Learning Visual Emotion Representations from Web Data

Zijun Wei\(^1\), Jianming Zhang\(^1\), Zhe Lin\(^1\), Joon-Young Lee\(^1\), Niranjan Balasubramanian\(^2\), Minh Hoai\(^2\), Dimitris Samaras\(^2\)
\(^1\)Adobe Inc., \(^2\)Stony Brook University

Abstract

We present a scalable approach for learning powerful visual features for emotion recognition. A critical bottleneck in emotion recognition is the lack of large scale datasets that can be used for learning visual emotion features. To this end, we curated a webly derived large scale dataset, StockEmotion, which has more than a million images. StockEmotion uses 690 emotion related tags as labels giving us a fine-grained and diverse set of emotion labels, circumventing the difficulty in manually obtaining emotion annotations. We used this dataset to train a feature extraction network, EmotionNet, which we further regularized using joint text and visual embedding and text distillation. Our experimental results establish that EmotionNet trained on the StockEmotion dataset outperforms SOTA models on four different visual emotion tasks. An added benefit of our joint embedding training approach is that EmotionNet achieves competitive zero-shot recognition performance against fully supervised baselines on a challenging visual emotion dataset, EMOTIC, which further highlights the generalizability of the learned emotion features.

1. Introduction

Understanding the emotion conveyed in an image or a video is an important computer vision task, one that has a wide range of applications from digital content management \([3, 6, 19, 45]\) and marketing \([17, 27, 50]\) to education \([10, 35]\) and healthcare \([7]\). In this paper, we address the need for a general visual emotion representation. We propose EmotionNet, a convolutional network that can take any input image and output a feature vector representing the emotion conveyed in the input image. The output feature vector can then be used for various downstream tasks such as emotion recognition, conditional image captioning and generation—much like how the feature vector from ResNet \([13]\) pretrained on ImageNet can be used for many downstream visual recognition tasks such as image classification, object detection, person tracking, and semantic segmentation. In other words, EmotionNet for visual emotions is analogous to a pretrained VGG16 for object categories.

EmotionNet is an emotion specific feature extraction network. One might question its merits over other general feature extraction networks such as a ResNet pretrained on ImageNet \([9]\). Unfortunately, such general feature extraction networks are not suitable for emotion analysis, as demonstrated in our experiments. This is understandable because ImageNet pretraining mainly forces the networks to distinguish between object categories, not visual emotions. Detecting emotion requires more than being able to recognize object classes—the same object can evoke different emotions depending on the context in which it appears.

To build a useful feature extraction model, it is crucial to have relevant training data; in our case, an emotion dataset at the scale of ImageNet with a million images and a well-defined taxonomy over hundreds of categories is desirable. Unfortunately, it is difficult to use the same approach as ImageNet to collect an annotated dataset for emotion. Due to language ambiguities and the abstract nature of emotion definitions, identifying emotion in an image is a much harder task than labeling object categories when there is no definitive emotion taxonomy over hundreds of categories. Most existing visual emotion datasets only provide annotations for a small set of emotion categories on a limited scale. Thus, features learned on such limited datasets generalize poorly to other emotion datasets \([16, 33]\).

In this paper, we propose to learn EmotionNet by leveraging web data. We use commercial stock images and their associated tags as our data source and annotation. Different from previous datasets that are manually labeled based on predefined emotion taxonomies with limited categories, we curate our stock image dataset based on 690 common tags that are related to more fine-grained and open categories of emotions. The resulting dataset, StockEmotion, is composed of over one million stock images, covering diverse emotion concepts related to humans, scenes, and symbols.

However, annotated stock image tags can be incomplete and noisy. The owner or creator of an image might only provide a few tags for each image, or might associate an image with concepts that are unrelated or only remotely related.
lated to the image. So, we need to address the technical challenge of how to learn from noisily and partially labeled images. Based on the fact that the representations of visual data (e.g., the input image) and text data (the associated tags) should be semantically close to each other, correlating information in the tags and the images can act as a regularizer for the image representation. To this end, we propose an approach for training a joint text and visual embedding that (1) reduces noise in the webly annotated tags and (2) induces a joint space that can be used for cross-modal tasks.

Empirically, we show that EmotionNet, a standard ConvNet architecture trained on the StockEmotion dataset, is indeed useful for various emotion recognition benchmarks. In addition, EmotionNet can be further enhanced by leveraging the image tags through knowledge distillation from text models. We investigate text models and embeddings learned in unsupervised and semi-supervised settings. The text models are used to denoise the keyword labels and enforce joint visual-text embeddings to regularize the visual feature learning.

The induced joint visual-text emotion embedding space can also be used for zero-shot emotion recognition. We achieve competitive performance against fully-supervised methods on the challenging EMOTIC dataset [21].

To sum up, our contributions are as follows:
1. We introduce a large-scale image dataset for visual emotion content.
2. We provide a general feature extraction network for emotion. This feature representation achieves state-of-the-art performance on several visual emotion benchmarks across different domains. The learned joint vision-text embedding achieves competitive zero-shot learning performance.
3. We propose methods to handle noisily, partially annotated data, improving visual feature learning through text model distillation and joint visual-text embedding.

2. Related Work

**Emotion in Psychological Research.** Studying emotions and their relations is an important research area in psychology. Two competing approaches are used in describing emotion: categorical [8, 11, 34, 36] that classifies emotions into basic categories and dimensional [38, 51] that projects emotions into a continuous manifold. Our work sidesteps this debate in that we construct a large collection of emotional words and learn an emotion representation in a data-driven approach. With a large number of emotional words, our model implicitly has much higher dimension than the traditional two to three dimension models used in psychology, allowing us to capture subtle differences in emotions.

Language plays a fundamental role in experiencing and perceiving emotions [24]. With this in mind, our approach connects visual emotion features to a latent emotion space learned from a textual embedding. Our work uses language models to learn an emotion embedding from the text keywords associated with images. See [49] for a detailed review on emotion detection in text.

**Visual Emotion Datasets.** Visual emotion detection is often framed as a classification problem defined over a small number of predefined emotion classes [21, 22, 25, 32, 33]. However, such a limited categorical taxonomy fails to capture the rich variation and mixture of emotions expressed in images and limits the diversity of retrieved images. There are some datasets with a larger number of categories that combine emotion words with nouns and their descriptive context [2, 5]. Our work goes further and introduces a richer descriptive set for modeling emotions using the natural distributions of keywords assigned to images.

**Learning from noisy data.** In this paper, we develop a method for training a feature extraction network from noisily annotated web data. Handling noisily labeled data is a well-studied area with many solutions (e.g., [14, 42, 43]), and we refer the reader to [12] for a comprehensive overview of label noise and robust algorithms.

In our work, we constrain images to be close to their keywords in the joint space induced by the transformation from visual space to textual space. Such multi-view structure preservation constraints have been explored in the metric learning literature [15, 28, 39]. However, different from previous work which requires a small set of clean data [48], our work does not need any clean labels as it is difficult to collect clean labels for stock images. We therefore develop a training method where a regularization term on the noisy labels is added to mitigate the label noise itself.

3. The StockEmotion Dataset

We have collected a large-scale dataset of images from Adobe Stock with emotion keywords extracted from the original image keywords provided by the image uploaders. Some samples are shown in Fig. 1.

3.1. Data Collection

We used Adobe Stock to search images using an overcomplete set of emotion keywords to cover diverse emotion concepts. Initially, we constructed a list of emotion keywords using linguistic emotion lexicons such as NRC-emotion [31] and WordNet-Affect [40]. However, we found that these emotion lexicons are not suitable for computer vision tasks. For example, many adjectives such as beautiful and white are labeled with emotions in these lexicons, but these keywords are often associated with images that do not convey the corresponding emotion information. To get
Our approach for collecting the stock images is motivated by the fact that emotion tags map poorly to existing emotion taxonomies. Many category names in the emotion taxonomies are rarely used for tagging stock images, leading to poor image retrieval results for data collection. Moreover, there are also many emotional tags that are not included in the taxonomies’ vocabulary, e.g. abuse, danger and challenge, which can provide useful semantic context for identifying fine-grained emotions.

3.2. Statistics

StockEmotion consists of 1.17 million images which we split, at random, into training (1.06 M), validation (33K), and testing (71K) subsets. Each image on average has 48.9 keywords, among which 7.04 are emotional keywords included in our 690-keyword list.

Since StockEmotion is curated through web search, it includes noisy labels. To estimate the amount of noise in the labels, we randomly sampled a subset of 1000 images and asked our lab colleagues to manually check the correctness of the weak emotion labels. The error rate of the emotion labels turned out to be around 15%, making it suitable for training deep convolutional networks [37].

There are around 600K images with one or more people detected by an open-source face/body detector [4]. About 280K out of the 600K images have one single clear face in the image. A significant portion of the images do not contain humans, and can be scenes, objects and symbols related to emotions, as shown in Fig. 1.

Looking at the co-occurrence matrix for the keywords, we found that only a small portion of the keywords co-occur frequently. Most of the 690 categories are independent of each other. A visualization for the co-occurrence matrix is...
3.3. SE30K8 – A Manually Annotated Subset

For verification and controlled studies, we collected ‘cleaner’ annotations, albeit of a different type, for a subset of the images. Starting from Ekman’s emotion taxonomy [11]: anger, happiness, surprise, disgust, sadness, fear, we added a neutral category and divided the surprise category into surprise-positive and surprise-negative. This led to a set of eight emotion categories.

We collected human annotation for the eight emotion categories for a subset of 33K images, using Amazon Mechanical Turk (AMT). For each image, annotators were asked to select all the emotional categories expressed in the image. Each image was annotated by five AMT workers (after a qualification task). Annotations provided by the workers are reasonably consistent: more than 85% of the images had the same annotation by at least three annotators. Many of these images have clearly conveyed emotions, but it is difficult to describe them using basic categories [11, 36].

4. EmotionNet

EmotionNet is a general feature extraction network for emotion, trained on the StockEmotion dataset, which has emotion keywords for over a million images. As in most annotations derived from web, the list of emotion keywords for an image might be incomplete, erroneous, or both. The presence or absence of a keyword in the list does not necessarily mean that the image must or must not be associated with that keyword. This is referred to as label noise, and we estimate there is around 15% of label noise. Unfortunately, this will impact the performance of feature extraction networks trained on the StockEmotion dataset, especially those trained by minimizing the data negative log-likelihood.

To mitigate the noise problem, we propose to use an additional data type that also comes with the stock images: text! In addition to the list of emotion keywords, each image in our dataset also comes with other non-emotion keywords, which should also be utilized to our benefit. Non-emotion keywords, by definition, do not convey emotions, but there exist correlations between them. For example, an image with keywords like sunday, young, outdoor is likely to evoke positive emotions. We propose to use the list of non-emotion keywords to infer the missing emotion keywords; in particular, we train a text-based classifier that predicts emotion keywords from the list of non-emotion keywords. The predicted distribution of emotion keywords in combination with the tagged emotion keywords are now used as the smoothed labels for training the feature extractor. Furthermore, we also regularize the visual features by forcing them to be compatible with the text-derived representation of the emotion keywords associated with them. The overview of our proposed model is shown in Fig. 2.

Figure 2: Training of EmotionNet. The non-emotion keywords of an image are used to predict the emotion keywords associated with the image. The predicted emotions and the original (noisy) emotion keywords are combined to form the target class distribution. EmotionNet is trained by minimizing two losses: the multi-label classification loss and the joint embedding loss. The joint embedding loss requires the visual embedding of the image to be compatible to the textual embedding of the associated keywords.

Figure 3: Architecture of text-to-emotion networks. This architecture is used by a text-to-emotion network, mapping an unordered list of keywords to a probability vector for multiple emotion categories.

Many word embeddings already exist, and state-of-the-art models often exploit the sequential and compositional nature of text [41, 44]. In our case, however, the text associated with each stock image is an unordered collection of keywords with no sequential or compositional aspects. We therefore use a simple model that combines the feature embeddings of multiple words to produce a fixed length feature vector. Although simple, such models have been shown to be effective for multiple text classification tasks [1, 18, 46].

Fig. 3 shows the components of our text-to-emotion classifier. The classifier is a mix of a text CNN [20] and a
deep averaging network (DAN) [18]. The classifier uses word2vec embeddings [30] to represent the keywords as rows in an embedding matrix. The CNN component uses a 1D convolution with kernel size one and a ReLU activation to transform the word embedding features into feature maps. The DAN component averages the feature maps using an averaging pooling layer and then applies one fully-connected layer for non-linear transformations. The resulting feature vector is projected onto the $K = 690$ emotion keywords categories. We denote this text model as TextCNN. We train this TextCNN model on the training set of the StockEmotion dataset.

The predicted probabilities from the text-to-emotion classifier are then combined with the original binary indicators to yield an augmented label distribution as follows:

$$y'_k = \frac{\hat{y}_k + y_k}{1 + \sum_{i=1}^{K} \hat{y}_i},$$ (1)

where $\hat{y}_k$ is the predicted probability by the text-to-emotion classifier for the emotion keyword $k$ and an input image $x$, $y_k$ is the binary indicator for whether the keyword $k$ is among the original keywords of image, and $y'_k$ is the resulting soft label. The multiple-label classification loss is then expressed as:

$$L_{cls} = -\frac{1}{K} \sum_{k=1}^{K} y'_k \log(P_k(x)).$$ (2)

The second type of regularization that we introduce is based on the observation that the tags provide an alternate view for the emotion conveyed in the image. As such, we can use the text-based embedding to aid the training of the visual embedding. The main idea is to ensure that the visual emotion features are compatible with the text-based features. We use the average of the keyword embeddings as our text-based representation and map the visual features into the same feature space. We add a regularization term into the training loss to encourage a small cosine distance between the text and the transformed visual features. Formally, the embedding loss for a pair of image $x$ and a list of keywords $y$ is given by:

$$L_{embed} = 1 - \cos(f_t(y), W f_v(x)), $$ (3)

where $f_t(y)$ is the average of all keyword features, $f_v(x)$ is the visual embedding of the input image $x$, and $W$ is a linear transformation that maps the visual features to the joint embedding space.

Finally, for a pair of image $x$ and associated keywords $y$ we minimize the combined loss function between the classification loss and the embedding loss:

$$L = L_{cls} + \lambda L_{embed},$$ (4)

where $\lambda$ controls the strength of the embedding loss term. We set $\lambda = 1$ in all of our experiments for simplicity and did not tune it for better performance. There are many other advanced solutions for this multimodal representation learning problem (detailed survey in [47]). Here we choose a simple but effective approach, as shown in our experiments.

5. Experiments

This section describes experiments to evaluate the benefits of EmotionNet for several emotion analysis tasks. First, we use EmotionNet as a feature extractor and train simple linear classifiers on emotion datasets and measure the recognition performance on those datasets. Second, we evaluate EmotionNet on zero-shot learning. Finally, we compare qualitatively between the features from EmotionNet and another generic feature extraction network for the task of image retrieval.

5.1. Network and implementation details

We use ResNet50, a residual network with 50 layers [13], as our backbone network. We initialize the model with ImageNet pretrained weights and continue to train on StockEmotion for 30 epochs using stochastic gradient descent with a mini batch size of 256, learning rate 0.001, momentum 0.9, and weight decay $10^{-5}$. We reduce the learning rate by a factor of 10 at epochs 10 and 20. When training converges, top-1 prediction accuracy for the 690 emotional categories on the test set stabilizes around 50%. Our experiments suggest that EmotionNet models trained from scratch on StockEmotion achieve similar accuracy values, but their training takes longer to converge.

The TextCNN model was trained following [18] using AdaGrad with an initial learning rate of 1 and dropped by a factor of 10 every 10 epochs for 30 epochs. We used the publicly available word2vec [30] trained with GoogleNews to generate word embeddings. We also experimented with word2vec embeddings learned from our dataset by regarding the keyword list associated with each image as a sentence but no improvement was observed.

5.2. Evaluation of learned image features

We evaluate the learned features by using them for emotion category prediction tasks defined by other emotion datasets. We use ResNet50 trained on StockEmotion to extract image features. The extracted features are used as is, without any fine-tuning on the target task datasets. We use simple linear classifiers for emotion category prediction in order to demonstrate the utility of the visual features returned by EmotionNet.

**Evaluation Protocol.** We freeze all the layers of EmotionNet and replace the last fully-connected layer with a new one that projects the learned features to the output cate-
Table 2: Emotion detection performance on multiple emotion datasets: A simple linear classifier trained with the Visual features learned on StockEmotion surpasses SOTA results on four of the five datasets. Our proposed unsupervised text regularization method provides minor modest additional gains.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous SOTA</td>
<td>61.13</td>
<td>74.30</td>
<td>57.31</td>
<td>-</td>
<td>25.44</td>
<td>22.48</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>58.30</td>
<td>60.26</td>
<td>40.17</td>
<td>52.52</td>
<td>24.34</td>
<td>26.03</td>
</tr>
<tr>
<td>EmotionNet</td>
<td>65.81</td>
<td>81.45</td>
<td>53.43</td>
<td>69.78</td>
<td>29.24</td>
<td>30.96</td>
</tr>
</tbody>
</table>

Datasets. We evaluate on the following datasets:

DeepEmotion [52] uses eight emotions derived from a recent psychological study [29]. It has 23K images collected from Flicker and Instagram that were annotated by Amazon Mechanical Turk workers. We followed the experiment setup for the emotion recognition task in [33] in which the authors used 80% of the 23K images for training and the remaining 20% for testing.

UnBiasedEmotion [33] contains 3000 images downloaded from Google with different emotions for the same objects to reduce object bias. Each image is labeled with one of six emotional categories. We follow the evaluation setup in [33].

EMOTIC [21] consists of a mixture of images from MSCOCO [23], Ade20k [55], and images that were manually downloaded using Google search. The dataset is a collection of images of people in real environments and includes annotations of their apparent emotions drawn from a set of 26 emotion categories. It includes 18,316 images with a total of 23,788 annotations. We report performance of our models on both cases denoted as EMOTIC-B(ody) and EMOTIC-I(mage). We follow training and evaluation procedures used in [21].

AffectNet [32] contains around 400K annotated facial images, each labeled by a single coder. It includes 5K labeled images in 10 categories as the validation set. Following [53], we selected around 280K images as training samples and 3.5K images for validation. The labels include six basic emotions and a neutral category. For efficiency, in each training epoch, we sample 30K images uniformly at random covering the seven categories and trained the final fully connected layer for 10 epochs.

SE30K8 is the manually annotated subset of our StockEmotion dataset as described in Sec. 3.3. We use a randomly selected subset of 22K images as training samples and 3K for validation. We test on 5K images. We again follow the evaluation setup in [33].

The datasets listed above are diverse in terms of image sources, emotion categories, and exhibiting locations. The emotion conveyed in an image could be inferred from the expression on a face, or the pose of a human body, or from the overall scene.

Comparison methods and results. We directly compare to previous state-of-the-art algorithms on each dataset: [33] achieves state-of-the-art performance on DeepEmotion [52] and UnBiasedEmotion [33] using curriculum training algorithms. Kosti et al. [21] report state-of-the-art performance on EMOTIC [21] by combining both categorical and continuous emotion information. Zeng et al. [53] report best performance on [32] by training on multiple datasets and automatically filtering inconsistencies. Compared to these methods, ours model is relatively simpler, a linear classifier on top of the visual emotion features from EmotionNet. To establish the utility of visual emotion features over general purpose image features, we also compare with features from a generic feature extractor ResNet-50 pre-trained on ImageNet.

The results in Tab. 2 show that: (1) the classifiers trained using the features from EmotionNet outperform four of the five previous state-of-the-art algorithms; and (2) the features from ResNet-50, a network trained for object recognition (ImageNet), are not useful for emotion prediction.

<table>
<thead>
<tr>
<th>Ablation experiment</th>
<th>DE</th>
<th>UBE</th>
<th>EMTIC-B</th>
<th>EMTIC-I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extra anno.</td>
<td>65.53</td>
<td>81.45</td>
<td>53.43</td>
<td>29.24</td>
</tr>
<tr>
<td>Soft loss</td>
<td>64.76</td>
<td>80.13</td>
<td>52.66</td>
<td>28.61</td>
</tr>
<tr>
<td>Embed loss</td>
<td>65.85</td>
<td>80.29</td>
<td>52.71</td>
<td>28.74</td>
</tr>
<tr>
<td>Embed &amp; Soft</td>
<td>65.29</td>
<td>78.98</td>
<td>52.51</td>
<td>28.58</td>
</tr>
</tbody>
</table>

5.3. Ablations studies

We conduct ablation studies to further understand the values of the StockEmotion dataset and the components of EmotionNet.
Benefits of extra supervision. Can we improve the performance of EmotionNet with extra supervision? To answer this question, we perform an experiment where we also train EmotionNet on SE30K8, a subset of the StockEmotion dataset with human annotation for eight basic emotions. We first train a text-based classifier that predicts the eight emotion categories from a list of keywords. The representation produced by this text-classifier is an alternate view of the emotion conveyed by the image. We use it to guide the learning of the visual embedding network, forcing the transformed feature vector to be compatible with the 8-emotion embedding feature vector. Further details on this setup can be found in the supplementary material. Tab. 3 compares the performance of EmotionNet trained with and without extra supervision. As can be seen, adding extra supervision does not provide consistent benefits. The extra supervision provides minor gains in two out of five cases, while slightly degrades performance in the others. This can be attributed to the limited size of the extra annotation (only 30K) or to the small number of emotion categories (only 8). In either case, it is time-consuming and costly to either increase the number of manually annotated images or the number of manually specified annotations. On the other hand, EmotionNet trained on our webly derived StockEmotion dataset does not suffer from these scalability issues.

Benefits of different loss functions. In addition to the original loss associated with predicting the emotion keywords that come with the images, EmotionNet is also trained with an embedding loss, which aims to minimize the distance between the visual representation and the textual representation of the associated tags and a soft label loss. The soft label loss refers to the difference between the label distribution predicted by EmotionNet and the emotion probabilities predicted by the text-to-emotion classifier. Tab. 3 shows the ablation study where we evaluate the contribution of the soft-label classification loss and the embedding loss. As can be seen, removing either or both of these loss terms degrades performance.

Benefits of a large emotion taxonomy. StockEmotion has 690 emotion categories. We perform experiments to understand the benefits of having such a large number of categories. We consider two variants of the feature extraction network, trained with different supervision signals: (1) Use the full set of 30K keywords, rather than using just the 690 emotion related keywords, for training the feature extractor. (2) Use only eight basic emotion categories. We first learn a text classifier that predicts the eight emotion categories given image keywords as input. We train this classifier on the 30K images of SE30K8 and use it to predict emotion categories for rest of StockEmotion. We use these predicted labels as emotion pseudo-labels for the images (since the image keywords often contain clear indicators of emotion, this pseudo-labeling is of high accuracy, yielding up to 90% in top-2 accuracy). We then train the image feature extractor to predict these emotion pseudo-labels.

Tab. 4 compares the performance of the feature extraction networks trained with different sets of emotion keywords or labels. As can be seen, the feature extraction network trained with 30K labels is substantially worse than the network trained with emotion-related keywords alone, either with 690 or 8 emotions. This suggests the benefits of focusing on the emotion-related concepts. The feature extraction network trained with 8 emotion labels is not as good as the network trained with 690 emotions. This indicates the benefits of having a fine-grained list of emotion categories.

<table>
<thead>
<tr>
<th>% of StockEmotion</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>52.45</td>
<td>66.34</td>
<td>72.24</td>
<td>76.63</td>
<td>78.98</td>
</tr>
</tbody>
</table>

Table 5: Ablation Experiment. Performance of feature extraction networks trained with various subsets of StockEmotion on UnBiasedEmotion (UBE) dataset.

For all experiments thus far, we used the publicly available word2vec trained with GoogleNews to generate word embeddings. We also experimented with word2vec embeddings learned from our dataset and variants of the text classification model [18], but there were no significant improvements. We report these detailed experiments in the supplementary material.

5.4. Zero-Shot Learning Performance

EmotionNet is trained with both classification and joint vision-text embedding losses. One benefit of this approach is that the feature vectors returned by EmotionNet can be
used for zero-shot learning, given the ability to map the image features into the same space as the text features. We evaluate EmotionNet for zero-shot learning on the EMOTIC dataset [21], in which each of the 26 emotion categories comes with a brief textual description. We create a representation for each of these categories in the text emotion space by processing the emotion keywords mentioned in their descriptions through the aforementioned text-to-eight-emotion classifier. To classify any image, we first use the text distillation model to produce a representation of the image in the text emotion space. We then score each category based on its cosine similarity to the representation of the image in the text emotion space.

Results in Tab. 6 show that zero-shot learning using EmotionNet gets close to the fully supervised SOTA method on EMOTIC-B and outperforms the SOTA method on EMOTIC-I. Note that in these experiments, we do not perform any training on the EMOTIC dataset. The results show the strong generalizability of EmotionNet and the representations learned on the StockEmotion dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>EMOTIC-B</th>
<th>EMOTIC-I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous SOTA</td>
<td>25.44</td>
<td>22.48</td>
</tr>
<tr>
<td>EmotionNet</td>
<td>23.29</td>
<td>24.24</td>
</tr>
</tbody>
</table>

Table 6: Zero-Shot Learning results on EMOTIC

5.5. Image Retrieval and Qualitative Results

The feature representations produced by EmotionNet can be used to find images with similar emotion content. Given a query image, we can retrieve the nearest neighbors to the query in the emotion feature space. Fig. 4 compares the performance of ImageNet and EmotionNet features for image retrieval. The figure shows four query examples on the left. The images on the right are the nearest neighbors obtained by either ImageNet features (top row) or EmotionNet features (bottom row). As can be seen, ImageNet features return nearest neighbors that have relevant object categories but unrelated emotion attributes. For example, for the query image on the bottom left, none of the images returned by ImageNet features conveys the emotion horror. In contrast, using EmotionNet features, we can retrieve other horror images.

6. Conclusion

Advances in many computer vision tasks have been built on top of large scale datasets such as ImageNet. Such large datasets enable learning effective representations that are transferable to a variety of downstream tasks. In this work, we introduced a scalable method for acquiring a large-scale image dataset with rich emotion related tags. Using this method, we created EmotionStock, a dataset with more than a million images and 690 emotion-related keywords. We also proposed text-based distillation methods to mitigate the problem of label noise, creating EmotionNet, a general feature extraction network for emotion content. Experiments on a number of datasets showed that EmotionNet is useful for various downstream emotion analysis tasks, including emotion recognition, zero-shot learning, and image retrieval.

Acknowledgements. This project is partially supported by NSF IIS-1763981, the Partner University Fund, the SUNY2020 Infrastructure Transportation Security Center, and a gift from Adobe. This material is also based on research that is in part supported by the Air Force Research Laboratory (AFRL), DARPA, for the KAIROS program under agreement number FA8750-19-2-1003. The U.S. Government is authorized to reproduce and distribute reprints for Govermental purposes notwithstanding any copyright notation thereon.
References


[28] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Metric learning for large scale image clas-


