Smile, Be Happy :) Emoji Embedding for Visual Sentiment Analysis

Ziad Al-Halah ziadlhlh@gmail.com
Andrew Aitken aaitken@twitter.com
Wenzhe Shi wshi@twitter.com
Jose Caballero jcaballero@twitter.com
Twitter

Abstract

Due to the lack of large-scale datasets, the prevailing approach in visual sentiment analysis is to leverage models trained for object classification in large datasets like ImageNet. However, objects are sentiment neutral which hinders the expected gain of transfer learning for such tasks. In this work, we propose to overcome this problem by learning a novel sentiment-aligned image embedding that is better suited for subsequent visual sentiment analysis. Our embedding leverages the intricate relation between emojis and images in large-scale and readily available data from social media. Emojis are language-agnostic, consistent, and carry a clear sentiment signal which make them an excellent proxy to learn a sentiment aligned embedding. Hence, we construct a novel dataset of 4 million images collected from Twitter with their associated emojis. We train a deep neural model for image embedding using emoji prediction task as a proxy. Our evaluation demonstrates that the proposed embedding outperforms the popular object-based counterpart consistently across several sentiment analysis benchmarks. Furthermore, without bell and whistles, our compact, effective and simple embedding outperforms the more elaborate and customized state-of-the-art deep models on these public benchmarks. Additionally, we introduce a novel emoji representation based on their visual emotional response which supports a deeper understanding of the emoji modality and their usage on social media.

1. Introduction

Analyzing people's emotions, opinions, and attitudes towards a specific entity, an event or a product is referred to as sentiment analysis [29, 25]. Sentiment can be reduced to *positive*, *neutral*, and *negative*, or can be extended to a richer description of fine-grained emotions, such as *happiness*, *sadness*, or *fear*. Summarizing and understanding sentiment has important applications in various fields like interpretation of customer reviews, advertising, politics, and social studies. Thus, automated sentiment analysis is an active subject of research to devise methods and tools to enable such applications [20, 35, 2].

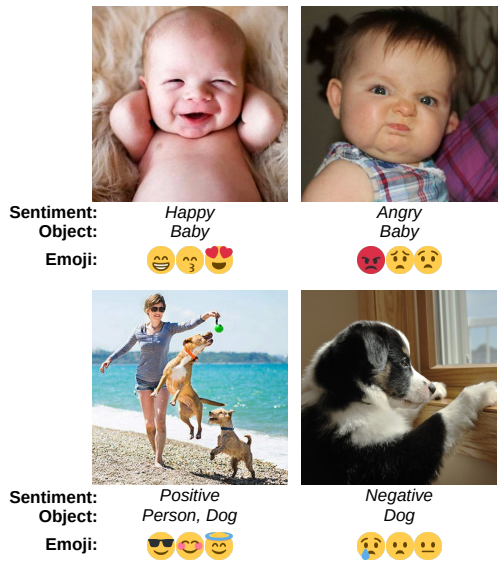


Figure 1: Images with similar objects may show different sentiments. Unlike the object neutral representation, emoji embedding is well aligned with the sentiment label space. Hence, it is expected to generalize well in transfer learning settings for visual sentiment and emotion analysis.

Driven by the availability of large-scale annotated datasets [15, 40] along with modern deep learning models, language sentiment analysis witnessed great improvements over the last few years [32]. However, *visual* sentiment analysis still lags behind. This is mainly due to the lack of large-scale image datasets with sentiment labels. Current datasets (*e.g.*, [43, 33, 2, 23, 28]) are scarce and too small to appropriately train deep neural networks, which are prone to overfitting the small training data.

To overcome the previous problem, the dominant approach currently is to employ cross-domain transfer learning methods. This is achieved by *pretraining* a deep neural network on a large-scale dataset for object classification, such as ImageNet [38], and then *fine-tuning* the network for sentiment classification on the small target dataset. This ap-

proach is unanimously adopted by recent visual sentiment models and has led to improved results, *e.g.* [43, 7, 33]. Nonetheless, object categories and sentiment labels are not aligned and rather orthogonal. Object labels are sentiment neutral; *i.e.* objects of the same category can exhibit various emotions (Fig. 1). Hence, the domain gap between object recognition and sentiment analysis is significant. Pretraining a model with an object-focused embedding may not be the most useful representation for subsequent transfer learning for sentiment or emotion classification.

Given that collecting data for the target task is impractical, is there an alternative representation which 1) is better aligned with sentiments and 2) can be learned efficiently with minimum overhead? Emojis, with the advent of social media, became a prevailing medium to emphasize emotions in our communications such as happiness 😀, anger 😡, or fear 😱. Not only do emojis carry a clear sentiment signal by themselves (see Fig. 1), they also act as sentiment magnifiers or modifiers of surrounding text [34]. Additionally, due to their prominent use in social media like Facebook, Twitter and Instagram, one can relatively easily tap into large amounts of readily available data without the need for any manual labeling. All these factors turn an emoji-based representation into an attractive candidate for our target task of visual sentiment analysis. In fact, emojis have been successfully leveraged for *language* sentiment analysis recently [17, 14, 36].

However, the interaction among emojis and the corresponding images in social media remains elusive. Is there a strong correlation between an emoji and a visual signal? And if so, do emojis capture the visual sentiment exhibited in images? The answer to these questions is not straightforward. Social media data is known to be noisy [3], and the use of emojis is influenced by the user's cultural background [4, 26] and major temporal events [39]. These hurdles represent important challenges to learning an effective emoji representation that can generalize well across domains. In this paper, we present the *first* work to address the previous questions with a thorough analysis of emojis and their visual sentiment connotation.

To that end, we leverage weakly labeled data collected from social media (*e.g.* Twitter) to build a large-scale dataset of 4 million images and their corresponding emoji annotation. Through extensive experiments, we demonstrate that an emoji based representation can be effectively learned from such noisy data. Moreover, using off-the-shelf deep neural models and without bells and whistles, we show that our emoji embedding exhibits remarkable generalization properties across domains and outperforms state-of-the-art in visual sentiment and fine-grained emotion recognition. Additionally, we introduce a new perspective on emoji interpretation using their visual emotional signature and their perceived similarity in the visual emotion space.

2. Related Work

Visual sentiment analysis While sentiment analysis from text has been extensively studied, extracting sentiment from visual data has proven to be more challenging, primarily due to the lack of large-scale datasets suited for advanced models like deep neural networks. Most available datasets are small and contains only hundreds (*e.g.* [30, 43]) or a few thousands (*e.g.* [44]) samples. Hence, many visual sentiment methods rely on hand-crafted features (*e.g.* color histograms, SIFT) to train simple models with few parameters in order to avoid the risk of overfitting the training data [28, 27, 45]. However, it is hard for such low-level features to effectively capture the higher level concept of sentiment. One way to overcome the previous problem is by learning an intermediate representation from external data that helps bridging the gap between low-level features and sentiment. For example, this can be achieved by learning an intermediate concept classifier for Adjective Noun Pairs (ANP) as in the SentiBank model [6]. However, the most common approach is to take advantage of powerful models, *i.e.* deep neural networks, in a transfer learning setting [43, 7, 42]. In this case, the neural network model is initially trained on a large-scale dataset for object classification [38]. Afterwards, the model is fine-tuned on the target task for sentiment prediction.

However, while ANP- and object-based embedding lead to improved performance, both are still not ideal for sentiment analysis. It is not clear how to select a good ANP vocabulary that can generalize well to various tasks requiring the inference of emotions from images. Additionally, object-based models are not suited for capturing sentiment since they are trained for *sentiment neutral* object classification. In this work, we propose to learn an emoji-based embedding for cross-domain sentiment and emotion analysis. Unlike objects and ANPs, emojis carry a strong sentiment signal which leads to a compact and powerful representation outperforming the previous methods as demonstrated by our evaluation.

Emojis Due to the increasing popularity of emojis, there is great interest in analyzing and studying their usage, *e.g.* [19, 24, 31, 26]. Most of this work is carried from a natural language processing (NLP) point of view, *e.g.* [5, 10]. More relevant to our work is the analysis of emojis and sentiment. Emojis can be shown to act as a strong sentiment signal that generalizes well when analyzed from a NLP perspective [37, 16, 34, 14, 36, 17]. However, whether the same can be said for a visual sentiment perspective is still to be determined. Recently, few studies attempted to learn the correlations between the emoji and image modalities. In [11], a model is developed to predict the proper emoji matching a facial expression input. On the other hand, [8] propose to handle emojis as new modality and introduce a model to predict visual or textual concepts by using emo-

jis correlations, *e.g.* learn a ship classifier by leveraging the ship emoji 🚢. In contrast to previous work, and to the best of our knowledge, this work is the first to propose emoji embedding for cross-domain visual sentiment analysis and provide an in depth analysis of their visual sentiment and emotional interpretation.

3. Emoji for Visual Sentiment Analysis

We aim in this work to learn an efficient and low-dimensional embedding of images in the emoji space. This embedding is well aligned with and encodes the visual sentiment exhibited in an image. Moreover, it can be learned efficiently from large-scale and weakly labeled data. To that end, we introduce a large-scale benchmark for visual emoji prediction (Sec. 3.1) along with deep neural model for efficient emoji embedding and transfer learning (Sec. 3.2).

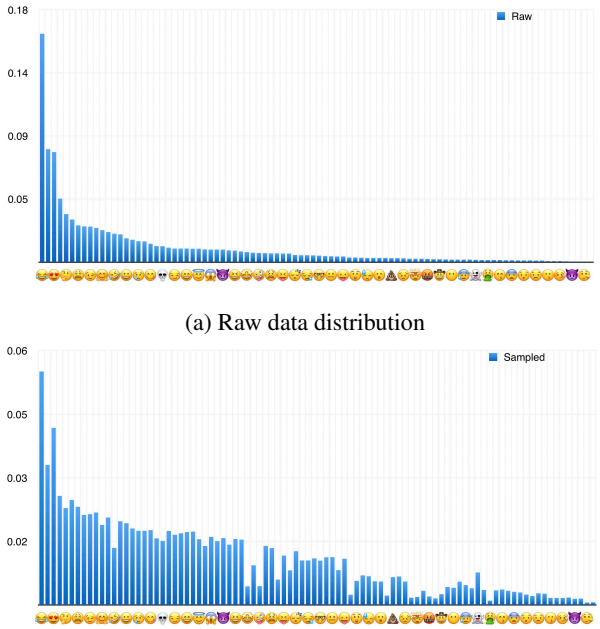
3.1. Visual Smiley Dataset

In this section, we describe our method for data collection from social media, including a) the selection of emoji categories; b) the analysis of the sample distribution; and c) a temporal sampling strategy that suits our learning task[1].

Categories The emoji list has grown from 76 entries in 1995 to 3019 in the latest *Emoji v12.0* in 2019 [41]. Many of these emojis represent objects categories (*e.g.* 📝📢🚢), abstract concepts (*e.g.* 🆘❓♉) or animals and plants (*e.g.* 🐄🦄🌵). These types of emojis are either sentiment neutral or have weak correlation with sentiment that usually arise from users cultural background or personal preferences, *e.g.* towards certain animal classes. Since our goal is to have an emoji-based representation for sentiment analysis these types are excluded from our selection. As our target categories, we chose a subset of 92 popular emojis which commonly referred to as *Smileys* (*e.g.* 😫😄😭). These smileys show a clear sentiment or emotional signal which make them adequate for our cross domain sentiment analysis. Moreover, they are among the most frequently used emojis in social media which further facilitates data collection and aids the learning process.

Sample Distribution Social media such as Instagram, Flickr and Twitter represent a rich source for large-scale emoji data. It is estimated that more than 700 million emojis are sent daily over Facebook while half the posts in Instagram contains emojis [12]. Here, we select our samples from Twitter such that we target only tweets that contain emojis and are associated with at least one image. Furthermore, to increase the relevance between the emojis and the associated image in the samples we constrain the selected tweets to those that do not contain urls, hashtags nor user mentions. This is motivated by the observation that these elements usually represent important context cues to understand the use of the selected emoji that goes beyond the associated visual data. We additionally ignore tweets that are quotes or replies to other tweets to reduce redundancy.

[1]The visual smiley dataset collected and used as part of this work will be released as a public benchmark.



(a) Raw data distribution

(b) Temporally balanced data distribution

Figure 2: Emoji frequency in (a) a raw sample of data and (b) the temporal balanced sampled dataset. Dataset (b) is used in this study.

Given the previous criteria, we retrieve a collection of 2.8 million Tweets from the first six months of 2018. Fig. 2a shows the label distribution of the data. We see that this data has a long-tail distribution and is heavily biased towards a few categories, with the top 5 most frequent emojis (*i.e.* 😂😭😍😊🤔) representing around 40% of the retrieved samples. This poses a great challenge for most standard machine learning methods as an imbalanced training dataset may lead a training process to trivially predict the most frequent labels instead of learning a more meaningful representation. Additionally, we notice that when collecting the data from a relatively short time period the content of samples tends to be heavily biased towards a few major temporal events (*e.g.* USA presidential elections or World Cup). This in turn reduces the variability of the images and hence the ability of the model to generalize well across domains.

Temporal Sampling To overcome content homogeneity, we propose to retrieve the samples from a relatively large time period while uniformly sampling the tweets from smaller temporal windows. Specifically, we collect tweets from January 1^{st} 2016 till July 31^{st} 2018. We split the time range to sequential time windows of 30 days. Fur-

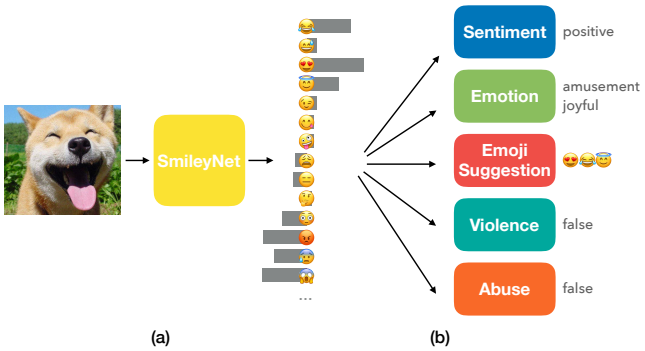


Figure 3: Our model (SmileyNet) (a) learns to embed images in the low-dimensional emoji space from large-scale and noisy data collected from social media. This embedding can subsequently be leveraged via transfer learning (b) for many target tasks in which deriving emotions from visual data is needed, such as sentiment and emotion analysis.

thermore, to alleviate label imbalance we randomly select a maximum of 4000 tweets for each emoji category within each window. We additionally allow valid samples to have a maximum number of 5 emojis, meaning that certain samples will contain multiple labels. In total, this methodology led to about 4 million images with 5.2 million emoji labels. Fig. 2b shows the label distribution of the sampled dataset. We see that compared to the raw data distribution, our dataset is more balanced across the various categories. Nonetheless, some emojis still occur relatively more often than others due to the multi-label nature of the data and the innate inter-emoji correlations.

To get a better notion of the correlation between labels, we construct the normalized correlation matrix of all emojis in the collected data[2]. As expected, by analyzing the correlation matrix we see that the two most frequent emojis 😂 and 😍 co-occur with most of the categories. Additionally, the correlation matrix reveals some semantically related groups like [😊😒😤😢😭] and [🐶🐱🐭].

3.2. Smiley Embedding Network

Given the large-scale nature of the collected dataset, it is possible to leverage deep neural network architectures for effective learning of the emoji embedding with reduced risks of data overfitting. Formally, our goal is to learn an embedding function $f(\cdot)$ that maps an image $\mathbf{x} \in \mathcal{X}^{d_x}$ to an embedding in the emoji space $\mathbf{e} \in \mathcal{E}^{d_e}$, *i.e.* $f : \mathcal{X}^{d_x} \rightarrow \mathcal{E}^{d_e}$. Such that d_x and d_e are the dimensionality of the image and emoji spaces respectively. An efficient option to realize $f(\cdot)$ is through the proxy task of explicit emoji prediction (Fig. 3a). This has two main advantages compared to other options like metric learning in the emoji space. Firstly, it is more computationally efficient compared to Siamese and Triplet networks that are usually employed for metric learning. Hence, it scales easily to large datasets while using less resources. Secondly, the learned embedding through the emoji prediction task is interpretable since each dimension in $\mathbf{e}$ corresponds to one of the emoji categories, *i.e.* $d_e = C$ where C is the number of emoji categories. This enables subsequent analysis of the embedding, better understanding of model properties, and a novel zero-shot visual sentiment learning task as we will see in Sec. 4.

To that end, we train an emoji prediction model $h(\cdot)$ such that: $h(\mathbf{x}) = \sigma(f(\mathbf{x}))$, where σ is the sigmoid activation function since our task is a multi-label classification problem. Then $h(\cdot)$ can be optimized using the binary cross entropy loss:

$$\mathcal{L}(\mathbf{x}_i, \mathbf{y}_i) = -\sum_{c=1}^{C} \mathbf{y}_{i,c}\, log(h(\mathbf{x}_i)_c), \tag{1}$$

where $\mathbf{y}_{i,c}$ is the binary label for the emoji of class c, and $h(\mathbf{x}_i)_c$ is the probability of the model predicting class c for image $\mathbf{x}_i$.

Transfer learning Once $f(\cdot)$ is trained, we can easily adapt our model across domains for a target task $g(\cdot)$ such as sentiment or emotion prediction (Fig. 3b). This is achieved through $t(\cdot)$ that maps the emoji embedding to the target label space $\mathcal{T}$, such that $g = t \circ f : \mathcal{X} \rightarrow \mathcal{E} \rightarrow \mathcal{T}$. $t(\cdot)$ is realized using a multilayer perceptron and $g(\cdot)$ can then be learned using the small training data of the target task.

4. Evaluation

We evaluate our embedding model (*SmileyNet*) for three main tasks: 1) emoji prediction which is used as a proxy to train our embedding model; and the transfer learning tasks of 2) visual sentiment analysis and 3) fine-grained emotion classification. Furthermore, 4) we introduce and analyze a novel representation for emojis that captures their unique properties in the visual sentiment space.

4.1. Emoji Prediction

Implementation Given our visual smiley dataset, we select 45 thousand images for validation and 91 thousand for testing. Samples in the validation and testing splits are balanced such that each category has around 500 and 1000 samples, respectively. We use the remaining data to train our SmileyNet model. We adopt a residual neural network with 50 layers ResNet50 [18] as the base architecture for SmileyNet. The model parameters are estimated using Adam [21] for stochastic gradient descent optimization with an initial learning rate of $1e-4$. Furthermore, we leverage data augmentation during training by randomly selecting an image crop of size 224×224 pixels with random horizontal flipping and scaling. The model is trained for $320,000$ iterations with a batch size of 128 images.

[2]See supplementary for the full correlation matrix

Model	mTop-1	mTop-3	mTop-5	AUC
Random performance	1.7	3.3	5.4	50.0
SmileyNet (Raw-Dist.)	9.5	11.6	16.3	67.6
SmileyNet (Temp-Sampling)	**11.5**	**14.4**	**19.5**	**69.8**

Table 1: Emoji prediction performance of our SmileyNet on the proposed Visual Smiley Datasets.

Evaluation metric Since emoji prediction is a multi-label task, we adopt a variant of the Top-k accuracy that accounts for the number of correct emojis in the top k predictions out of the set of ground truth emoji of each sample. Formally:

$$\text{mTop-}k_i(p_i, y_i) = \frac{|\text{ind}_\text{k}(p_i) \cap \text{ind}(y_i = 1)|}{\min(k, |\text{ind}(y_i = 1)|)}, \quad (2)$$

where $p_i = p(y|x_i)$ is the model prediction given image x_i, $\text{ind}_\text{k}(p_i)$ are the indexes of the top k predictions, and $\text{ind}(y_i = 1)$ are the indexes of the ground truth labels. Notice that here $p_i \in \mathbb{R}^C$ and $y_i \in \mathbb{R}^C$ are vectors in which $p_{i,c}$ and $y_{i,c}$ are individual entries. The final mTop-k is the average over all N samples in the test split:

$$\text{mTop-}k = \frac{1}{N}\sum_{i}^{N} \text{mTop-}k_i(p_i, y_i). \quad (3)$$

We also report the average area under curve (AUC) of the receiver operating characteristic (ROC) of all categories.

Results Along with the full model, we test two variants: 1) a random baseline and b) our SmileyNet trained with the raw emoji distribution (Raw-Dist.) without the proposed temporal sampling (Sec. 3.1). Table 1 shows the performance of these models in emoji prediction on the testing split. We notice that even with a noisy data source as social media, our model is able to predict emojis from images significantly better than a random baseline. Furthermore, our temporal sampling method leads to higher performance, *i.e.* better learned embedding, compared to a model learned with the raw and biased data distribution. In general, we see that the accuracy is relatively low. This can be attributed partly to the expected amount of noise in data annotations since it is collected automatically without any human intervention; and also to the strict evaluation metric adopted in this task which tend to underestimate the model performance. For example, a prediction of 😗 by our model for an image labeled with 😚 is considered wrong. Additionally the model needs to predict all annotated emojis for an image to get a full score on it. Nonetheless, our subsequent qualitative and transfer learning evaluation confirms that our SmileyNet in fact learns a compelling visual embedding with high performance.

Qualitative results Fig. 5 shows the top predictions of our SmileyNet for some test images from Twitter [43]. Our model produces sensible predictions that capture the general sentiment in the image. Unlike a model trained for object classification, SmileyNet output is not tailored to the object category but rather to the sentiment depicted by the object. This can be best observed by checking the model output for similar objects, like the faces, the dogs and the cars images. Our model predicts emojis of sentiment with opposite polarities when the input image is composed of sub-images (like the car accident and the child, 3^{rd} row) or when the main sentiment region is not in focus (like the image of the damaged road, 3^{rd} row). This can be related to the holistic approach of the SmileyNet. We hypothesize that an attention or region based processing might help in prioritizing the most influential image area for final predictions. Finally, predictions on images similar to those in the 4^{th} row, suggest that SmileyNet might be helpful not only for sentiment analysis but also for novel applications such as detecting violence or abuse in images.

Figure 4: Low dimensional representation using the first two principal components of the emoji embedding and the corresponding sentiment label (blue for negative and yellow for positive sentiment).

4.2. Visual Sentiment

Dataset We evaluate our model on the Twitter dataset [43]. The dataset contains 1269 images collected from Twitter and labeled manually by several annotators with positive and negative sentiment. It has 3 splits based on the degree of agreement among the annotators: "5 agrees", "4 agrees", and "3 agrees". For example, *4 agrees* split has images that at least 4 human annotators agreed upon their sentiment label.

Emojis & sentiment We use our SmileyNet to embed all images of the "5 agrees" split in the emoji space without *any* further training. Fig. 4 shows the projection of these embeddings in 2D using the first 2 dimensions of principle

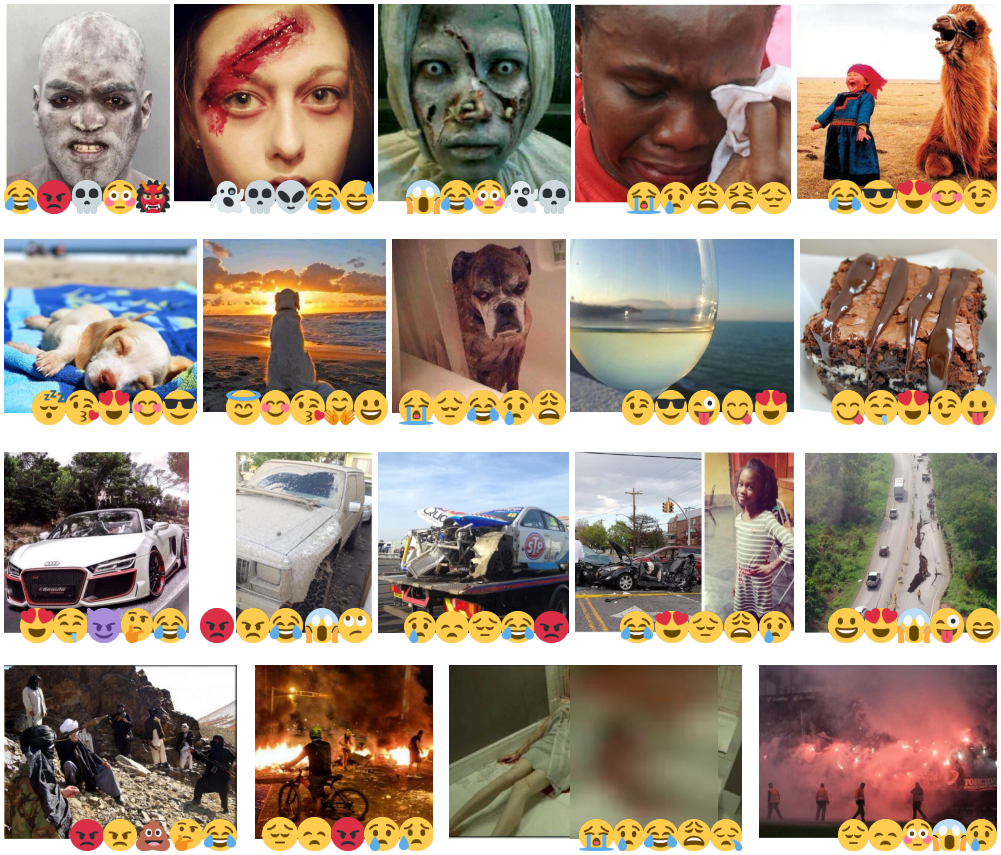


Figure 5: Qualitative results for the top 5 emojis predicted per image using our SmileyNet (ordered left to right). In contrast to a *sentiment neutral* object representation, our model produces diverse output for objects of the same category depending on the emotion conveyed in the image, *e.g.* see predictions on faces, dogs & cars in 1^{st}, 2^{nd} & 3^{rd} rows.

	Twitter Visual Sentiment [43]		
Model	3 agrees	4 agrees	5 agrees
ObjectNet	74.0	79.0	82.1
SmileyNet (ours)	**76.5**	**80.0**	**84.7**

Table 2: 1-Nearest neighbor sentiment prediction accuracy.

component analysis (PCA). One can clearly see that samples of both positive and negative sentiments are well separated in this low dimensional space. This indicates that our emoji embedding does indeed capture the visual sentiment exhibited in the image. Furthermore, using the Spearman's rank-order correlation analysis, we analyze the relations between the individual emoji dimensions and the sentiment labels. We find out that emojis with the highest correlation with the *positive* sentiment are: (😍, 0.62), (🤗,0.62), (🤩, 0.58), (😘, 0.56) and (😚, 0.53), whereas emojis with the highest correlation to *negative* sentiment are: (👿, 0.67), (😟,0.66), (😰, 0.65), (😠, 0.64) and (😖, 0.64).

Emojis & objects To evaluate the quality of the embedding quantitatively, we use 1 nearest neighbor classification and do 5-fold cross validation over the sentiment dataset for each of the 3 splits. We compare our emoji-embedding to an embedding produced by a model with the same base architecture (*i.e.* ResNet50) but trained over the ImageNet dataset (ObjectNet). As expected, our SmileyNet produces better embeddings for sentiment analysis than ObjectNet and outperforms it on all three splits (see Table 2), while SmileyNet's embedding is 10 times smaller compared to that of ObjectNet.

Transfer learning Alternatively, we can adopt a transfer learning scheme and finetune our model on the target set to see how well our model can adapt to the target data distribution from a few samples. We realize $t(\cdot)$ as a fully connected layer (Sec. 3.2) and use 5-fold cross validation to finetune and test our model as in [43].

Table 3 compares the accuracy of our model to state-of-the-art (SOTA) models. Our SimleyNet outperforms the SentiBank models [6, 9] which embed images in Adjective-Noun pairs (ANP) space that is learned as well from social media data. This indicate that emojis are better in capturing sentiment than text-based cues. We speculate emoji labeling has the advantage of being universal, finite, and offers an unambiguous one-to-one mapping between label and emo-

Model	Twitter Visual Sentiment [43] 3 agrees	4 agrees	5 agrees
PAEF [45]	67.92	69.61	72.90
SentiBank [6]	66.63	68.28	71.32
DeepSentiBank [9]	71.25	70.15	76.35
PCNN [43]	76.36	76.52	82.54
Campos *et al*. [7]	74.90	78.70	83.00
AR+Concat(K=1) [42]	77.79	83.25	86.10
AR+Concat(K=8) [42]	81.06	**85.10**	88.65
ObjectNet	78.28	82.73	87.67
SmileyNet (ours)	**82.69**	84.87	**89.16**

Table 3: State-of-the-art comparison of SimleyNet for visual sentiment prediction.

Model	Twitter Visual Sentiment [43] 3 agrees	4 agrees	5 agrees
SmileyNet - Con.	73.4	76.0	80.0
SmileyNet - Bin.	74.2	77.1	81.2

Table 4: Zero-shot visual sentiment prediction accuracy.

tion, whereas words carry rich connotations that may make the design of an effective lexicon mapping words to emotions more difficult. Moreover, our SmileyNet outperforms the advanced AR model [42] that employs a customized approach with attention mechanisms when using a single model ($K = 1$), like ours, and even when using an ensemble of $K = 8$ models. This is significant given that our model leverages off-the-shelf neural architecture and trained using noisy social media data. This further demonstrates the effectiveness of the learned embedding. We hypothesize that our model can be improved even further by employing an ensemble of models like in [42] or customized attention modules such as [13].

Zero-shot visual sentiment prediction Unlike other representations, our embedding is interpretable and each dimension can be easily related to a certain sentiment class. That is we can construct a sentiment classifier without using any training images, *i.e*. zero-shot learning (ZSL) [22, 1]. To our knowledge, ours is the *first* work to attempt ZSL for visual sentiment. We ask 4 annotators to label each of the emojis in our representation with a positive or negative sentiment based solely on the emoji's visual depiction. Then we use the average annotation as a mapping $t(\cdot)$ that will ensemble the emoji's prediction scores to estimate whether an image x has a positive or a negative sentiment. Table 4 shows the performance of our model in ZSL setting. Interestingly, while using *no training images* at all our model is still capable of producing reliable sentiment prediction that is competitive with many of the SOTA models in Table 3. We also see that using equal weighting to each emoji (the binary version "Bin.") lead to higher accuracy in comparison to using the average annotation to weight the emoji's prediction in the ensemble (the continuous model "Con.").

Emotion	Most Correlated Emojis						
amusement	😎	😆	😄	😃	😻	😀	😁
	0.31	0.30	0.29	0.27	0.27	0.26	0.26
anger	👺	😈	👹	💀	😠	🤡	💩
	0.18	0.18	0.17	0.17	0.16	0.16	0.15
awe	🙂	😲	😊	😯	😃	☺️	😃
	0.28	0.24	0.24	0.23	0.23	0.22	0.21
contentment	😘	😚	😌	😢	😁	😕	😇
	0.27	0.26	0.24	0.24	0.23	0.23	0.22
disgust	😷	🤢	😹	😵	😖	😐	😟
	0.29	0.28	0.23	0.20	0.20	0.18	0.17
excitement	🤑	😜	😝	😻	😈	😏	😛
	0.22	0.20	0.20	0.19	0.17	0.17	0.17
fear	💀	☠️	👹	😈	👺	😧	🤡
	0.21	0.18	0.17	0.17	0.16	0.16	0.15
sadness	😶	😪	😔	💀	😓	😣	😤
	0.26	0.25	0.24	0.23	0.22	0.21	0.21

Table 5: Top correlated Emojis with each emotion class.

Model	Multi-Class Emotions	Sentiment
You *et al*. [44]	48.30	-
DeepSentiBank [9]	-	61.54
PCNN [43]	-	75.34
AR+Concat(K=1) [42]	-	84.83
AR+Concat(K=8) [42]	-	86.35
ObjectNet	54.42	83.81
SmileyNet (ours)	**55.81**	**87.01**

Table 6: Fine-grained emotion classification accuracy on the Flickr&Instagram dataset [44].

4.3. Fine-grained Emotions

Dataset Finally, we evaluate our model for fine-grained emotion classification on the Flickr&Instagram dataset [44]. The dataset contains $23,308$ images queried from Flickr and Instagram and labeled by Amazon Mechanical Turk with 8 emotion classes: *amusement*, *anger*, *awe*, *contentment*, *disgust*, *excitement*, *fear* and *sadness*.

Emojis & emotions We analyze first the correlations between emojis and emotion classes. Table 5 ranks the most correlated emojis per emotion class. Interestingly, many of the top ranked emojis correspond to our intuition of the emotion depicted by the emoji's image itself. However, the ranking also reveals some unexpected correlations like 🤡 with *anger* and *fear*, 😐 with *disgust*, 😬 with *anger*, and 😪 with *sadness*. Some of these come form cultural context (like 🤡), while others we expect from common confusion of similarly looking emojis (like the sleepy face 😪 and crying face 😢).

Transfer learning Table 6 shows the performance of our SmileyNet in predicting the 8 emotion classes in a transfer learning setting. Similar to the previous section, we compare our model to ObjectNet which has been trained previously on the ImageNet dataset as it is commonly the

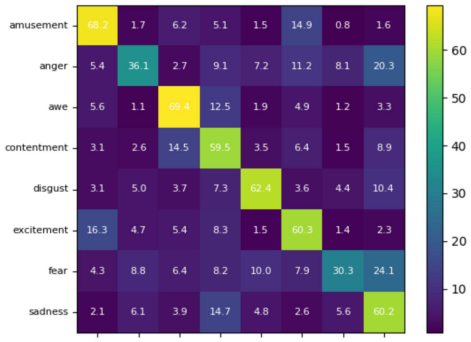
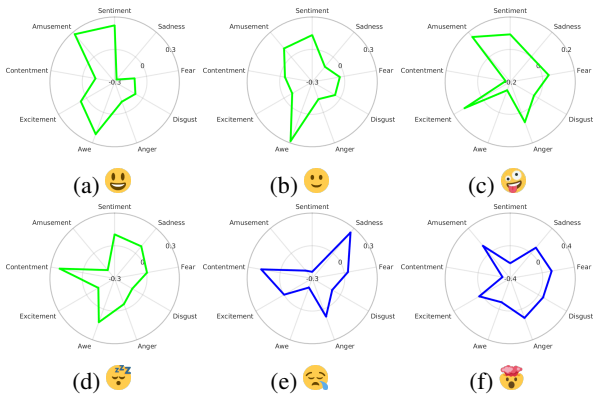


Figure 6: Confusion matrix of our SmileyNet predictions of the 8 emotion classes in Flickr&Instagram dataset.



(a) (b) (c) (d) (e) (f)

Figure 7: Emoji's emotional fingerprint. Our model reveals a unique emotional response for each emoji. Fingerprints with general positive or negative sentiment are colored with green and blue respectively.

case in literature. As hypothesized previously, SmileyNet is more suitable for fine-grained emotion prediction and outperforms a similar model transferred from an object classification task (*i.e*. ObjectNet). Moreover, our model outperforms SOTA in this task as well and shows that our compact embedding is highly effective for fine-grained emotion prediction. Fig. 6 gives us a deeper insight on the performance of each of the emotion classes. Most of the emotions are predicted with equal accuracy except for *anger* and *fear* which show high confusion with the *sadness*. Finally, similar to [42], we map the 8 emotion classes to positive and negative sentiment and report classification accuracy. Our model outperforms SOTA for this derivative task too, in accordance to our previous results from Sec. 4.2.

4.4. Emoji's Emotional Fingerprint

Given our previous analysis, we notice that each emoji in our representation has a unique signature in the emotional space. Fig. 7 shows a sample of 6 emojis and their corresponding emotional fingerprint (EEF). We see that even emoji that have similar portrayal such as 😃 & 🙂 or similar

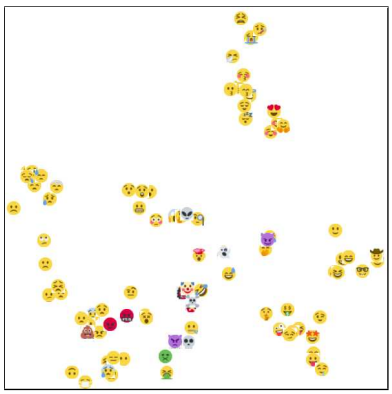


Figure 8: Low dimensional embedding of the emojis using t-SNE and based on their emotional fingerprint.

semantics like the sleepy 😪 & sleeping 😴 face have different emotional response both in intensity and bias towards certain type of emotions. Furthermore, projecting the emojis in 2D space based on their emotional fingerprints reveals further interesting findings (Fig. 8). For example, [😂🤗😈] has similar EEF, the EEF of 🙃 is closer to 😁😭 than to 🙂, and 🤨 shows bias towards anger, disgust and fear in its EEF similar to 🤬😵😟. We believe this novel representation can be of great interest for further research in behavioral studies in social media and deeper understanding of the emoji modality and its usage.

5. Conclusion

We propose to circumvent current limitations of small visual sentiment analysis datasets by learning a compact image embedding from readily available data in social media. Unlike the common object-based embedding, the proposed embedding is well aligned with the visual sentiment label space and generalizes better in transfer learning settings. Furthermore, our embedding can be efficiently learned from noisy data in social media by leveraging the intricate relation between emojis and images. To that end, we build a novel dataset, the Visual Smiley Dataset, which we use to learn an emoji-based image embedding. The evaluation on sentiment and emotion recognition shows that our low-dimensional embedding consistently outperforms the commonly used object-based embedding and the more elaborate and customized SOTA models. Furthermore, due to its interpretability we demonstrate that our embedding can be used for sentiment analysis without any further training in a zero-shot learning setting. Finally, initial results show that our embedding can aid novel applications for which inferring emotion from visual data is relevant, *e.g*. visual abuse and violence detection. We expect this work findings to be of interest not only for computer vision and visual sentiment analysis communities but also for social media studies and emoji modality understanding.

References

[1] Z. Al-Halah, M. Tapaswi, and R. Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *CVPR*, 2016. 7

[2] X. Alameda-Pineda, E. Ricci, Y. Yan, and N. Sebe. Recognizing emotions from abstract paintings using non-linear matrix completion. In *CVPR*, 2016. 1

[3] T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 2013. 2

[4] F. Barbieri, G. Kruszewski, F. Ronzano, and H. Saggion. How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics. In *ACM MM*, 2016. 2

[5] F. Barbieri, F. Ronzano, and H. Saggion. What does this emoji mean? a vector space skip-gram model for twitter emojis. In *LREC*, 2016. 2

[6] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013. 2, 6, 7

[7] V. Campos, B. Jou, and X. Giro-i Nieto. From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction. *Image and Vision Computing*, 65:15–22, 2017. 2, 7

[8] S. Cappallo, S. Svetlichnaya, P. Garrigues, T. Mensink, and C. G. Snoek. New modality: Emoji challenges in prediction, anticipation, and retrieval. *IEEE Transactions on Multimedia*, 21(2):402–415, 2018. 2

[9] T. Chen, D. Borth, T. Darrell, and S.-F. Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014. 6, 7

[10] B. Eisner, T. Rocktäschel, I. Augenstein, M. Bošnjak, and S. Riedel. emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359*, 2016. 2

[11] A. El Ali, T. Wallbaum, M. Wasmann, W. Heuten, and S. C. Boll. Face2emoji: Using facial emotional expressions to filter emojis. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2017. 2

[12] Emojipedia. Emoji statistics. https://worldemojiday.com/statistics, 2018. [Online; accessed 25-Feb-2019]. 3

[13] S. Fan, M. Jiang, Z. Shen, B. L. Koenig, M. S. Kankanhalli, and Q. Zhao. The Role of Visual Attention in Sentiment Prediction. In *ACM MM*, 2017. 7

[14] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*, 2017. 2

[15] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 2011. 1

[16] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12), 2009. 2

[17] B. Guthier, K. Ho, and A. El Saddik. Language-independent data set annotation for machine learning-based sentiment analysis. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2017. 2

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[19] R. Kelly and L. Watts. Characterising the inventive appropriation of emoji as relationally meaningful in mediated close personal relationships. *Experiences of Technology Appropriation: Unanticipated Users, Usage, Circumstances, and Design*, 2015. 2

[20] H.-R. Kim, Y.-S. Kim, S. J. Kim, and I.-K. Lee. Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Transactions on Multimedia*, 2018. 1

[21] D. P. Kingma and J. L. Ba. ADAM: A Method for Stochastic Optimization. In *ICLR*, 2015. 4

[22] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 7

[23] P. J. Lang. International affective picture system (iaps): Affective ratings of pictures and instruction manual. *Technical report*, 2005. 1

[24] L. Lebduska. Emoji, emoji, what for art thou? *Harlot: A Revealing Look at the Arts of Persuasion*, 1(12), 2014. 2

[25] B. Liu and L. Zhang. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer, 2012. 1

[26] N. Ljubešić and D. Fišer. A global analysis of emoji usage. In *Proceedings of the 10th Web as Corpus Workshop*, 2016. 2

[27] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. Li, M. G. Newman, and J. Z. Wang. On shape and the computability of emotions. In *ACM MM*, 2012. 2

[28] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 2010. 1, 2

[29] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014. 1

[30] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior research methods*, 37(4):626–630, 2005. 2

[31] H. J. Miller, D. Kluver, J. Thebault-Spieker, L. G. Terveen, and B. J. Hecht. Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication. In *ICWSM*, 2017. 2

[32] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018. 1

[33] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the International Conference on Multimodal Interaction*. ACM, 2015. 1, 2

[34] P. K. Novak, J. Smailović, B. Sluban, and I. Mozetič. Sentiment of emojis. *PloS one*, 10(12):e0144296, 2015. 2

[35] K.-C. Peng, T. Chen, A. Sadovnik, and A. C. Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *CVPR*, 2015. 1

[36] M. Rathan, V. R. Hulipalled, K. Venugopal, and L. Patnaik. Consumer insight mining: aspect based twitter opinion mining of mobile phone reviews. *Applied Soft Computing*, 68:765–773, 2018. 2

[37] J. Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL student research workshop*, 2005. 2

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1, 2

[39] S. Santhanam, V. Srinivasan, S. Glass, and S. Shaikh. I stand with you: Using emojis to study solidarity in crisis events. In *Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media*, 2018. 2

[40] D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou. Coooolll: A deep learning system for twitter sentiment classification. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, 2014. 1

[41] Unicode. Emoji list v12.0. https://unicode.org/emoji/charts/emoji-counts.html, 2019. [Online; accessed 25-Feb-2019]. 3

[42] J. Yang, D. She, M. Sun, M.-M. Cheng, P. Rosin, and L. Wang. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia*, 2018. 2, 7, 8

[43] Q. You, J. Luo, H. Jin, and J. Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, 2015. 1, 2, 5, 6, 7

[44] Q. You, J. Luo, H. Jin, and J. Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI*, 2016. 2, 7

[45] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun. Exploring principles-of-art features for image emotion recognition. In *ACM MM*, 2014. 2, 7