Contemplating Visual Emotions: Understanding and Overcoming Dataset Bias

Anonymous ECCV submission

Paper ID 2418

Abstract. While machine learning approaches to visual emotion recognition offer great promise, current methods risk training and testing models on small scale datasets covering limited visual emotion concepts. Our analysis identifies an important but long overlooked issue of existing visual emotion benchmarks in the form of dataset biases. We design a series of tests to show and measure how such dataset biases obstruct learning a generalizable emotion recognition model. Based on our analysis, we propose a webly supervised approach by leveraging a large quantity of stock image data. Our approach uses a simple yet effective curriculum guided training strategy for learning discriminative emotion features. We discover that the models learned using our large scale stock image dataset exhibit significantly better generalization ability than the existing datasets without the manual collection of even a single label. Moreover, visual representation learned using our approach holds a lot of promise across a variety of tasks on different image and video datasets.

Keywords: Emotion Recognition, Webly Supervised Learning

1 Introduction

Recently, algorithms for object recognition and related tasks have become sufficiently proficient that new vision tasks beyond objects can now be pursued. One such task is to *recognize emotions expressed by images* which has gained momentum in last couple of years in both academia [1–6] and industries, e.g., Apple's recent acquisition of Emotient startup[1]. Teaching machines to recognize diverse emotions is a very challenging problem with great application potential.

Let us consider the image shown in Figure 1.a. Can you recognize the basic emotion expressed by this image? Practically, this should not be a difficult task as a quick glance can well reveal that the overall emotional impact of the image is negative (i.e., sadness) (9 out of 10 students in our lab made it correct!). In fact, this is the image of an Six Flags theme park at New Orleans which has been closed since Hurricane Katrina struck the state of Louisiana in August 2005. [2]

Intrigued, we decided to perform a toy experiment using Convolutional Neural Networks (CNNs) to recognize emotions. A ResNet-50 [7] model that we

[1] `https://goo.gl/nnzJ5E`

[2] The image is taken from Google Images with the search keyword *sad amusement park*. Source: `https://goo.gl/AUwoPZ`

Fig. 1. (a) An example image of an amusement park with negative emotion (sadness) (Source: Google Images). (b)-(c) Nearest neighbor images extracted from "amusement" and "sadness" category in the Deep Emotion dataset [1], which show a strong data bias. We use the pool5 features from our ResNet-50 trained on Deep Emotion dataset to extract these nearest neighbor images.

trained on the current largest Deep Emotion dataset [1] predicts an emotion of "amusement/joy" with 99.9% confidence from the image in Figure 1.a. Why is this happening? Our initial investigation with the nearest neighbour images in Figure 1.b/c show that the dataset bias appears to be the main culprit. Specifically, the Deep Emotion dataset [1] suffers from two types of biases. The first is the positive set bias, which makes the *amusement* category in the dataset full of photos of amusement parks (see Figure 1.b). This is due to the lack of diversity in visual concepts when collecting the source images. The second is the negative set bias, where the rest of the dataset does not well represent the rest of the world, i.e., no images of sad park in the dataset (see Figure 1.c).

In this paper, instead of focusing on beating the latest benchmark numbers on the latest dataset, we take a step back and pose an important question: *how well do the existing datasets stack up overall in the emerging field of visual emotion recognition?* We first conduct a series of tests including a novel correlation analysis between emotion and object/scene categories to analyze the presence of bias in existing benchmarks. We then present a number of possible remedies, mainly proposing a new weakly-labeled large-scale emotion dataset collected from a stock website and a simple yet effective curriculum guided training strategy for learning discriminative features. Our systematic analysis, which is first in emotion recognition, will provide insights to the researchers working in this area to focus on the right training/testing protocols and more broadly simulate discussions in the community regarding this very important but largely neglected issue of dataset bias in emotion recognition. We also hope our efforts in releasing several emotion benchmarks in this work will open up avenues for facilitating progress in this emerging area of computer vision.

The key takeaways from this paper can be summarized as follows:

- **Existing visual emotion datasets appear to have significant bias.** We conduct extensive studies and experiments for analyzing emotion recognition datasets (Sec. 3). Our analysis reveals the presence of significant biases in current benchmark datasets and calls for rethinking the current methodology for training and testing emotion recognition models.
- **Learning with large amounts of web data helps to alleviate (at least minimize) the effect of dataset bias.** We demonstrate that models learned using large-scale stock data exhibit significantly better generalization ability while testing on new unseen datasets (Sec. 5.1). We further propose a simple yet effective curriculum guided training strategy (Sec. 4) for learning

discriminative emotion features that achieves state-of-the-art performance on various tasks across different image and video datasets (Sec. 5.2). E.g., we show improved performance (∼3% in top-5 mAP) of a state-of-the-art video summarization algorithm [8] by just plugging in our emotion features.
– **New Datasets.** We introduce multiple image emotion datasets collected from different sources for model training and testing. Our stock image dataset is one of the largest in the area of visual emotion analysis containing about 268,000 high quality stock photos across 25 fine-grained emotion categories. All our datasets and models are publicly available on our project website. [3]

2 Related Work

Emotion Wheels. Various types of emotion wheels have been studied in psychology, e.g., Ekmans emotions [9] and Plutchik's emotions [10]. Our work is based on the popular Parrott's wheel of emotions [11] which organizes emotions in the form of a tree with primary, secondary and tertiary emotions. This hierarchical grouping is more interpretable and can potentially help to learn a better recognition model by leveraging the structure.

Image Emotion Recognition. A number of prior works studying visual emotion recognition focus on analyzing facial expressions [12–17, 13, 18]. Specifically, these works mainly predict emotions for images that involve a clear background with people as the primary subject. Predicting emotions from user-generated videos [19–21], social media images [22, 21, 23] and artistic photos [24, 25] are also some recent trends in emotion recognition. While these approaches have obtained reasonable performance on such controlled emotion datasets, they have not yet considered predicting emotions from natural images as discussed in this paper. Most related to our work along the direction of recognizing emotions from natural images are the works of [1, 26, 2, 4] which predict emotions from images crawled from Flickr and Instagram. As an example, You et al. [1] learn a CNN model by supervised training to recognize emotions in natural images and performs reasonably well on the Deep Emotion dataset [1]. However, it requires expensive human annotation and is difficult to scale up to cover the diverse emotion concepts. Instead, we focus on webly supervised learning of CNNs which can potentially avoid (at least minimize) the dataset design biases by utilizing vast amount of weakly labeled data from diverse concepts.

Webly Supervised Learning. There is a continued interest in the vision community on learning recognition models directly from web data since images in web can cover a wide variety of visual concepts and, more importantly, can be used to learn computational models without using instance-level human annotations [27–36]. While the existing works have shown advantages of using web data by either manually cleaning the data or developing a specific mechanisms for reducing the noise level, we demonstrate that noisy web data can be surprisingly effective with a curriculum guided learning strategy for recognizing fine-grained emotions from natural images.

Curriculum Learning. Our work is related to curriculum learning [37–42] that learns a model by gradually including easy to complex samples in training so as to

[3] To be added after acceptance --

Fig. 2. (a) Confusion matrix, (b) From top to bottom, depicted are examples of high confident correct predictions from Deep Sentiment, Deep Emotion and Emotion-6 datasets respectively.

increase the entropy of training samples. However, unlike these prior works that typically focus on the evolution of the input training data, our approach focuses on the evolution of the output domain, i.e., evolution of emotion categories from being easy to difficult in prediction.

Hierarchical Recognition. Category hierarchies have been successfully leveraged in several recognition tasks: image classification [43–48], object detection [49, 50], image annotation [51], and concept learning [52] (see [53] for an overview). CNN based methods [54, 43, 44, 55] have also used class hierarchy for large scale image classification. Unlike these methods that mostly use clean manually labeled datasets to learn the hierarchy, we adopt an emotion hierarchy from psychology [11] to guide the learning with noisy web data. Our basic idea is that the emotion hierarchy can provide guidance for learning more difficult tasks in a sequential manner and also provide regularization for label noises.

3 Understanding Bias in Emotion Datasets

Goal. Our main goal in this section is to identify, show and measure dataset bias in existing emotion recogntion datasets using a series of tests.

Datasets. We pick three representative datasets including one newly created by us: (1) Deep Sentiment [5] dataset containing 1269 images from Twitter, (2) the current largest Deep Emotion dataset [1], (3) our Emotion-6 dataset of 8350 images (*anger*: 1604, *fear*: 1280, *joy*: 1964, *love*: 724, *sadness*: 2221, *surprise*: 557) labeled by five human subjects from intially 150K images collected from Google and Flickr (see supp). Our main motivation on creating Emotion-6 dataset is to repeat the standard data collection/annotation protocol used by existing works [1, 5] and see how well it performs regarding the dataset biases.

Test 1. Name That Dataset Game. With the aim of getting an initial idea on the relation among different datasets, we start our analysis by running *Name That Dataset Game*, as in [56–58]. We randomly sample 500 images from the training portions of each of the three datasets and train a 3-class linear classifier over the ResNet-50 features. The classifier is tested on 100 random images from each of the test sets. We observe that the classifier is reasonably good at telling different datasets apart, giving 63.67% performance. The distinct diagonal in confusion matrix (Figure 2.a) shows that these datasets possesses an unique signature leading to the presence of bias. E.g., visually examining the high confident correct predictions from the test set in Figure 2.b indicate that Deep Emotion

Table 1. Binary Cross-Dataset Generalization. Diagonal numbers refer to training and testing on same dataset while non-diagonal numbers refer to training on one dataset and testing on others. % Drop refers to the performance drop across the diagonal and the average of non-digonal numbers.

Train on: \ Test on:	Deep Sentiment	Deep Emotion	Emotion-6	% Drop
Deep Sentiment	78.74	68.38	49.76	**24.98**
Deep Emotion	61.41	84.81	69.22	**22.99**
Emotion-6	54.33	64.28	77.72	**23.69**

dataset has a strong preference for outdoor scenes mostly focusing on parks (2nd row), while Emotion-6 tend to be biased toward images where a single object is centered with a clean background and a canonical viewpoint (3rd row).

Test 2. Binary Cross-Dataset Generalization. Given all three datasets, we train a ResNet-50 classifier to show cross-dataset generalization i.e., training on one dataset while testing on the other. For both Deep Emotion and Emotion-6, we randomly sample 80% of images for training and keep rest 20% for testing, while on Deep Sentiment, we use 90% of images for the training and keep the rest for testing, as in [5]. Since, exact emotion categories can vary from one dataset to another, we report binary classification accuracies (positive vs negative) which are computed by tranforming the predicted labels to two basic emotion categories, following Parrott's emotional grouping [11]. We call this *Binary Cross-Dataset Generalization Test*, as it asks the CNN model to predict the most trivial basic emotion category from an image. If a model cannot generalize well in this simple test, it will not work on more fine-grained emotion categories. Moreover, the binary generalization test only involves minimum post-processing of the model predictions, so it can evaluate different datasets more fairly.

Table 1 shows a summary of results. From Table 1, the following observations can be made: (1) As expected, training and testing on the same dataset provides the best performance on all cases (marked in red). (2) Training on one dataset and testing on the other shows a significant drop in accuracy, for instance, the classifier trained on Deep Emotion dataset shows a average drop of 22.99% in accuracy while testing on other two datasets. Why is this happening? Our observations suggest that the answer lies in the emotion dataset itself: it's size is relatively small, which results in the positive set bias due to the lack of diversity in visual concepts. As a result, models learned using such data essentially memorize all it's idiosyncrasies and lose the ability to generalize.

Test 3. Quantifying Negative Bias. We choose three common emotion categories across Deep Emotion and Emotion-6 datasets (*anger*, *fear* and *sadness*) to measure negative set bias in different datasets. For each dataset, we train a binary classifier (e.g., anger vs non-anger) on its own set of positive and negative instances while for testing, the positives come from that dataset, but the negatives come from other datasets. We train the classifiers on 500 positive and 2000 images randomly selected from each dataset. Then for testing, we use 200 positive and 4000 negative images from other datasets.

Table 2 summarizes the results. For both datasets, we observe a significant decrease in performance (maximum of about 25% for Deep Emotion dataset on *sadness* emotion), suggesting that some of the new negative samples coming from other datasets are confused with positive examples. This indicates that

Table 2. Quantifying Negative Bias. Self refers to testing on the original test set while Others refer to the testing on a set where positives come from the original dataset but negatives come from the other. % Drop refers to the performance drop across the self and others. Values in Others represent the average numbers. WEBEmo refers to our released dataset that we will discuss in next section.

Task	+ve set: / -ve set:	Deep Emotion	Emotion-6	WEBEmo
anger vs non-anger	Self/Others/% Drop	90.64/78.98/**12.86**	92.40/83.56/**9.57**	83.90/83.37/**0.63**
fear vs non-fear	Self/Others/% Drop	85.95/80.77/**6.05**	81.14/76.02/**2.56**	82.97/84.79/**-2.19**
sadness vs non-sadness	Self/Others/% Drop	81.90/61.35/**25.09**	89.20/82.07/**7.99**	89.89/90.55/**-0.73**

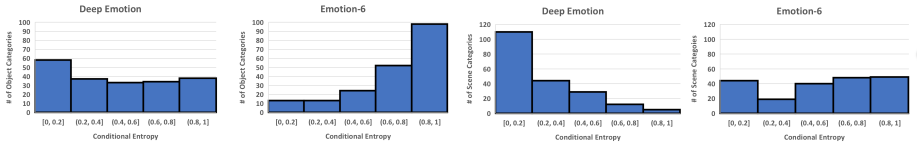
rest of the dataset does not well represent the rest of the visual world leading to overconfident and not very discriminative classifiers.

Test 4. Correlation Analysis with Object/Scene Categories. Given existing object/scene recognition models, the objective of this test is to see how well emotions are correlated with object/scene categories and whether analyzing the correlations can help to identify the presence of bias in emotion datasets. We use ResNet-50 pre-trained on ImageNet and ResNet-152 pre-trained on Places365 as object and scene recognition models respectively. We start our analysis by predicting object/scene categories from images of three common emotion categories used in previous task. We then select top 200 most occuring object/scene categories from each emotion class and compute the conditional entropy of each object/scene category across positive and negative set of a specific emotion. Mathematically, given an object/scene category c and emotion category e, we compute the conditional entropy as $\mathcal{H}(Y|X=c) = -\sum_{y\epsilon\{e_p,e_n\}} p(y|X=c)\log p(y|X=c)$, where e_p and e_n represent the positive and negative set of emotion e respectively (e.g., anger and non-anger). More number of object/scene categories with zero conditional entropy will most likey lead to a biased dataset as it shows the presence of these object/scene categories in either positive or negative set of an emotion resulting in an unbalanced representation of the visual world (Figure 1).

Figure 3 shows the distribution of object/scene categories w.r.t conditional entropy for both Deep Emotion and Emotion-6 datasets. While analyzing correlations between objects and *sadness* emotion in Figure 3.a, we observe that about 30% of object categories (zero conditional entropy) are only present in either sadness or non-sadness category and then further examining these categories, we find most of them will lead to a dataset bias (see supp). For example, objects like balloon, candy store and parachute are only present in negative set of *sadness*. Categories like balloon are strongly related to happiness, but still there should be a few negative balloon images such as sad balloon in the negative set[4]. Completely missing the negative balloon images will lead to dataset bias. Emotion-6 appears to be less biased compared to Deep Emotion but still it has 25% of object categories in the entropy range of [0,0.5]. Similarly, on analyzing scene categories for *anger* emotion in Fig. 3.b, we see that both datasets are biased towards to specific scene categories, e.g., for Deep Emotion, about 55% of scene categories have zero conditional entropy while about 20% of categories have zero entropy in Emotion-6. More Results are included in the supplementary.

Epilogue. Our main conclusions from these series of tests are the following:

[4] For example, see: `https://tinyurl.com/yazvkjmv`



(a) Object Categories for *Sadness* Emotion (b) Scene Categories for *Anger* emotion.

Fig. 3. Distribution of object/scene categories w.r.t conditional entropy. (a) objects in *sadness* emotion, (b) scenes in *anger* emotion. Both datasets show a strong presence of bias.

(a) Despite all three datasets are collected from Internet and labeled using a similar paradigm involving multiple humans, these datasets appear to have strong bias that severly obstruct learning a generalizable recognition model.

(b) In order to achieve better performance, the classifier (CNNs) should be trained on a very large-scale less-biased emotion datasets. However, emotional labeling of such large scale images can be very expensive, time-consuming and may often require specialists to avoid design bias. This begs an important question in one's mind whether this fully supervised paradigm of creating datasets is the right way forward to learning better emotion recognition models.

4 Curriculum Guided Webly Supervised Learning

Goal. The main goal of this section is to present possible remedies to the dataset bias issues described above, mainly proposing a large-scale web emotion database, called **WEBEmo** and an effective curriculum guided strategy for learning discrimative emotion features. Our basic idea is that we can potentially avoid (at least minimize) the effect of dataset design biases by exploiting vast amount of freely available web data covering a wide variety of emotion concepts.

Emotion Categories. Emotions can be grouped into different categories. Most prior works only consider a few independent emotion categories, e.g., Ekmas's six emotions [9] or Plutchik's eight emotion categories [10]. Instead, we opt for Parrott's hierarchical model of emotions [11] for two main advantages. First, by leveraging this hierarchy with associated lists of keywords, we are able to allieviate the search engine bias by diversifying the image search. Second, we are able to learn discriminative features by progressively solving different tasks.

Following [11], we design a three-level emotion hierarchy, starting from two basic categories (*positive* and *negative*) at level-1, six categories (*anger*, *fear*, *joy*, *love*, *sadness*, and *surprise*) at level-2 to 25 fine-grained emotion categories at level-3 (see Figure 5 for all categories). Note that while data-driven learning [59, 46] can be used for constructing such hierarchy, we chose to design it following prior psychological studies [11] as emotion has been well studied in psychology.

Retrieving Images from the Web. We use a stock website to retrieve web images and use those images without any additional manual labeling. Below, we provide a brief description of the dataset and refer to supplementary for details.

To collect web images for emotion recognition, we follow [11] to assemble a list of keywords (shown in supp) for each 25 fine-grained emotions, focusing on diverse visual concepts (see Figure 4). We then use the entire list of keywords to query a stock site and retrieve all the images (∼10,000) together with their tags

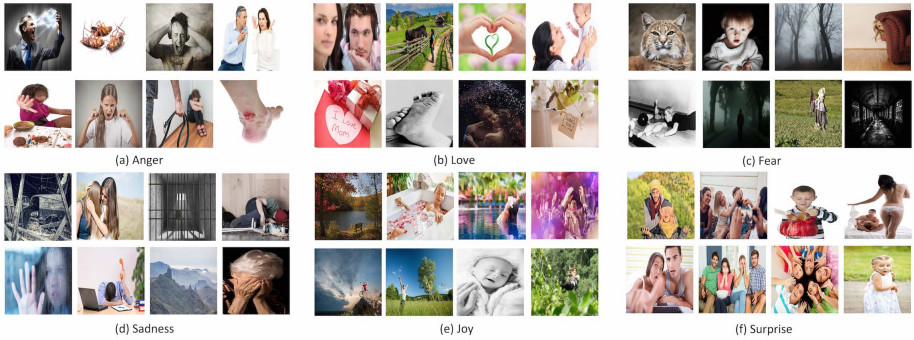
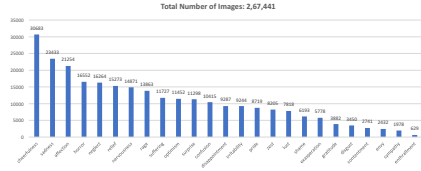


Fig. 4. Sample images from our **WEBEmo** dataset across six secondary emotion categories. These images cover a wide range of visual concepts. Best viewed in color.

Fig. 5. Category-wise distribution of images in **WEBEmo** dataset. The distribution is heavily long-tailed: e.g., there are more than 30K images on *cheerfulness* category but only 629 images on *enthrallment* emotion category.



returned for each query. In this way, we are able to collect about 300,000 weakly labeled images, i.e., labeled by the queries. We then remove images with non-english tags and also use captions with top-5 tags to remove duplicate images. After deduplication, we ended up with about 268,000 high-quality stock images.

Figure 5 shows category-wise distribution of images in **WEBEmo** dataset. The total number of images in our **WEBEmo** dataset is about 12 times larger than the current largest Deep Emotion dataset [1]. Though labels can be noisy, web searches unveil an order of magnitude more data which can be used to learn fine-grained emotion features.

Curriculum Guided Training. Our goal is to learn discriminative features for emotion recognition directly using our **WEBEmo** database. While it seems that one can directly train a CNN with such data, as in [35] for image classificaton, we found it is extremely hard to learn good features for our task, as emotions are intrinsically fine-grained, ambigious, and web data is more prone to label noise. However, as shown in psychology [11], emotions are organized in a hierarchy starting from basic emotions like postive or negative to more fine-grained emotions like affection, contentment, optimism and exasperation, etc. Categorizing images to two basic emotions is an easier task compared to categorizing images to such fine-grained emotions. So, what we want is an approach that can learn visual representation in a sequential manner like we humans normally learn difficult tasks in an organized manner.

Inspired by curriculum learning [42] and the emotion wheel from psychology [11], we develop a curriculum guided strategy for learning discriminative features in a sequential manner. Our basic idea is to gradually inject the information to the learner (CNN) so that in the early stages of training the coarse-scale properties of the data are captured while the finer-scale characteristics are learned in later stages. Moreover, since the amount of label noise is likey to be

much less in coarse categories, it can produce regularization effect and enhance the generalization of the learned representations.

Let C be the set of fine-grained emotion categories ($= 25$ in our case) and $k \in \{1 \dots K\}$ be the different stages of training. Assume $C_K = C$ is the fine-grained emotion categories that we want to predict; that is, our target is to arrive at the prediction of these emotion labels at the final stage of learning K. In our curriculum guided learning, we require a stage-to-stage emotional mapping operator $\mathcal{F}$ which projects C_k, the output labels at stage k, to a lower-dimensional C_{k-1} which is easier to predict compared to the prediction of C_k labels. We follow the Parrott's emotion grouping [11] as the mapping operator that groups C_K categories into six secondary and two primary level emotions as described earlier. Specifically, a CNN (pre-trained on ImageNet) is first fine-tuned with 2 basic emotions (positive/negative) at level-1 and then it serves to initialize a second one that discriminates six emotion categories at level-2 and the process is finally repeated for 25 fine-grained emotion categories at level-3.

5 Experiments

Goal. We perform rigorous experiments with the following two main objectives:

(a) How well our newly introduced **WEBEmo** dataset along with the curriculum guided learning help in reducing the dataset bias? (Sec. 5.1)

(b) How effective our visual representation learned using **WEBEmo** dataset in recognizing both image and video emotions? Does emotion features benefit other visual analysis tasks, say video summarization? (Sec. 5.2)

Implementation Details. All the networks are trained using the Caffe toolbox [60]. We choose ResNet-50 [7] as our default deep network and leave the study on impact of model capacity as future work. We follow [61] to initialize the networks from an ImageNet checkpoint while learning using web data. During training, all input images are resized to 256×256 pixels and then randomly cropped to 224×224. We use batch normalization after all the convolutional layers and train the networks using stochastic gradient descent with a minibatch size of 24, learning rate of 0.01, momentum of 0.9 and weight decay of 0.0001. We monitor the learning progress and continue the training until there is a plateau in validation accuracy for each task. We reduce the learning rate to its $\frac{1}{10}$ while making transition from one task to another in our curriculum guided training.

5.1 Revisiting Dataset Bias with Our Approach

Experiment 1: Quantifying Negative Bias. We use the same number of images (total 2500 for training and 4200 for testing) and follow the exact same testing protocol mentioned in Sec. 3: Test 2 to analyze negative bias on our **WEBEmo** dataset. Table 2 shows that classifiers trained on our dataset do not seem to be affected by a new external negative set across all three emotion categories (see right most column in Table 2). This is because **WEBEmo** dataset benefits from a large variability of negative examples and hence more comprehensively represent the visual world of emotions.

Experiment 2: Correlation Analysis with Object/Scene Categories. Figure 6 shows the correlation between emotion and object/scene categories in

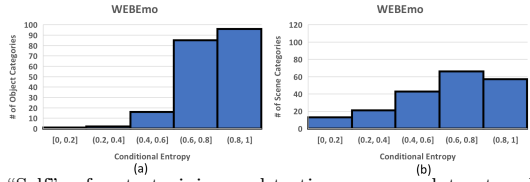


Fig. 6. Distribution of object/scene categories w.r.t conditional entropy on **WEBEmo** dataset. (a) objects in *sadness*, (b) scenes in *anger* emotion.

Table 3. Cross-Dataset Generalization. "Self" refers to training and testing on same dataset and "Mean Others" refers to the mean performance on all others. Model trained using curriculum guided webly supervised learning generalizes well to other datasets.

Train on: \ Test on:	Deep Sentiment	Deep Emotion	Emotion-6	WEBEmo	Self	Mean Others
Deep Sentiment [5]	78.74	68.38	49.76	47.79	78.74	55.31
Deep Emotion [1]	61.41	84.81	69.22	59.95	84.81	63.52
Emotion-6 (Sec. 3)	54.33	64.28	77.72	64.30	77.72	62.30
WEBEmo (Ours)	68.50	71.42	78.38	81.41	**81.41**	**72.76**

our **WEBEmo** dataset. As can be seen from Figure 6.a, less than 10% of object categories are within the entropy range [0,0.6] for *sadness* emotion leading to a much less biased dataset. This result is also consistent with the performance of the classifier trained for sadness vs non-sadness image classification in previous experiment (see Table 2). We also observe that more number of scene categories have entropy in the higher range (see Figure 6.b) showing that most of the scenes are well distributed across positive and negative emotion sets in our dataset. Note that the negative bias still persists regardless of the large size of our dataset covering a wide variety of concepts (some object/scene categories still have zero entropy). We can further minimize the bias by adding weakly labeled images associated with zero entropy categories such that both positive and negative set can have a balanced distribution. This experiment demonstrates that our correlation analysis can help to detect as well as reduce biases in datasets.

Experiment 3: Binary Cross-Dataset Generalization. Table 3 summarizes the results. We have the following key observations from Table 3: (1) Model trained using our **WEBEmo** dataset shows the best generalization ability compared to the models trained using manually labeled emotion datasets. We believe this is because learning by utilizing web data helps in minimizing the dataset biases by covering a wide variety of emotion concepts. (2) More interestingly, on Emotion-6 dataset, the model trained using our stock images even outperforms the model trained with images from the same Emotion-6 dataset (77.72% vs 78.38%). This is quite remarkable as our model has only been trained using the web images without any strong supervision.

Exploration Study. To better understand effectiveness of curriculum guided learning strategy, we analyze cross-dataset generalization performance by comparing with following methods: (1) Direct Learning – directly learning using the noisy web images of 25 fine-grained emotion categories, as in [1, 35, 30]; (2) Self-Directed Learning – start learning with a small clean set (500 images) and then progressively adapt the model by refining the noisy web data, as in [5, 33]; (3) Joint Learning – simultaneously learning with all the tasks in a multi-task setting. For details please refer to our supplementary material. We have the following key observations from Table 4: (1) Performance of direct learning baseline is much worse compared to our curriculum guided learning. This is not surprising

Table 4. Exploration study on different webly supervised learning strategies.

Methods	Deep Sentiment	Deep Emotion	Emotion-6	WEBEmo	Self	Mean Others
Direct Learning	62.20	67.48	74.73	76.65	76.65	68.13
Self-Directed Learning	64.56	68.76	76.15	78.69	78.69	69.82
Joint Learning	66.71	69.08	75.36	78.27	78.27	70.38
Curriculum Learning	68.50	71.42	78.38	81.41	**81.41**	**72.76**

since emotions are highly complex and ambigious that directly learning models to categorize such finegrained details fails to learn discriminative features. (2) Self-directed learning shows better generalization compared to the direct learning but still suffers from the requirement of initial labeled data. (3) The joint learning baseline is more competitive since it learns a shared representation from multiple tasks. However, the curriculum guided learning still outperforms it in terms generalization across other datasets (70.38% vs 72.76%). We believe this is because by ordering training from easy to difficult in a sequential manner, it is able to learn more discriminative feature for recognizing complex emotions.

Impact of Emotion Categories. We compare our three stage curriculum learning strategy (2-6-25) with a two stage one involving only six emotion categories (2-6). We found that the later produces inferior results, with an accuracy of 78.21% on the self test set and a mean accuracy of 70.05% on other two datasets, compared to 81.41% and 72.76% respectively by the three stage curriculum learning. Similarly, there is a drop of 2.31% in "self" test accuracy of the direct learning baseline while training with six emotion categories compared to the training with 25 emotion categories. In summary, we observe that the generalization ability of learned models increase with increased number of fine-grained emotion categories.

Impact of Dataset Size. We randomly sample a subset of 25,000 images (size similar to Deep Emotion dataset) from our **WEBEmo** dataset and follow the curriculum guided learning to train a model. We observe that the model trained using this subset produces an accuracy of 69.04% on the self test set and a mean accuracy of 64.49% on the other datasets, compared to 81.41% and 72.76% respectively by the full dataset which is about 10 times larger than this subset. Both self test and mean others accuracy increases as the size of dataset increases. Interestingly, model trained using the manually labeled Deep Emotion dataset only achieves a mean accuracy of 63.53% compared to 64.49% by the reduced subset while testing on the other datasets. This once again shows the effectiveness of our approach in learning a generalization recognition model.

State-of-the-Art Results. Note that all the numbers presented in Table 3 represent the binary accuracies we achieved without using any ground truth training data from the testing dataset. By fine-tuning, our model achieves a state-of-the-art accuracy of 61.13% in classifying eight emotions on Deep Emotion dataset [1] and an accuracy of 54.90% on Emotion-6 dataset. Similarly, by utilizing training data from Deep Sentiment dataset, our model achieves an accuracy of 82.67% which is about 8% improvement over the prior work [5].

5.2 Analyzing Effectiveness of Our Learned Emotion Features

Experiment 1: Testing on Cross-Domain Unbiased Data. In this experiment, we introduce a new unbiased emotion test set, **UnBiasedEmo** of about

Fig. 7. Sample images from our challenging **UnBiasedEmo** test set. See supplementary file for more example images on different object/scenes. Best viewed in color.

Table 5. Experimental results on our **UnBiasedEmo** test dataset. Features learned using curriculum learning outperforms all other basline features, including ImageNet.

Methods	Accuracy (%)
ImageNet	64.20
Direct Learning	71.64
Self-Directed Learning	72.45
Joint Learning	71.64
Curriculum Learning	**74.27**

3000 images dowloaded from Google to evaluate our learned models in recognizing very challenging emotions, e.g., different emotions with same object/scene (see Figure 7). Since source of this test set is different from our **WEBEmo** dataset, it helps us alleviate the dataset bias issue in evaluation, so we can compare the generalization ability of various learning strategies in a less biased manner. Note that developing a large-scale unbiased dataset containing hundred thousands of images like this is a very difficult task as it requires extensive effort and also provides poor scalability. For an example, we could only able to get 3045 emotional images across six emotion categories (same as Emotion-6 dataset) from a collection of about 60,000 images. More details on this unbiased dataset collection and annotations are included in the supplementary.

We freeze the weights and use our learned models as feature extractors. We use 80% of the images for training and keep rest 20% for testing. Table 5 shows the classification accuracies achieved by the features learned using different methods. We have the following observations from Table 5: (1) Our curriculum learning strategy significantly outperforms all other baselines in recognizing fine-grained emotions from natural images (see Figure 8 for some qualitative predictions by our model). (2) Among the alternatives, self-directed learning baseline is the most competitive. However, our approach still outperforms it due to the fact that we use the emotion hierarchy to learn discriminative features by focusing tasks in a sequential manner. (3) Performance of ImageNet features is much worse compared to the features learned using our curriculum guided webly supervised learning (64.20% vs 74.27%). This is expected as ImageNet features are tailored towards object/scene classification while emotions are more fine-grained and can be orthogonal to object/scene category, as shown in Figure 7.

We also inverstigate the quality of features learned using the current largest Deep Emotion dataset [1] in recognizing image emotions on this unbiased test set and found that it produces inferior results, with an accuracy of 68.88% compared to 74.27% by our curriculum guided webly-supervised learning strategy on the **WEBEmo** dataset. We believe this is because of the effective utilization of large scale web data covering a wide variety of emotion concepts.

Experiment 2: Sentiment Analysis. We perform this experiment to verify the effectivenss of our features in recognizing sentiments from online advertisement images. We conduct experiments using Image Advertisement dataset [63]

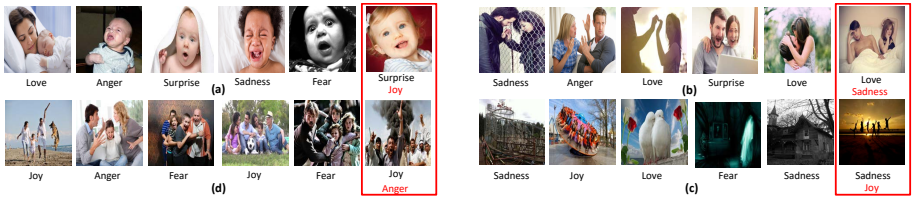


Fig. 8. Sample prediction results on **UnBiasedEmo** test dataset. From clockwise, depicted are examples of (a) "baby", (b) "couple", (c) "scenery" and (d) "group" with different emotions. The first examples in each row are correctly predicted by our curriculum guided learning model, while last example is an error, with our prediction in black and ground truth category printed below in red.

Methods	Accuracy (%)
ImageNet	23.42
Direct Learning	25.43
Self-Directed Learning	24.92
Joint Learning	26.18
Curriculum Learning	**27.96**

Table 6. Experimental results on Image Advertisement dataset. Our curriculum learning model performs the best.

Methods	Accuracy (%)
ImageNet	43.27
Direct Learning	45.67
Self-Directed Learning	46.18
Joint Learning	47.25
Knowledge Transfer [62]	45.10
Curriculum Learning	**49.22**

Table 7. Experimental results on VideoStory-P14 dataset. Features learned using our proposed curriculum learning outperforms the knowledge transfer approach by a margin of about 4%.

consisting of 30,340 online ad images labeled with 30 sentiment categories (e.g., active, alarmed, feminine, etc, – see [63] for more details). We use the model weights as initialization and fine-tune the weights [63]. We use 2403 images for testing and rest for training as in [63]. We follow [63] and chose the most frequent sentiment as the ground-truth label for each advertisement image.

Table 6 shows results of different methods on predicting image sentiments on the Advertisement dataset. From Table 6, the following observations can be made: (1) Once again, our curriculum guided learning significantly outperforms all other baselines in predicting sentiments from online ad images. (2) We achieve an improvement of about 6% over the ImageNet baseline showing the advantage of our learned features in automatic ad understanding tasks.

Experiment 3: Video Emotion Recognition. The goal of this experiment is to evaluate quality of our features in recognizing emotions from user videos. We conduct experiments on VideoStory-P14 emotion dataset [62] consisting of 626 user videos across Plutchik's 14 emotion classes. We fine-tune the weights using video datasets and use 80%/20% of the videos in each category for training/testing. To produce predictions for an entire video, we average the frame-level predictions of 20 frames which are randomly selected from the video.

From Table 7, the following observations can be made: (1) We can see that all the models trained using **WEBEmo** dataset outperforms both ImageNet and transfer encoding features [62] indicating the generalizability of our learned features in recognizing video emotions. (2) We further observe that curriculum guided learning provides about 2% improvement over the joint learning baseline. We also expect that further improvement can be obtained by incorporating audio features along with our emotion features while recognizing video emotions.

Experiment 4: Video Summarization. Our goal in this experiment is to see whether our learned features can benefit summarization algorithms in extracting

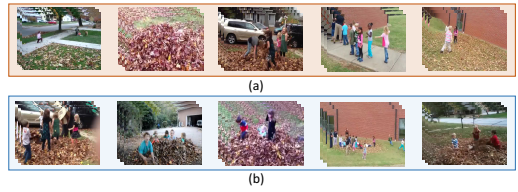


Fig. 9. Role of emotion in summarizing videos, "Kids Playing in Leaves" from CoSum dataset. Top row: [8] w/o emotion features, Bottom row: [8] w/ emotion features. Emotion-aware feature helps in extracting maximally informative shots by focusing central emotion ("joy") conveyed in videos.

high quality summaries from user videos. We believe this is possible since an accurate summary should keep emotional content conveyed by the original video.

We perform experiments on the CoSum dataset [64] containing 51 videos covering 10 topics from the SumMe benchmark [65]. We follow [8, 64] and segment the videos into multiple non-uniform shots for processing. We first extract pool5 features from the network trained with curriculum learning on our **WEBEmo** dataset and then use temporal mean pooling to compute a single shot-level feature vector, following [8]. We follow the exact same parameter settings of [8] and compare the summarization results by only replacing the visual features.

By using our learned emotion features, the top-5 mAP score of the recent summarization method [8] improves by a margin of about 3% over the C3D features [66] (68.7% vs 71.2%). This improvement is attributed to the fact that good summary should be succinct but also provide good coverage of the original video's emotion content (see Figure 9 for an example). This is an important finding in our work and we believe this can largely benefit researchers working in video summarization to consider the importance of emotion while generating video summaries, which has been largely ignored in the literature.

Additional Experiments in Supplementary. We analyze the effectivenss of our learned features in predicting communicative intents from persuasive images (e.g., politician photos) [67] and see that our approach outperforms all other baselines by a signifcant margin (∼8% improvement over ImageNet features). We also provide category-wise sample prediction results among the top-5K predictions by our curriculum guided learning model in the supplementary material.

6 Conclusion

In this paper, we have provided a thorough analysis of the existing emotion benchmarks and studied the problem of learning recognition models directly using web data without any human annotations. We introduced a new large-scale image emotion dataset containing about 268,000 high-quality images crawled from a stock website to train generalizable recognition models. We then proposed a simple actionable curriculum guided training strategy for learning discriminative emotion features that holds a lot of promise on a wide variety of visual emotion understanding tasks. Finally, we demonstrated that our learned emotion features can improve state-of-the-art methods for video summarization. Our approach has benefits in both performance and scalability. We hope that our learned model and datasets will serve as important resources to facilitate further research in the area of visual emotion analysis and beyond.

References

1. You, Q., Luo, J., Jin, H., Yang, J.: Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In: AAAI. (2016)
2. Kim, H.R., Kim, S.J., Lee, I.K.: Building emotional machines: Recognizing image emotions through deep neural networks. arXiv preprint arXiv:1705.07543 (2017)
3. Ng, H.W., Nguyen, V.D., Vonikakis, V., Winkler, S.: Deep learning for emotion recognition on small datasets using transfer learning. In: ICMI. (2015)
4. Peng, K.C., Chen, T., Sadovnik, A., Gallagher, A.C.: A mixed bag of emotions: Model, predict, and transfer emotion distributions. In: CVPR. (2015)
5. You, Q., Luo, J., Jin, H., Yang, J.: Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: AAAI. (2015)
6. Chen, T., Borth, D., Darrell, T., Chang, S.F.: Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. arXiv preprint arXiv:1410.8586 (2014)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016)
8. Panda, R., Roy-Chowdhury, A.K.: Collaborative summarization of topic-related videos. CVPR (2017)
9. Ekman, P.: An argument for basic emotions. Cognition & emotion (1992)
10. Plutchik, R., Kellerman, H.: Emotion, Theory, Research, and Experience: Theory, Research and Experience. Academic press (1980)
11. Parrott, W.G.: Emotions in social psychology: Essential readings. Psychology Press (2001)
12. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Emotion recognition in context. In: CVPR. (2017)
13. Du, S., Tao, Y., Martinez, A.M.: Compound facial expressions of emotion. Proceedings of the National Academy of Sciences (2014)
14. Fabian Benitez-Quiroz, C., Srinivasan, R., Martinez, A.M.: Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: CVPR. (2016)
15. Eleftheriadis, S., Rudovic, O., Pantic, M.: Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. TIP (2015)
16. Eleftheriadis, S., Rudovic, O., Pantic, M.: Joint facial action unit detection and feature fusion: A multi-conditional learning approach. TIP (2016)
17. Soleymani, M., Asghari-Esfeden, S., Fu, Y., Pantic, M.: Analysis of eeg signals and facial expressions for continuous emotion detection. TAC (2016)
18. Chu, W.S., De la Torre, F., Cohn, J.F.: Selective transfer machine for personalized facial expression analysis. TPAMI (2017)
19. Kahou, S.E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R.C., et al.: Combining modality specific deep neural networks for emotion recognition in video. In: ICMI. (2013)
20. Jou, B., Bhattacharya, S., Chang, S.F.: Predicting viewer perceived emotions in animated gifs. In: MM. (2014)
21. Xu, B., Fu, Y., Jiang, Y.G., Li, B., Sigal, L.: Video emotion recognition with ferred deep feature encodings. In: ICMR. (2016)
22. Wu, B., Jia, J., Yang, Y., Zhao, P., Tang, J., Tian, Q.: Inferring emotional tags from social images with user demographics. TMM (2017)
23. Wang, X., Jia, J., Tang, J., Wu, B., Cai, L., Xie, L.: Modeling emotion influence in image social networks. TAC (2015)

24. Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.S., Sun, X.: Exploring principles-of-art features for image emotion recognition. In: MM. (2014)
25. Alameda-Pineda, X., Ricci, E., Yan, Y., Sebe, N.: Recognizing emotions from abstract paintings using non-linear matrix completion. In: CVPR. (2016)
26. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: MM. (2010)
27. Li, W., Wang, L., Li, W., Agustsson, E., Van Gool, L.: Webvision database: Visual learning and understanding from web data. arXiv preprint arXiv:1708.02862 (2017)
28. Chen, X., Gupta, A.: Webly supervised learning of convolutional networks. In: ICCV. (2015)
29. Divvala, S.K., Farhadi, A., Guestrin, C.: Learning everything about anything: Webly-supervised visual concept learning. In: CVPR. (2014)
30. Joulin, A., van der Maaten, L., Jabri, A., Vasilache, N.: Learning visual features from large weakly supervised data. In: ECCV. (2016)
31. Gan, C., Sun, C., Duan, L., Gong, B.: Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In: ECCV. (2016)
32. Liang, J., Jiang, L., Meng, D., Hauptmann, A.G.: Learning to detect concepts from webly-labeled video data. In: IJCAI. (2016)
33. Gan, C., Yao, T., Yang, K., Yang, Y., Mei, T.: You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In: CVPR. (2016)
34. Sukhbaatar, S., Fergus, R.: Learning from noisy labels with deep neural networks. arXiv preprint arXiv:1406.2080 (2014)
35. Krause, J., Sapp, B., Howard, A., Zhou, H., Toshev, A., Duerig, T., Philbin, J., Fei-Fei, L.: The unreasonable effectiveness of noisy data for fine-grained recognition. In: ECCV. (2016)
36. Liang, J., Jiang, L., Meng, D., Hauptmann, A.: Exploiting multi-modal curriculum in noisy web data for large-scale concept learning. arXiv preprint arXiv:1607.04780 (2016)
37. Lee, Y.J., Grauman, K.: Learning the easy things first: Self-paced visual category discovery. In: CVPR. (2011)
38. Dong, Q., Gong, S., Zhu, X.: Multi-task curriculum transfer deep learning of clothing attributes. In: WACV. (2017)
39. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. arXiv preprint arXiv:1707.09465 (2017)
40. Gao, R., Grauman, K.: On-demand learning for deep image restoration. In: ICCV. (2017)
41. Pentina, A., Sharmanska, V., Lampert, C.H.: Curriculum learning of multiple tasks. In: CVPR. (2015)
42. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML. (2009)
43. Yan, Z., Zhang, H., Piramuthu, R., Jagadeesh, V., DeCoste, D., Di, W., Yu, Y.: Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition. In: ICCV. (2015)
44. Xiao, T., Zhang, J., Yang, K., Peng, Y., Zhang, Z.: Error-driven incremental learning in deep convolutional neural network for large-scale image classification. In: MM. (2014)
45. Griffin, G., Perona, P.: Learning and using taxonomies for fast visual categorization. In: CVPR. (2008) 1–8
46. Li, L.J., Wang, C., Lim, Y., Blei, D.M., Fei-Fei, L.: Building and using a semantivisual image hierarchy. In: CVPR. (2010)

47. Cesa-Bianchi, N., Gentile, C., Zaniboni, L.: Incremental algorithms for hierarchical classification. JMLR (2006)
48. Deng, J., Krause, J., Berg, A.C., Fei-Fei, L.: Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In: CVPR. (2012)
49. Deng, J., Satheesh, S., Berg, A.C., Li, F.: Fast and balanced: Efficient label tree learning for large scale object recognition. In: NIPS. (2011)
50. Marszalek, M., Schmid, C.: Semantic hierarchies for visual object recognition. In: CVPR. (2007)
51. Tousch, A.M., Herbin, S., Audibert, J.Y.: Semantic hierarchies for image annotation: A survey. PR (2012)
52. Jia, Y., Abbott, J.T., Austerweil, J.L., Griffiths, T., Darrell, T.: Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. In: NIPS. (2013)
53. Silla Jr, C.N., Freitas, A.A.: A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery (2011)
54. Srivastava, N., Salakhutdinov, R.R.: Discriminative transfer learning with tree-based priors. In: NIPS. (2013)
55. Wang, D., Shen, Z., Shao, J., Zhang, W., Xue, X., Zhang, Z.: Multiple granularity descriptors for fine-grained categorization. In: ICCV. (2015)
56. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR. (2011)
57. Khosla, A., Zhou, T., Malisiewicz, T., Efros, A.A., Torralba, A.: Undoing the damage of dataset bias. In: ECCV. (2012)
58. Tommasi, T., Patricia, N., Caputo, B., Tuytelaars, T.: A deeper look at dataset bias. In: Domain Adaptation in Computer Vision Applications. (2017)
59. Verma, N., Mahajan, D., Sellamanickam, S., Nair, V.: Learning hierarchical similarity metrics. In: CVPR. (2012)
60. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: MM. (2014)
61. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: ICCV. (2017)
62. Xu, B., Fu, Y., Jiang, Y.G., Li, B., Sigal, L.: Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. TAC (2016)
63. Hussain, Z., Zhang, M., Zhang, X., Ye, K., Thomas, C., Agha, Z., Ong, N., Kovashka, A.: Automatic understanding of image and video advertisements. In: CVPR. (2017)
64. Chu, W.S., Song, Y., Jaimes, A.: Video co-summarization: Video summarization by visual co-occurrence. In: CVPR. (2015)
65. Gygli, M., Grabner, H., Riemenschneider, H., Van Gool, L.: Creating summaries from user videos. In: ECCV. (2014)
66. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. (2015)
67. Joo, J., Li, W., Steen, F.F., Zhu, S.C.: Visual persuasion: Inferring communicative intents of images. In: CVPR. (2014)