Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark

Quanzeng You and **Jiebo Luo**
Department of Computer Science
University of Rochester
Rochester, NY 14623
{qyou, jluo}@cs.rochester.edu

Hailin Jin
Adobe Research
345 Park Avenue
San Jose, CA 95110
hljin@adobe.com

Jianchao Yang
Snapchat Inc
64 Market St
Venice, CA 90291
jianchao.yang@snapchat.com

Abstract

Psychological research results have confirmed that people can have different emotional reactions to different visual stimuli. Several papers have been published on the problem of visual emotion analysis. In particular, attempts have been made to analyze and predict people's emotional reaction towards images. To this end, different kinds of hand-tuned features are proposed. The results reported on several carefully selected and labeled small image data sets have confirmed the promise of such features. While the recent successes of many computer vision related tasks are due to the adoption of Convolutional Neural Networks (CNNs), visual emotion analysis has not achieved the same level of success. This may be primarily due to the unavailability of confidently labeled and relatively large image data sets for visual emotion analysis. In this work, we introduce a new data set, which started from 3+ million weakly labeled images of different emotions and ended up 30 times as large as the current largest publicly available visual emotion data set. We hope that this data set encourages further research on visual emotion analysis. We also perform extensive benchmarking analyses on this large data set using the state of the art methods including CNNs.

Introduction

Psychological studies have provided evidence that human emotions can be aroused by visual content, e.g. images (**?**; **?**; **?**). Based on these findings, recently computer scientists also started to delve into this research topic. However, differently from psychological studies, which mainly focus on studying the changes between physiological and psychological activities of human beings on visual stimuli, most of the works in computer science are trying to predict the aroused human emotion given a particular piece of visual content. Indeed, affective computing, which *aims to recognize, interpret and process human affects* (http://en.wikipedia.org/wiki/Affective_computing), has achieved significant progress in recent years. However, the problem of visual emotion prediction is more difficult in that we are trying to predict the emotional reactions given a general visual stimulus, instead of using the collected signals from human's physiological reactions of visual stimuli as studied in affective computing.

Figure 1: Example images of eight different emotion categories. *Top Row*: four positive emotions and *Bottom Row*: four negative emotions.

The increasing popularity of social networks attracts more and more people to publish multimedia content in online social network platforms. Online users can easily add textual data, e.g., title, descriptions, tags, to their uploaded images and videos. However, these textual information can only help the retrieval of multimedia content in the cognitive level (**?**), i.e., semantic level. The meta text data has limited help in bridging the *affective* semantic gap between images pixels and human feelings. In (**?**), the authors call visual emotion prediction *affective content analysis*.

Inspired by the psychology and art theory, different groups of manually crafted features are designed to study the emotional reactions towards visual content. For example, based on art theory (**?**; **?**), Machajdik and Hanbury (**?**) defined eight different kinds of pixel level features (e.g. color, texture and composition), which have been empirically proved to be related to emotional reactions. In another recent work (**?**), *principles-or-art* based features are extracted to classify emotions. Following their works, we study the same eight emotions, *Amusement*, *Awe*, *Contentment*, *Excitement*, *Anger*, *Disgust*, *Fear* and *Sadness*. Figure 1 shows the example images for these studied emotions. All these images are selected from the newly constructed

data set in this work, where each image is labeled by five Amazon Mechanical Turk workers.

Recently deep learning has enabled robust and accurate feature learning, which in turn produces the state-of-the-art performance on many computer vision related tasks, e.g. digit recognition (**?**; **?**), image classification (**?**; **?**), aesthetics estimation (**?**) and scene recognition (**?**). One of the main factors that prompt the success of deep learning on these problems is the availability of a large scale data set. From ImageNet (**?**) to AVA dataset (**?**) and the very recent Places Database (**?**), the availability of these data sets have significantly promoted the development of new algorithms on these research areas. In visual emotion analysis, there is no such a large data set with strong labels. More recently, You et al. (**?**) employed CNNs to address visual sentiment analysis, which tries to bridge the high-level, abstract sentiments concept and the image pixels. They employed the *weakly* label images to train their CNN model. However, they are trying to solve a binary classification problem instead of a multi-class (8 emotions) problem as studied in this work.

In this work, we are interested in investigating the possibility of solving the challenging visual emotion analysis problem. First, we build a large scale emotion data set. On top of the data set, we intend to find out whether or not applying CNNs to visual emotion analysis provides advantages over using a predefined collection of art and psychology theory inspired visual features or visual attributes, which have been done in prior works. To that end, we make the following contributions in this work.

- We collect a large number of weakly labeled emotion related images. Next, we employ Amazon Mechanical Turk to manually label these images to obtain a relatively strongly labeled image data set, which makes the usage of CNN for visual emotion analysis possible. All the data set will be released to the research community upon publishing this work.
- We evaluate the performance of Convolutional Neural Networks on visual emotion analysis and establish it as the baseline for future research. Compared with the state-of-the-art manually crafted visual features, our results suggest that using CNN can achieve significant performance improvement on visual emotion analysis.

Related Work

Our work is mostly related to both visual emotion analysis and Convolutional Neural Networks (CNNs). Recently, deep learning has achieved massive success on a wide range of artificial intelligence tasks. In particular, deep Convolutional Neural Networks have been widely employed to solve traditional computer vision related problems. Deep convolutional neural networks typically consist of several convolutional layers and several fully connected layers. Between the convolutional layers, there may also be pooling layers and normalization layers. In early studies, CNNs (**?**) have been very successful in document recognition, where the inputs are relatively small images. Thanks to the increasing computational power of GPU, it is now possible to train a deep convolutional neural network on large collections of images (e.g. (**?**)), to solve other computer vision problems, such as scene parsing (**?**), feature learning (**?**), visual recognition (**?**) and image classification (**?**).

However, to the best of our knowledge, there are no related works on using CNNs for visual emotion analysis. Currently, most of the works on visual emotion analysis can be classified into either dimensional approach (**?**; **?**) or categorical approach (**?**; **?**; **?**), where the former represents emotion in a continuous two dimensional space and in the later model each emotion is a distinct class. We focus on the categorical approach, which has been studied in several previous works. Jia et al. (**?**) extract *color* features from the images. With additional social relationships, they build a factor graph model for the prediction of emotions. Inspired by art and psychology theory, Machajdik and Hanbury (**?**) proposed richer hand-tuned features, including *color*, *texture*, *composition* and *content* features. Furthermore, by exploring the principles of art, Zhao et al. (**?**) defined more robust and invariant visual features, such as *balance*, *variety*, and *gradation*. Their features achieved the best reported performance on several publicly accessible emotion data sets.

Those hand-tuned visual features have been validated on several publicly available small data sets (see following sections for details). However, we want to verify whether or not deep learning could be applied to this challenging problem and more importantly, on a much larger scale image set. The main issue is that there are no available well labeled data sets for training deep neural networks. Our work intends to provide such a data set for the research community and verify the performance of the widely used deep convolutional neural architecture on this emotion data set.

Visual Emotion Data Sets

Several small data sets have been have been used for visual emotion analysis (**?**; **?**; **?**), including (1) **IAPS-Subset:** This data set is a subset of the International Affective Picture System (IAPS) (**?**). This data set is categorized into eight emotional categories as shown in Figure 1 in a study conducted in (**?**). (2) **ArtPhoto:** Machajdik and Hanbury (**?**) built this data set, which contains photos by professional artists. They obtain the ground truth by the labels provided by the owner of each image. (3) **Abstract Paintings:** These are images consisting of both color and texture from (**?**). They obtain the ground truth of each image by asking people to vote for the emotions of each image in the given eight emotion categories. Table 1 shows the statistics of the existing three data sets. The numbers show that each data set only consists of a very small number of images. Meanwhile, images in all the three different data sets are highly imbalanced. All three data sets are relatively small with images coming from a few specific domains. In particular, for some categories, such as *Anger*, there are less than 10 images. Therefore, if we employ the same methodology (5-fold Cross Validation within each data set) (**?**; **?**), we may have only several images in the training data. This may lead to the possibility that their trained models may have been either over or under fitted.

Data Set	Amusement	Anger	Awe	Contentment	Disgust	Excitement	Fear	Sadness	Sum
IAPS-Subset	37	8	54	63	74	55	42	62	395
ArtPhoto	101	77	102	70	70	105	115	166	806
Abstract Paintings	25	3	15	63	18	36	36	32	228
In Total	163	88	171	196	162	196	193	260	1429

Table 1: Statistics of the three existing data sets. The three data sets are imbalanced across the 8 categories.

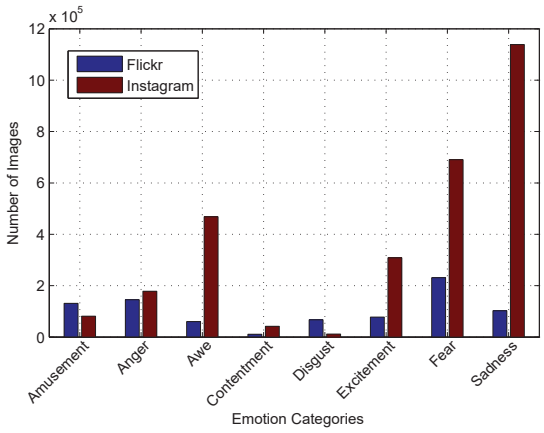


Figure 2: Statistics of the images downloaded from Flickr and Instagram.

The above results suggest that the previous efforts on visual emotion analysis deal with small emotion-centric data sets compared with other vision data sets, such as ImageNet (**?**) and Places (**?**). In this work, we present here a new emotion data set, which is by far the largest available emotion-centric database.

Building An Image Emotion Dataset from the Wild

There are many different emotion definition systems[1] from psychological and cognitive science. In this work, we use the same eight emotions defined in Table 1, which is derived from a psychological study in (**?**). Using the similar approach in (**?**), we query the image search engines (Flickr and Instagram) using the eight emotions as keywords. In this way, we are able to collect a total of over 3 million *weakly* labeled images, i.e., labeled by the queries. Next, we delete images which have tags of any two different emotions. We also remove duplicate images using *fdupes*[2]. Figure 2 shows the statistics of the remaining images. As we can see, the number of images in different categories is imbalanced. In particular, there are only small numbers of *contentment* and *disgust* images in both social media platforms. Meanwhile, the number of per category images from Instagram varies significantly. There are much more images from both *Fear* and *Sadness*. This agrees with the finding (http://goo.gl/vhBBF6) that people are likely to share sadness from their Instagram accounts.

[1]http://en.wikipedia.org/wiki/Emotion

[2]https://code.google.com/p/fdupes/

Next, we employ Amazon Mechanical Turk (AMT) to further label these *weakly* labeled images. In particular, we design a qualification test to filter all workers who want to work on our tasks. The qualification test is designed as an image annotation problem. We randomly select images from the publicly available *ArtPhoto* data set and use the ground-truth labels as the answers. For each given image, we ask the workers to choose the emotion they feel from the eight emotion categories. At first, we conduct experiments within members in our research group. Indeed, the results suggest that this qualification is challenging difficult, in particular when you have to choose only one emotion for each image. Therefore, we design our AMT tasks (HITs) as a much easier verification task instead of the annotation task. Since we have crawled all the images with emotion queries, we have a *weakly* labeled data set. In each HIT, we assign *five* AMT workers to verify the emotion of each image. For each given image and its *weak* label, we ask them to answer a question like *Do you feel anger seeing this image?* The workers are asked to choose a YES or NO for each image. All workers have to meet the rigorous requirement of correctly answering at least half of the questions (a total of 20) in our qualification test. By the time of finishing work, we have over 1000 workers on our qualification task. Among them, 225 workers meet our qualification criteria.

To start the verification task, we randomly select $11,000$ images for each emotion category. After collecting the batch results from AMT, we keep those images which receive at least three Yeses from their assigned fiver AMT workers. In this way, we are able to build a relatively strongly labeled data set for visual emotion analysis. Table 2 summarizes the number of images for our current data set. The numbers show that different categories may have different acceptance rates for workers to reject or accept the *positive* samples of each verification task. In particular, we add another $2,000$ images to make sure that the number of images in *Fear* category is also larger than $1,000$ images. In total, we collect about $23,000$ images, which is about 30 times as large as the current largest emotion data set, i.e., *ArtPhoto*.

Fine-tuning Convolutional Neural Network for Visual Emotion Analysis

Convolutional Neural Networks (CNN) have been proven to be effective in image classification tasks, e.g., achieving the state-of-the-art performance in ImageNet Challenge (**?**). Meanwhile, there are also successful applications by fine-tuning the pre-trained ImageNet model, including recognizing image style (**?**) and semantic segmentation (**?**). In this work, we employ the same strategy to fine-tune the pre-

Data Set	Amusement	Anger	Awe	Contentment	Disgust	Excitement	Fear	Sadness	Sum
Submitted	11,000	11,000	11,000	11,000	11,000	11,000	13,000	11,000	90,000
Labeled	4,942	1,266	3,151	5,374	1,658	2,963	1,032	2,922	23,308

Table 2: Statistics of the current labeled image data set. Note that all emotion categories have 1000+ images.

trained ImageNet reference network (**?**). The same neural network architecture is employed. We only change the last layer of the neural network from 1000 to 8. The remain layers keep the same as the ImageNet reference network. We randomly split the collected $23,000$ samples into training (80%), testing (15%) and validating sets (5%).

Meanwhile, we also employ the *weak* labels (**?**) to fine-tune another model as described in (**?**). We exclude those images that have been chosen to be submitted to AMT for labeling. Next, since *contentment* contains only about $16,000$ images, we randomly select $20,000$ images for other emotion categories. In this way, we have a total of $156,000$ images. We call this model *Noisy-Fine-tuned CNN*. We fine-tune both models using Caffe with a Linux server with 2 NVIDIA TITAN GPUs on top of the pre-trained ImageNet CNN model.

Performance of Convolutional Neural Networks on Visual Emotion Analysis

After the fine-tuning of the pre-trained CNN model, we obtain two new CNN models. To compare with the ImageNet-CNN, we also show the results of using the SVM trained on features extracted from the second to the last layer of the pre-trained ImageNet-CNN model. In particular, we employ PCA to reduce the dimensionality of the features. We also try several different numbers of principal components. The results are almost the same. To overcome the imbalance problem in the data, we adjust the weights of SVM for different classes (in our implementation, we use LIBSVM[3], which provides such a mechanism). Table 3 summarizes the performance of the three groups of features on the 15% randomly chosen testing data. The overall accuracy of the Fine-tuned-CNN is almost 60%. As a baseline, the visual features extracted from ImageNet-CNN only lead to an overall accuracy of about 30%, which is half of Fine-tuned-CNN. The Noisy-Fine-tuned-CNN model has an overall accuracy of about 46%, which suggests that this model can learn some knowledge from the noisily labeled images. However, even though it has much more training samples compared with *Fine-tuned CNN*, it fails to outperform *Fine-tuned CNN*, which is trained on strongly labeled samples.

We also calculate the confusion matrix of the three algorithms from their prediction results on the testing data to further analyze their performance. Figure 3(c) shows the confusion matrix of the Fine-tuned CNN model. Compared with the other two models, the true negative rates from *Fine-tuned-CNN* are the best in most emotion categories. Meanwhile, the confusion matrix of *Noisy-Fine-tuned CNN* seems to be more balanced, except for the *contentment* emotion

[3]http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

Algorithms	Correct Samples	Accuracy
ImageNet-CNN	1120/3490	32.1%
Noisy-Fine-tuned-CNN	1600/3490	45.8%
Fine-tuned-CNN	2034/3490	58.3%

Table 3: Classification accuracy on the 15% randomly selected testing set labeled by the Amazon Mechanical Turk.

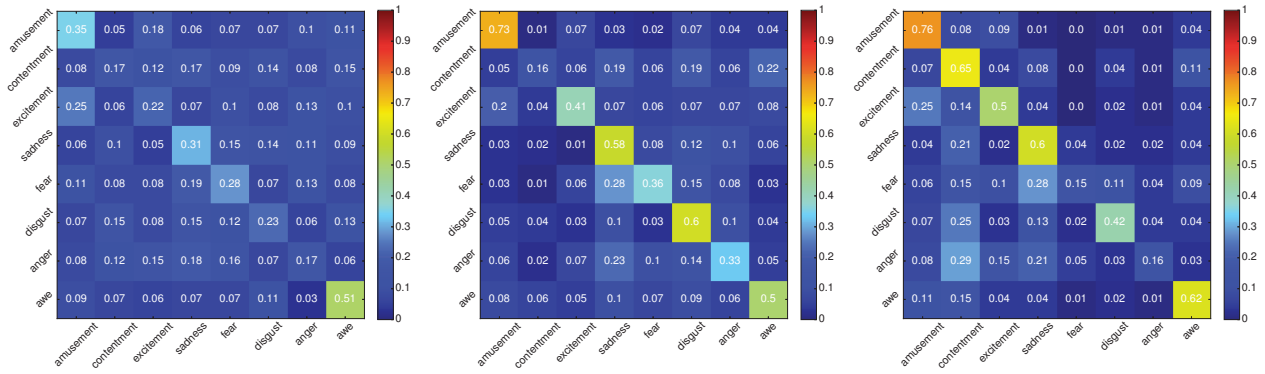
(seeFigure 3(b)). Indeed, these findings are consistent with the number of available labeled samples (see Table 2). The more the labeled images, the higher probability that the corresponding emotion will receive a higher true positive rate. Figure 3(a) shows the confusion matrix using the more general ImageNet-CNN features. It is interesting to see that overall the performance is worse than the Fine-tuned CNN features. However, the true positive rate of *fear* is higher than that of using the Fine-tuned features.

The embedding of the testing images using deep visual features (we do not show the embedding for Noisy-Fine-tuned-CNN due to space arrangement) is shown in Figure 4. The features are also processed using t-SNE (**?**). The embedding using ImageNet-CNN shows that images from the same scene or of similar objects are embedded into neighboring areas. However, the embedding using Figure 4(b) seems to make the images more diverse in terms of objects or scenes. This is indeed comply with the fact that even the same object could lead to different visual emotion at its different state, e.g., angry dog and cute dog.

Performance of Convolutional Neural Networks on Public Existing Visual Emotion Data Set

We have described several existing data sets in Section . Table 1 summarizes the statistics of the three data sets. To the best of our knowledge, no related studies have been conducted on evaluating the performance of Convolutional Neural Networks on visual emotion analysis. In this section, we evaluate all the three deep neural network models on all the three data sets and compare the results with several other state-of-the-art methods on these data sets.

In particular, we extract deep visual features for all the images in the three data sets using the trained deep neural network models from the second to the last layer. In this way, we obtain a 4096 dimensional feature representation for each image from each deep model. Next, we follow the same evaluation routine described in (**?**) and (**?**). At first, PCA is employed to reduce the dimensions of the features respectively. For all the three data sets, we reduce the number of feature dimensions from 4096 to 20, which is capable of keeping at least 90% variance. Next, a linear SVM is trained on the reduced feature space. Following the same ex-



(a) ImageNet-CNN (avg: 0.28) (b) Noisy-Fine-tuned-CNN (avg: 0.459) (c) Fine-tuned-CNN (avg: 0.483)

Figure 3: Confusion matrix for ImageNet-CNN, Noisy-Fine-tuned-CNN and Fine-tuned-CNN on the testing Amazon Mechanical Turk (AMT) labeled images.

(a) Embedding of the *testing* images using general ImageNet-CNN features (b) Embedding of the *testing* images using Fine-tuned-CNN features.

Figure 4: Visualization of learned filters for both ImageNet-CNN and the fine-tuned CNN (best viewed on screen with zoom).

perimental approach, the *one v.s. all* strategy is employed to train the classifier. In particular, we randomly split the data into 5 batches such that 5-fold Cross Validation is used to obtain the results. Also, we assign larger penalties to *true negative* samples in the SVM training stage in order to optimize the *per class true positive rate* as suggested by both (**?**) and (**?**).

We compare the performance of deep features on visual emotion analysis with several other baseline features, including Wang et al. (**?**), Yanulevskaya et al. (**?**), Machajdik and Hanbury (**?**) and Zhao et al. (**?**). Figures 5, 6 and 7 show the performance of these features on the three data sets respectively. Note that since emotion *anger* only contains 8 and 3 images in IAPS-Subset and Abstract Paintings data sets, which are not enough to perform the 5-fold Cross Validation. We do not report the true positive rates for emotion *anger* on these two data sets.

It is interesting to find out that deep visual features significantly outperform the state-of-the-art manually crafted visual features in some emotion categories. However, the performance of using deep visual features are not consistent across the emotion categories at present. In particular, the performance of directly employing deep visual features from ImageNet-CNN and Noisy-Fine-tuned-CNN differ significantly among categories as well as across data sets. The performance of deep visual features from Fine-tuned-CNN is relatively more consistent. However, it has poor performance on emotions *Contentment* and *Fear* in the ArtPhoto data. These results suggest that it is still challenging to solve visual emotion analysis even with the state-of-the-art deep visual features. Meanwhile, the performance of deep visual features also suggests the promise of using CNNs in visual

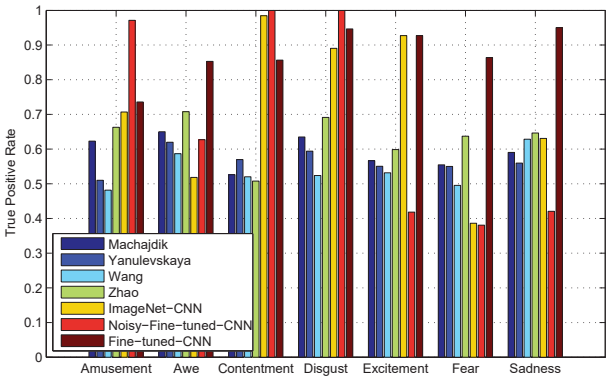


Figure 5: Per-class true positive rates of *Machajdik* (**?**), *Yanulevskaya* (**?**), *Wang* (**?**), *Zhao* (**?**), *ImageNet-CNN*, *Noisy-Fine-tuned-CNN* and *Fine-tuned-CNN* on IAPS-Subset data set.

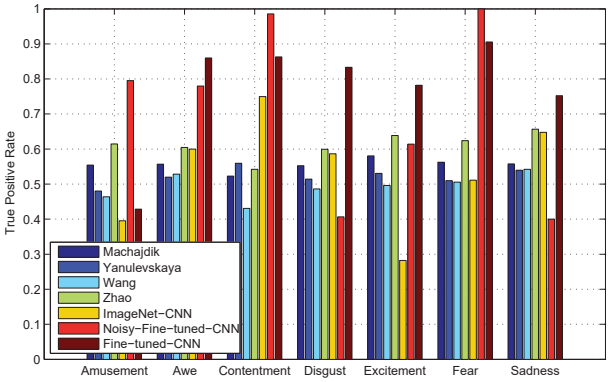


Figure 6: Per-class true positive rates of *Machajdik* (**?**), *Yanulevskaya* (**?**), *Wang* (**?**), *Zhao* (**?**), *ImageNet-CNN*, *Noisy-Fine-tuned-CNN* and *Fine-tuned-CNN* on Abstract Paintings data set.

emotion analysis. Overall, this may encourage the development of more advanced deep architectures for visual emotion analysis, as well as development of other approaches.

Conclusions

In this work, we introduce the challenging problem of visual emotion analysis. Due to the unavailability of a large scale well labeled data set, little research work has been published on studying the impact of Convolutional Neural Networks on visual emotion analysis. In this work, we are introducing such a data set and intend to release the data set to the research community to promote the research on visual emotion analysis with the deep learning and other learning frameworks. Meanwhile, we also evaluate the deep visual features extracted from differently trained neural network models. Our experimental results suggest that deep convolutional neural network features outperform the state-of-the-art hand-tuned features for visual emotion analysis. In addition, fine-tuned neural network on emotion related data

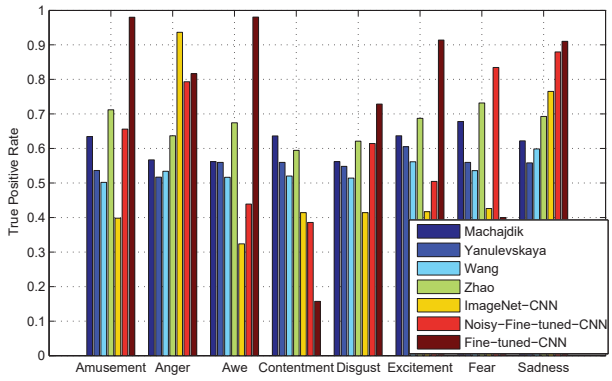


Figure 7: Per-class true positive rates of *Machajdik* (**?**), *Yanulevskaya* (**?**), *Wang* (**?**), *Zhao* (**?**), *ImageNet-CNN*, *Noisy-Fine-tuned-CNN* and *Fine-tuned-CNN* on ArtPhoto data set.

sets can further improve the performance of deep neural network. Nevertheless, the results obtained in this work are only a start for the research on employing deep learning or other learning frameworks for visual emotion analysis. We will continue the collection of labeled data from AMT with a plan to submit additional 1 million images for labeling. We hope our visual emotion analysis results can encourage further research on online user generated multimedia content in the wild. Better understanding the relationship between emotion arousals and visual stimuli and further extending the understanding to valence are the primary future directions for visual emotion analysis.

Acknowledgement

This work was generously supported in part by Adobe Research, and New York State CoE IDS. We thank the authors of (**?**) for providing their algorithms for comparison.

Culpa ex numquam sed velit, iusto iure tenetur eos magni magnam ipsa culpa nemo ab, earum id excepturi quidem tempore magni illo, repellat id mollitia incidunt tempora dolorum magnam maxime magni perspiciatis quos, vitae cumque earum consequuntur odio quis magni id cum asperiores facilis.Et deleniti eum hic, possimus reiciendis delectus illo numquam magnam maiores adipisci, laudantium nam odio et ullam accusamus, perferendis similique itaque veritatis earum quia, necessitatibus impedit ad soluta nisi quo ipsam architecto tempore.Voluptate ad repellat distinctio nulla tempora ratione in repellendus sunt ducimus ipsa, officiis odio neque qui itaque consectetur accusamus repellat sequi, qui aliquid atque assumenda minus laboriosam deserunt at nemo quibusdam officiis doloribus, nihil iure sequi magnam mollitia hic commodi, autem sunt error sint animi quam nobis impedit consectetur voluptas.Quaerat ex consectetur magnam veniam culpa ipsum, recusandae excepturi quod tempore doloribus, eum quam minus maxime quisquam eligendi eaque fuga quas consequuntur iste.Assumenda quis magnam itaque adipisci fuga sunt, dignissimos nulla sint dolores assumenda ratione consectetur ipsam inventore, ullam ea quidem soluta laudantium, vel perferendis doloribus voluptatibus magni in eaque voluptas architecto laborum

nulla.Explicabo minima rem fugiat excepturi cum vero, asperiores inventore cumque, unde eveniet cumque veniam libero corporis odio ratione vitae