*colorful clouds*" and "*crying baby*". From the dataset, we chose a set of images for each emotion category based on sentiment scores given for ANPs. The sentiment scores were calculated using SentiWordNet scores [9] for adjectives and nouns. For example, a sentiment score of "*colorful clouds*" is 1.53, while that of "*crying baby*" is -1.00. To improve the reliability of sentiment labels, we selected images such that the absolute values of the sentiment scores are larger than a certain threshold. In experiments, the threshold was set to 1.0. The resulting dataset, denoted by Dataset 1, consists of 424,212 positive emotional images and 222,109 negative emotional images. All images in the dataset are resized to 64×64 before training DCGANs.

[1]http://www.ee.columbia.edu/ln/dvmm/vso/download/flickr_dataset.html



(a)



(b)

Figure 1. Architectures of the generator and discriminator in DCGAN [8]. (a) Generator and (b) discriminator.

Dataset 2: Flickr images labeled by [3]

The authors in [3] assigned sentiment labels to a set of Flickr images via crowdsourcing. For each image, three workers were asked to choose its sentiment score on a discrete five-point scale labeled with "highly positive," "positive," "neutral," "negative," and "highly negative." The dataset is available on the web[2]. Because our experiments focus on only two emotion categories, i.e., positive and negative, we regard both "highly positive" and "positive" as positive emotions and discard "neutral". Similarly for negative emotions. The resulting dataset, denoted by Dataset 2, consists of 48,138 positive emotional images and 12,606 negative emotional images. All images in the dataset are also resized to 64×64 before training DCGANs.

Note that Dataset 1 depends on user-generated tags, while Dataset 2 was manually labeled by majority voting.

2.2 DCGAN Training

DCGANs are a variant of CNNs for unsupervised representation learning, which exploit an idea of GAN [10]. GAN generates a model distribution that represents a generation process of images using two neural networks, called generator and discriminator, respectively. The generator has a function that produces images using random vectors from a uniform or normal distribution, while the discriminator has a function that judges whether the input image given by the generator is real or fake. These two networks compete with each other: if images given by the generator can deceive the discriminator, then the generator wins. On the other hand, if the discriminator can accurately judge whether the input image is real or fake, then the discriminator wins. The training is performed so as to increase the probability of winning. Finally, the generator can produce images which seem real, and the discriminator can judge fake images with high possibility. The network architectures of DCGAN are shown in Fig. 1. We train

[2]http: //mm.doshisha.ac.jp/senti/CrossSentiment.html

Figure 2. Examples of positive emotional images generated from DCGAN learned by Dataset 1.

Figure 3. Examples of negative emotional images generated from DCGAN learned by Dataset 1.

DCGANs for emotion categories in each dataset. The batch size and the epoch size are set to 100 and 250, respectively. By providing a set of images generated by DCGANs, we can analyze what types of visual features are learned as emotional features in the given dataset.

2.3 Similarity matrix calculation

To analyze the difference of visual features captured for different emotion categories or datasets, we focus on the visual similarity between images generated from arbitrary two DCGANs. Specifically, a similarity matrix is calculated by the following process: we randomly generate 30 images from each DCGAN, and then extract their visual features from a pre-trained eight-layer CNN[3] [11]. Finally, we calculate the cosine similarity of visual features among the images.

3. Experiment

In this section, we perform experiments using Datasets 1 and 2 to show how DCGANs learn the visual features for each emotion category. We first show images generated from DCGAN for a pair of a dataset and an emotion category (see 3.1). Then, we investigate the difference between DCGANs learned for different emotion categories (see 3.2) and the dif-

[3]We used an output of the seventh fully-connected layer of the CNN as visual features.

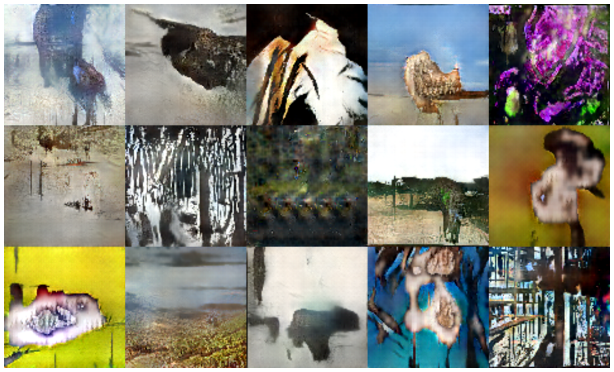


Figure 4. Examples of positive emotional images generated from DCGAN learned using Dataset 2.

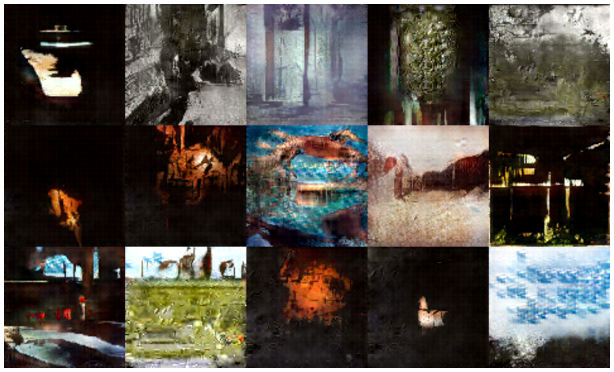


Figure 5. Examples of negative emotional images generated from DCGAN learned using Dataset 2.

ference between those for different datasets (see 3.3). Finally, to support our findings, we also evaluate the sentiment classification performance across the datasets (see 3.4).

3.1 Analyzing images generated by DCGANs

In this experiment, we show a set of images generated from DCGANs for two emotion categories in each dataset. Examples of positive and negative emotional images generated using Dataset 1 are shown in Figs. 2 and 3, respectively. Figure 2 consists of bright colors and specific objects related to nature, e.g., sky and flowers, while Fig. 3 consists of dark colors and seems to be noisy and complex. Similarly, examples of positive and negative emotional images generated using Dataset 2 are also shown in Figs. 4 and 5, respectively. The above-mentioned tendency can also be found in these figures. Comparing the resulting images in the Dataset 1 and 2, Fig. 4 is darker than Fig. 2, and Fig. 4 has a brightness similar to Fig. 3. From these results, we can consider as follows: (i) positive and negative emotions evoked from images mostly depend on colors; (ii) negative images tend to consist of complex objects compared with positive images; and (iii) although the difference between the emotion categories can be confirmed within a dataset, the visual appearances of the same emotion are not similar across different datasets.

3.2 Investigating the similarity between positive and negative emotional images

Using the approach described in Section 2.3, we calculated the visual similarity between the two emotion categories memorized from Dataset 1. The resulting matrix is shown in Fig. 6 (a). From this figure, we can see that images generated as positive emotion do not have strong correlations with those generated as negative. Thus, DCGAN learned for a specific emotion category in a given dataset is able to capture the visual characteristics from its images.

3.3 Investigating the similarity of positive emotional images across datasets

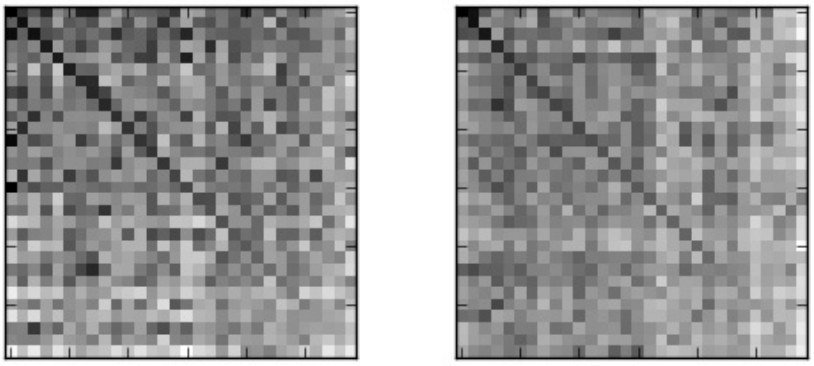
Next, we calculated the visual similarity between two sets of positive emotional images memorized from the two datasets. The resulting matrix is shown in Fig. 6 (b). The correlation between two DCGANs trained for the same emotion is higher than the correlation seen in Fig. 6 (a). However, it is not sufficiently high to show the visual consistency across datasets. Thus, we can consider that the emotional features captured by models significantly depend on how to construct the dataset.

3.4 Sentiment classification across different datasets

Finally, to confirm the above findings, we evaluate the performance of sentiment classifiers trained across different datasets. Here, we introduce an additional Twitter dataset [4], which contains 882 images (581 positive, 301 negative emotional images) that built consensus among five annotators. Following to [5], we fine-tuned an eight-layer CNNs using one of the datasets for binary sentiment classification. Note that each dataset was randomly divided into training and testing subsets before fine-tuning. Table 1 shows the classification accuracy in different combinations of datasets used for training and testing. The best accuracy was usually achieved when the same dataset was used for both training and testing: the performance got worse when different datasets were used for training and testing. This is because how to prepare sentiment labels for images is quite different: Dataset 1 was automatically labeled based on tags, while Dataset 2 and Twitter dataset were manually labeled. As seen, testing in Twitter dataset always shows the highest accuracy, while classifiers learned using Twitter dataset could not achive reasonable results in other datasets. One of possible reasons is that Twitter dataset has strong consensus among annotators, and the emotions in this dataset can be stronger than those in other datasets: i.e., the classifiers trained using Twitter dataset cannot capture weak emotions well. From these results, we should carefully design how to assign sentiment labels to a dataset.

4. Conclusion

In this paper, we explore visual features of emotional images on social media through visualization of representation learning. Results of the experiments show that DCGAN learned for a positive or negative emotion in each dataset is able to capture its visual characteristics: positive emotional images consist of bright colors, while negative images tend to have



(a) (b)

Figure 6. Similarity matrices for different DCGANs. (a) Similarity between positive and negative emotional images generated from Dataset 1 and (b) similarity between two sets of positive emotional images generated from Datasets 1 and 2.

Table 1. Accuracy of sentiment classification in different combinations of datasets used for training and testing.

	Training		
Testing	Dataset 1	Dataset 2	Twitter dataset
Dataset 1	0.678	0.593	0.522
Dataset 2	0.653	0.726	0.670
Twitter dataset	0.566	0.759	0.815

dark colors and complex objects. However, the results in Section 3.3 show that the emotions expressed in different datasets are not highly consistent. The reasons are not only that emotions have diverse concepts but also that how to label images with sentiment is quite different across datasets. Thus, towards sentiment classifier training, we must carefully design a way to assign the sentiment labels.

In future work, we will investigate whether sophisticated color features can improve the performance of image sentiment analysis. In addition, we should develop a sentiment classifier that has high generalization ability in different datasets.

References

[1] Instagram Blog, "Celebrating a community of 400 million," http://blog.instagram.com/post/129662501137/150922-400million, Sep 2015, Last accessed: 04/13/2016.

[2] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. of the 21st ACM Int. Conf. on Multimedia*. ACM, 2013, pp. 223–232.

[3] M. Katsurai and S. Satoh, "Image sentiment analysis using latent correlations among visual, textual, and sentiment views," in *Proc. of the 41st Int. Conf. on Acoustics, Speech, and Signal Processing*. ICASSP, 2016, pp. 2837–2841.

[4] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Twenty-Ninth AAAI Conf. on Artificial Intelligence*, 2015.

[5] V. Campos, A. Salvador, X. G.-i. Nieto, and B. Jou, "Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction," in *Proc. of the 1st Int. Workshop on Affect & Sentiment in Multimedia*. ACM, 2015, pp. 57–62.

[6] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *Proc. of the 18th ACM Int. Conf. on Multimedia*. ACM, 2010, pp. 715–718.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conf. on*. IEEE, 2009, pp. 248–255.

[8] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[9] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," in *Proc. Int. Conf. Language Resources and Evaluation (LREC)*, 2006, pp. 417–422.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems (NIPS 2014)*, 2014, pp. 2672–2680.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. Int. Conf. Multimedia (MM)*, 2014, pp. 675–678.