Visual Sentiment Analysis for Few-Shot Image Classification Based on Metric Learning

Tetsuya Asakawa * and Masaki Aono†
*† Toyohashi University of Technology, Aichi, Japan
* E-mail: asakawa@kde.cs.tut.ac.jp
† E-mail: aono@tut.jp

***Abstract*—Visual sentiment analysis is an interesting and challenging research problem. that investigates sentiment estimation from images. Most studies have focused on estimating a few specific sentiments and their intensities, using several complex convolutional neural network (CNN) models. However, sentiment estimation from a small number of images using few-shot 1-way learning has not been sufficiently investigated. This research aims to accurately estimate sentiment using few-shot 1-way learning from given images that evoke different emotions. We first introduce a visual sentiment dataset based on Plutchik' s wheel of emotions, called Senti8PW. We perform a few-shot image classification using Senti8PW, where we present a highly accurate deep neural network model with a small number of parameters and convolutions. We then use Senti8PW to perform experiments using few-way 1-shot learning. We also employ the Euclidean distance and Cosine similarity as a metric of our proposed model. Each emotion is assumed to have a probability distribution. After training our deep neural network, we predict an evoked emotion for a given unknown image. We also perform experiments to compare our proposed model with existing models. The classification system of four layers of convolutions with 5-way 1-shot learning proves to be the best in terms of balancing accuracy and a number of model parameters. Thus, results demonstrate that our model outperforms existing state-of-the-art algorithms with regard to the of balance between accuracy and parameter number.**

I. Introduction

With the proliferation of social networking serivices (SNS), such as Twitter, Instagram, Facebook and TikTok), , as well as smartphones, and the internet, SNS users capture numerous images as they go about their daily lives to record all kinds of activities. Therefore, a vast number of images exists on the internet, resulting in an urgent need for image indexing and retrieval techniques. An image can elicit several emotions, both positive or negative.. In other words, different visual images have different emotional triggers. For instance, an image of a snake or a bee will most likely trigger a bad feeling like "disgust" or "fear", whereas an image of a funny object or a laughing face will most likely trigger a good feeling like"Joy" or "Surprise".

Visual sentiment prediction investigates sentiment estimation from images, which is an interesting and challenging research problem. Furthermore, most previous studies have focused on estimating a few specific sentiments using several complex convolutional neural network (CNN) models. However, sentiment estimation from a small numbers of images using few-shot 1-way learning has not been sufficiently investigated. The purpose of this research is to estimate few-shot image sentiment classification accurately, using metric learning from given images that evoke different emotions.

Fan et al. [1] performed sentiment prediction using the Emotion6 dataset, and Peng et al. [2] performed sentiment prediction using the EmotionROI dataset. However, the existing Emotion6 dataset considers a small number of items, and has difficulty predicting multiple emotions. Therefore, we will first introduce a visual sentiment dataset, based on Plutchik' s wheel of emotions, called Senti8PW. No previous research to our knowledge has used Senti8PW.

n this paper, we employ few-shot 1-way learning, whereas most existing methods neglect to fulfill very important requirements for a good few-shot 1-way learning system. We perform experiments using few-way 1-shot learning by Senti8PW. We also employ the Euclidean distance and cosine similarity as a metric of our proposed model. In other words, we introduce a new neural network model. In addition, deep learning has weak classifications, therefore, we propose two layers that are fully-connected. The new contributions of this paper include the following: (1) a novel feature considering CNN features to predict the sentiment of images, unlike most recent research that only adopts CNN features, and (2) a few-shot 1-way learning object recognition system that is capable of dynamically learning novel categories from little training data while retaining the base categories on which it was trained.

We first survey related work in Section II, followed by introducing the proposed research method in Section III. Section IV presents the dataset and experiment settings. We describe the experiments we have carried out in Section V, and we conclude this paper in Section VI.

II. related work

Visual sentiment analysis and the building of a visual sentiment dataset are important tasks that have seen rapid development in recent years. In the visual sentiment dataset, there are image emotion datasets. Data was collected from social networking websites such as Flickr and Instagram (FI) [3], using the names of emotion categories as search terms. We collected 23,308 well-labeled images for emotion recognition in 8 categories (Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear and Sadness). EmotionRoI [2] contains 1,980 affective images from Flickr with labeled emotional regions in 2 categories (Positive and Negative). The

International Affective Picture System subset (IAPSsubset) [4] was a standard stimulus image set, which has been widely used in affective image classification in 8 categories (Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear and Sadness). IAPS consists of 1,182 documentary-style natural color images depicting complex scenes, such as portraits, puppies, babies, animals, and landscapes. In the ArtPhoto dataset [4] of eight categories, 806 photos were selected from some art sharing sites using the names of emotion categories as the search terms. The Abstract dataset [4] consists of 228 abstract paintings in 8 categories. Unlike the images in the IAPS-Subset and ArtPhoto dataset, the images in the Abstract dataset represent emotions through overall color and texture, instead of some emotional objects. The LUCFER dataset [5] is a collection of 3.6 million affective images using Bing's Cognitive Services API in eight categories. However, to our knowledge, no dataset has been built based on Plutchik' s wheel of emotions, which captures a myriad of human emotions.

In visual sentiment analysis, research on sentiment analysis is presented in terms of how many specific sentiments should be classified. Examples of the research group includes Katsurai et al. [6], who exploited latent correlations among visual, textual, and sentiment features for image sentiment classification, Fan el al. [7], who studied the relation between image sentiment and visual attention, and Cordel et al. [8], who proposed emotion-aware human attention prediction. Kim et al. [9] used feed forward deep learning for 8-class visual sentiment classification and Yang et al. [10], who proposed weakly supervised coupled networks for visual sentiment analysis using four emotions ("Joy", "Fear", "Anger" and "Sadness"). Recently, Zhang et al. [11] used a concatenated neural network model for eight labels visual sentiment classification. However, to our knowledge, none of the above research has dealt with "few-way 1-shot image" classification of emotions.

We built a visual sentiment dataset and performed single-label and multi-label visual sentiment analyses. Moreover, we utilized from the Plutchik's wheel of emotions[12]. We used the Euclidean distance[13] and Cosine similarity[14] on how to evaluate the metric of our proposed model' s metric. We adopted accuracy[15] to evaluate few-way image classifications.

III. PROPOSED METHOD

We propose a visual sentiment analysis system to predict emotion inspired by Plutchnik's wheel of emotions using the few-shot classification. To this end, we built a dataset that included eight emotion categories After building the dataset, we will describe our deep neural network model that enables single outputs using the few-shot classification, given images that evoke emotions.

A. Building Visual Sentiment Dataset

We collected a large set of images to build the visual sentiment dataset, We first searched for emotional images on Flickr [16] using the eight keywords (Surprise, Sadness, Joy, Disgust, Anger, Fear, Trust and Anticipation). These keywords are based on Plutchik's wheel of emotions[12]. Using these keywords, we collected more than three thousand images. Since the goal is to obtain emotional images, we manually eliminated non-emotional images that consist of simple monochromatic or textual images. The keywords used for the search and the number of images in the dataset are as follows: Surprise (300), Sadness (300), Joy (300), Disgust (300), Anger (300), Fear (300), Trust (300) and Anticipation (300). The total number of images was 2,400.

Second, we employed Amazon Mechanical Turk (AMT) workers to manually annotate each image in this dataset with a sentiment label by AMT workers. Specifically, five AMT workers were recruited to generate a sentiment label (Surprise, Sadness, Joy, Disgust, Anger, Fear, Trust and Anticipation) for each image. We call this visual sentiment dataset **Senti8PW**, which we have published on github [1]. An example of images corresponding to each Senti8PW emotion is depicted in Figure 1.

Table I presents the labeling results from AMT, where "Three Agree" means that the three or more AMT workers provided a consistent sentiment label for an identical image.

Fig. 1. Image for each emotion of Senti8PW.

TABLE I
STATISTICS OF THE CURRENT LABELED SENTI8PW.

Sentiment	Three Agree
Surprise	43
Sadness	379
Joy	45
Disgust	80
Anger	62
Fear	196
Trust	176
Anticipation	163
Total	1444

B. Proposed Deep Neural Network Model

We propose new neural network models which allow inputs coming from End-to-end (CNN) features to solve our few-shot classification problem. We first present a brief review of our network models and describe CNN-based neural networks, including C32F, C64F, C128F, VGG16, and ResNet12. Finally,

[1] https://github.com/KDE-LAB-sentimentdata/Senti8PW_dataset

we demonstrate how to use the Euclidean distance and Cosine similarity function to handle few-way 1-shot learning with our proposed structured fully connected layer. The overview of the framework for 5-way classification is shown in Figure 2.

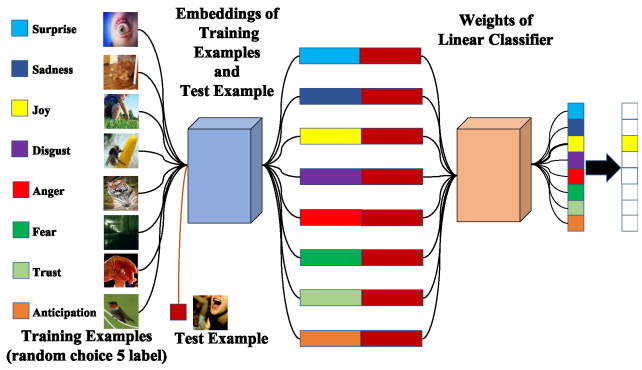

Fig. 2. Overview of 5-way image classification. The meta-training objective is to learn the parameters of a feature embedding model that generalizes well across tasks when used with regularized linear classifiers

1) Our Proposed Network Models: The overview of our proposed model for 5-way classification is shown in Figure 3. As shown in the figure, we uses two layers to model this constraint: an L2-Normalization layer and a Scale layer to model this constraint. The constraint equation is shown below:

$$\|f(x)_i\|_2 = \alpha \tag{1}$$

is equal to:

$$y = \frac{x}{\|x\|_2} \tag{2}$$

$$z = \alpha * y \tag{3}$$

which corresponds to what these two layers do. The L2 - Normalization layer normalizes the input feature vector x to a unit vector y. Then the Scale layer scales the input unit vector y to a fixed radius given by parameter alpha. There is only one scalar parameter introduced to the network, which can be trained and also fixed manually. The two modules are fully differentiable and can be integrated into an end-to-end training network. The following equations are the gradient with respect to alpha and input feature vector x.

As illustrated in Figure 4, we present a 5-way image classification in detail. We propose 4 convolutional modules, with 10×10 convolutions, followed by ReLU nonlinearity, and 2×2 max-pooling 2D, with 7×7 convolutions, followed by ReLU nonlinearity, and 2×2 max-pooling 2D, with 4×4 convolutions, followed by ReLU nonlinearity, and 2×2 max-pooling 2D, with 4×4 convolutions, followed by ReLU nonlinearity, and 2×2 max-pooling 2D. Input images of size 105×105 it yields feature maps with spatial size 3× 3. In addition, we used on flatten. Finally, the second to last convolutional layer from the last has 4096 feature channels, and the last layer has 256 feature channels.

2) CNN-Based Neural Networks: In addition to our proposed model described above, our system incorporates CNN features, which can be extracted from pre-trained deep convolutional neural networks with ImageNet [17] such as VGG16 [18], ResNet12[19]. Because prior approaches use several different network architectures to implement the feature extractor of the neural network model, we evaluate our model with each of those architectures. Specifically, the models that we evaluated are: C32F, which is a 4 module ConvNet network with 32 feature channels on each convolutional layer[20], C64F, which has 64 feature channels on each layer[20], and C128F, where the first two layers have 64 channels and the latter two have 128 channels[21]. Our applied CNN network does not retrain like VGG and ResNet.

We decrease the dimensions of the fully-connected layers used in CNN models. Specifically, for VGG16, we extract a 4096 dimensional vector from the 'fc2' layer (or the second to the last fully-connected layer), and reduce the vector to 256 dimensions by applying a fully-connected layer. Similarly, for ResNet12, we extract a 2046 dimensional vector from 'avg_pool' layer (or GlobalAveragePooling2D layer), and reduce it to 256 dimensions. Note that the output of 256 dimensions is determined empirically. This was implemented to reduce the number of parameters and unifying the dimensions.

C. Distance Metrics for Classification

We used Euclidean distance[13] and Cosine similarity[14] to detect 5-way image classification.

Algorithm 1 illustrates our proposed method. The input is a collection of features extracted from each Image data i, j including 8 kinds of sentiments (random choice N label) while the output is Feature extractor S. Our proposed network consists of two modules: an embedding module f_φ and a relation module g_ϕ. Samples x_j in the query set, and samples x_i in the sample set are fed through the embedding module f_φ, which produces feature maps $f_\varphi(x_i)$ and $f_\varphi(x_j)$. The feature maps $f_\varphi(x_i)$ and $f_\varphi(x_j)$ are combined with operator $C(f_\varphi(x_i), f_\varphi(x_j))$. In this work we assume $C(\cdot, \cdot)$ to be a concatenation of feature maps in depth, although other choices are possible.

The combined feature map of the sample and query is fed into the relation module g_ϕ, which eventually produces a scalar in range of 0 to 1 representing the similarity between x_i and x_j, which is called relation score. Thus, in the N-way 1-shot learning setting[22], we generate N relation scores for the relation between one query input x_j and training sample set examples x_i. We use compute distance such as Euclidean distance and Cosine similarity to evaluate distance. Finally, we generate feature extractor S. We repeat this computation to process all the test (unknown) images are processed.

IV. EXPERIMENTS

Here, we describe the experiments and the evaluations. We used the dataset Senti8PW for our basis, as described before. The dataset initially consisted of 2,400 images, from which

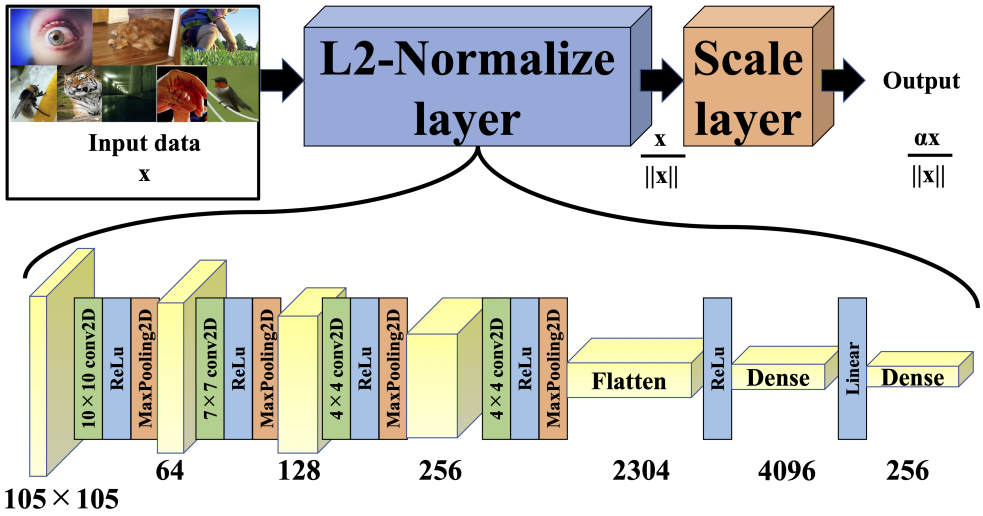

Fig. 3. Our proposed model for 5-way 1-shot image classification.

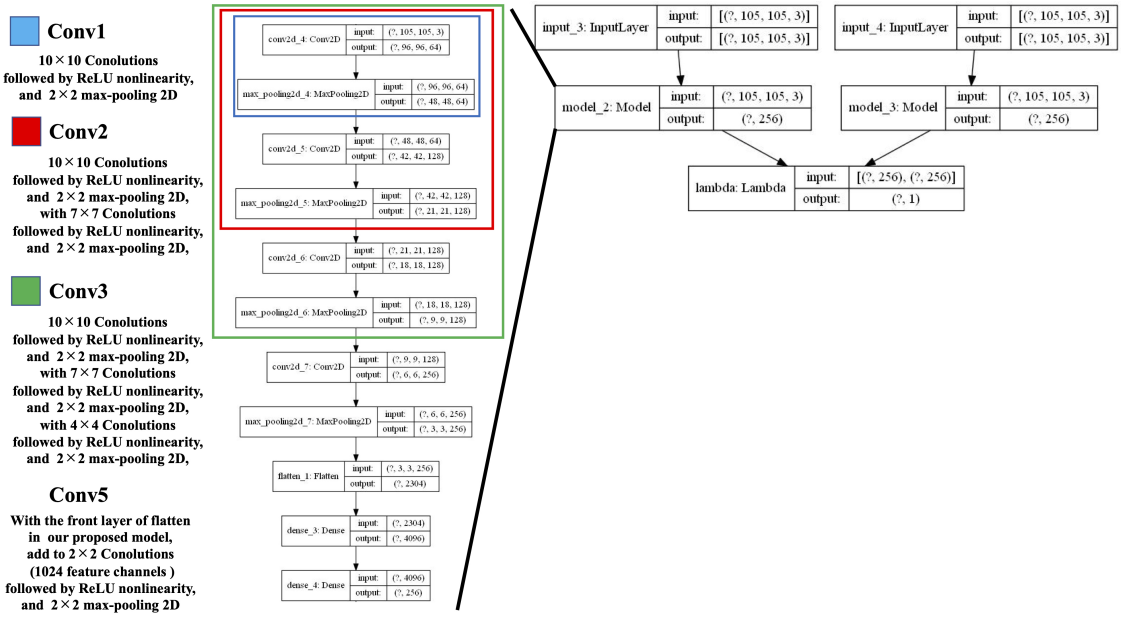

Fig. 4. Our proposed model for 5-way 1-shot image classification.

The left side is the expanded view of model_2. In addition, Conv1, Conv2, and Conv3 are highlighted by individual colors, and the explanation is shown on the left side. Conv5 added a convolutional layer to our proposed model.

we reduced the number of images in each Agree, as mention in Section III.

We have divided the reduced data into training and testing data with an 8:2 ratio. We determined the following hyper-parameters; batch size was 256, optimization function was "SGD" with a learning rate of 0.001 and momentum 0.9, and the number of epochs was 200. We employed TensorFlow[23] as our deep learning framework for the implementation.

We employed one measure of accuracy for the evaluation of few-shot classification accuracy. In the table, we compare in terms of accuracy. and include several baseline methods: 32F, 64F, 128F, VGG16, ResNet12, and our proposed model. The "Params" column of the table represents the parameters, and the "Dim" column of the table represents the feature dimension.

Table II illustrates the results, where we compare in terms of accuracy. The results show that our proposed model is the most accurate, with a Euclidean metric producing the highest accuracy.

Figure 5 illustrates accuracy and parameter with 10 models (Conv1, Conv2, Conv3, ResNet12, Conv5, VGG16, C32F, C64F, C128F, and our proposed model) in Senti8PW. The

Algorithm 1 Predict N-way 1-shot image classification

Input: Image data i, j including 8 kinds of sentiments
Output: Extracted feature S

1: Define embedding module f_φ
2: Randomly sample x_i in the query set
3: Randomly sample x_j in the sample set
4: Define the feature maps $f_\varphi(x_i)$ and $f_\varphi(x_j)$
5: Define Operator $C(f_\varphi(x_i), f_\varphi(x_j))$
6: Apply relation module g_ϕ in the combined feature map of the sample and query
7: Compute the similarity between x_i and x_j with N-way 1-shot learning
8: Feature extractor S_i, S_j=$g_\phi(C(f_\varphi(x_i), f_\varphi(x_j)))$
9: S = Compute distance(S_i, S_j)

TABLE II
THE RESULTS OF DOING EXPERIMENTS FOR 5-WAY 1-SHOT CLASSIFICATION ACCURACY IN SENTI8PW

Model	Params	Metric	Dim	Accuracy
C32F[20]	2K	Euclidean	32	0.250
C32F[20]	2K	Cosine	32	0.240
C64F[20]	3K	Euclidean	64	0.250
C64F[20]	3K	Cosine	64	0.240
C128F[21]	3,278K	Euclidean	128	0.275
C128F[21]	3,278K	Cosine	128	0.230
VGG16[18]	138,613K	Euclidean	256	0.305
VGG16[18]	138,613K	Cosine	256	0.350
ResNet12[19]	138,614K	Euclidean	256	0.235
ResNet12[19]	138,614K	Cosine	256	0.250
Our proposed Model	11,697K	Euclidean	256	**0.845**
Our proposed Model	11,697K	Cosine	256	**0.570**

figure demonstrates that our proposed model has the highest accuracy and the fourth smallest parameter.

Table III shows the results of comparing Emotion6 [24], EmotionROI [2], and Senti8PW in terms of accuracy. Senti8PW has the highest accuracy and a higher number of labels than other datasets. Thus, Senti8PW is more focused on human-specific sentiment than other datasets.

TABLE III
COMPARE WITH EMOTIONROI, SENTI8PW AND EMOTION6 OF ACCURACY IN OUR PROPOSED METHOD.

Dataset	accuracy
Emotion6 [24]	0.570
EmotionROI[2]	0.230
Senti8PW	**0.845**

V. CONCLUSIONS

We have proposed a few-shot 1-way classification framework that employs the Euclidean and Cosine Distance as the distance metric. The implicit theorem allows our network to be end-to-end trainable. Built on top of this generalized feature embedding, we can vastly improve the few-shot classification accuracy compared to the existing models. Additionally, the classification system of four layers of convolutions with 5-way 1-shot learning is the best in terms of balancing accuracy and a number of model parameters. The fully connected layer can directly classify dense representations of images in the 5-way 1-shot settings.

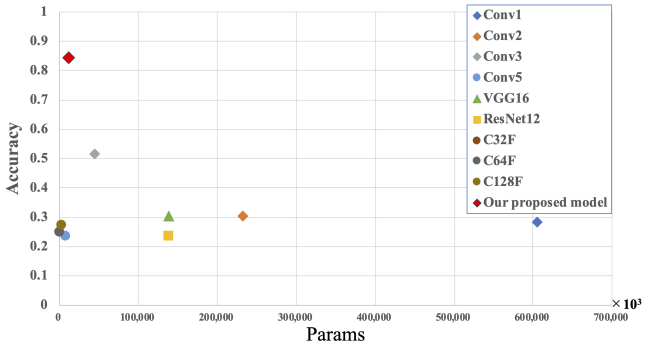

Fig. 5. The accuracy and parameter with 10 models (Conv1, Conv2, Conv3, ResNet12, Conv5, VGG16, C32F, C64F, C128F, and our proposed model).

In the future direction, we could compare our proposed methods to other related state-of-the-art models. In addition, our simple yet effective model achieves new, state-of-the-art performance on multiple datasets.

ACKNOWLEDGMENT

A part of this research was carried out with the support of the Grant-in-Aid for Scientific Research (B) (issue number 17H01746), and Grant for Education and Research in Toyohashi University of Technology.

REFERENCES

[1] Yangyu Fan, Hansen Yang, Zuhe Li, and Shu Liu. Predicting image emotion distribution by emotional region. *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–9, 2018.
[2] K. Peng, A. Sadovnik, A. Gallagher, and T. Chen. Where do emotions come from? predicting the emotion stimuli map. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 614–618, Sep. 2016.
[3] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark, 2016.
[4] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. *Proceedings of the international conference on Multimedia - MM '10*, 2010.
[5] P. Balouchian, M. Safaei, and H. Foroosh. Lucfer: A large-scale context-sensitive image dataset for deep learning of visual emotions. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1645–1654, Jan 2019.
[6] Marie Katsurai, Takahiro Ogawa, and Miki Haseyama. A cross-modal approach for extracting semantic relationships between concepts using tagged images. *IEEE Transactions on Multimedia*, 16(4):1059–1074, Jun 2014.
[7] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan S. Kankanhalli, and Qi Zhao. Emotional attention: A study of image sentiment and visual attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
[8] Macario O. Cordel, II, Shaojing Fan, Zhiqi Shen, and Mohan S. Kankanhalli. Emotion-aware human attention prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
[9] H. Kim, Y. Kim, S. J. Kim, and I. Lee. Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Transactions on Multimedia*, 20(11):2980–2992, Nov 2018.

[10] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L. Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[11] W. Zhang, X. He, and W. Lu. Exploring discriminative representations for image emotion recognition with cnns. *IEEE Transactions on Multimedia*, 22(2):515–523, Feb 2020.

[12] R. PLUTCHIK. Emotion : A general psychoevolutionary theory. *Approaches to emotion*, pages 197–219, 1984.

[13] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. *CoRR*, abs/1703.05175, 2017.

[14] Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *CoRR*, abs/1606.04080, 2016.

[15] R. Venkatesan and M. J. Er. Multi-label classification method based on extreme learning machines. In *2014 13th International Conference on Control Automation Robotics Vision (ICARCV)*, pages 619–624, Dec 2014.

[16] USA. [Online] San Francisco, CA. flicker. http://www. flickr.com., 2017.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[19] Boris N. Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. TADAM: task dependent adaptive metric for improved few-shot learning. *CoRR*, abs/1805.10123, 2018.

[20] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. *CoRR*, abs/1804.09458, 2018.

[21] Yann Lifchitz, Yannis Avrithis, Sylvaine Picard, and Andrei Bursuc. Dense classification and implanting for few-shot learning. *CoRR*, abs/1903.05050, 2019.

[22] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. *CoRR*, abs/1711.06025, 2017.

[23] Google. Tensorflow. https://github.com/tensorflow.

[24] K. Peng, T. Chen, A. Sadovnik, and A. Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868, June 2015.