Cross-Media Learning for Image Sentiment Analysis in the Wild

Lucia Vadicamo
ISTI-CNR, Pisa, Italy
lucia.vadicamo@isti.cnr.it

Fabio Carrara
ISTI-CNR, Pisa, Italy
fabio.carrara@isti.cnr.it

Andrea Cimino
ILC-CNR, Pisa, Italy
andrea.cimino@ilc.cnr.it

Stefano Cresci
IIT-CNR, Pisa, Italy
stefano.cresci@iit.cnr.it

Felice Dell'Orletta
ILC-CNR, Pisa, Italy
felice.dellorletta@ilc.cnr.it

Fabrizio Falchi
ISTI-CNR, Pisa, Italy
fabrizio.falchi@cnr.it

Maurizio Tesconi
IIT-CNR, Pisa, Italy
maurizio.tesconi@iit.cnr.it

Abstract

Much progress has been made in the field of sentiment analysis in the past years. Researchers relied on textual data for this task, while only recently they have started investigating approaches to predict sentiments from multimedia content. With the increasing amount of data shared on social media, there is also a rapidly growing interest in approaches that work "in the wild", i.e. that are able to deal with uncontrolled conditions. In this work, we faced the challenge of training a visual sentiment classifier starting from a large set of user-generated and unlabeled contents. In particular, we collected more than 3 million tweets containing both text and images, and we leveraged on the sentiment polarity of the textual contents to train a visual sentiment classifier. To the best of our knowledge, this is the first time that a cross-media learning approach is proposed and tested in this context. We assessed the validity of our model by conducting comparative studies and evaluations on a benchmark for visual sentiment analysis. Our empirical study shows that although the text associated to each image is often noisy and weakly correlated with the image content, it can be profitably exploited to train a deep Convolutional Neural Network that effectively predicts the sentiment polarity of previously unseen images.

1. Introduction

Everyday, billions of user-generated contents, such as text posts and digital photos, are created and shared on social media platforms. Therefore, blogs, microblogs, and social networks are now considered vital channels for communication and information exchange. The large amount of shared data, on the one hand, is a treasure trove of people's sentiments and opinions about a vast spectrum of topics. On the other hand, it opens challenges for the exploration of large multimedia datasets that were not previously available. The possibility of analyzing online users' opinions has attracted the attention of many researchers in both academia and industry, thanks to the close correlation between user sentiment dynamics and their real life activities. So, opinion mining and sentiment analysis have been applied in a broad set of domains, such as market prediction [32], political elections [24, 35], and crisis management [3, 14], to name but a few.

From its origin, sentiment analysis has mainly relied on textual contents, while only recently researchers have begun working on automatically detecting sentiments from visual and multimodal contents [8, 10, 22, 42, 50, 49]. The visual sentiment analysis is receiving increasing attention also due the tendency for microblog users to post pictures and/or videos with short textual descriptions or no text at all. As the old saying goes "*A picture is worth a thousand words*", in fact from the earliest cave paintings to the most recent posts, sharing images is one of the most immediate and effective way to communicate. If, on one hand, people with different backgrounds can easily understand the main content of an image or video, on the other hand, emotions and sentiments that arise in the human viewer are highly subjective. Decades of research on image annotation and search have helped to mitigate the *semantic gap* problem [15, 26]. However, the subjectivity of sentiment and the *affective gap* between image features and the affective content of an image [42], make the visual sentiment analysis a very challenging task. This reflect also the difficulties in generating high-quality labeled images for training visual sentiment classifiers.

In this paper, we focus on discovering the sentiment polarity (*positive*, *negative* or *neutral*) of a given image. We follow the current trend of facing sentiment "in the wild", that is, in all sort of varying conditions of the everyday world that we share with others (as opposed to testing situation in laboratories). For this scope, social networks, such as Twitter and Facebook, are particularly suitable as sources of data for our analysis, thanks to the huge variation of contents shared by their users.

The vast majority of the existing visual sentiment classifiers have been trained on sentiment-related images, typically annotated using crowd-sourcing. As other state-of-the-art approaches, we exploit deep learning models to build our visual sentiment classifiers. However, differently from previous supervised approaches, we propose an end-to-end pipeline to train the visual classifier starting from a large-scale dataset of unlabeled tweets (text and images). First, we use a tandem Long Short Term Memory Recurrent Neural Network-Support Vector Machine (LSTM-SVM) architecture to classify the sentiment polarity of the texts. Then, we exploit the images of the tweets, labeled according to the sentiment polarity of the associated text, to fine-tune a deep Convolutional Neural Network (CNN) architecture, by leveraging on transfer learning. Although the text of the tweets is often noisy or misleading with respect to the image content (*e.g.* irrelevant comments), we show that our cross-media approach can be profitably used for learning visual sentiment classifiers in the wild. In fact, our results on a manually annotated benchmark for visual sentiment analysis show that the prediction accuracy of our trained CNN models is better than, or in the worst cases, in line with, state-of-the-art classifiers trained on sentiment-related data sets.

The main contributions of our work can be summarized as follows:

- We present an empirical study to analyze visual sentiments in the wild, starting from a large-scale dataset of *unlabeled* user-generated content. To the best of our knowledge, this is the first work that proposes a cross-media approach to learn a classifier for predicting the sentiment polarity of a given image. In particular, we show that we are able to train visual sentiment classifiers that are comparable to, or better than, state-of-the-art classifiers trained on sentiment-related labeled images.

- Overall, we collected and analysed more that 3 million tweets in order to construct the *Twitter for Sentiment Analysis* (T4SA) dataset. T4SA is composed of about 1 million high-confidence tweets for which we provide the textual sentiment classification, and the corresponding 1.4M images. We make all the collected tweets as well as the T4SA dataset publicly available to encourage further research[1].

- We publicly release our trained visual sentiment classifiers, namely *Hybrid-T4SA* and *VGG-T4SA*, and we conduct comparative studies and evaluations on benchmarks for visual sentiment analysis.

2. Related Work

For what concerns the detection of sentiment polarity in texts extracted from Twitter, first works considered simple linguistic features, such as word unigrams, word bigrams and parts-of-speech, using machine learning algorithms like Naive Bayes, Maximum Entropy and Support Vector Machines (SVM). Among the pioneers of these approaches are Go *et al*. [18] and Bermingham and Smeaton [7]. In the last few years, deep neural networks contributed to considerable performance improvements with respect to other popular learning algorithms, such as SVM. In particular, Long Short Term Memory Networks (LSTM), which we employed for this work, and Convolutional Neural Networks (CNN) were tested on several shared tasks, such as the Semeval Sentiment analysis in Twitter [34], which is considered by the Natural Language Processing community the most important benchmark for this task. The results obtained by the submitted systems, such as the ones by Deriu *et al*. [16] and Rouvier *et al*. [39], have shown that the combination of these learning techniques with the adoption of word embeddings, a compact real-value vector word representation which takes into account word similarity [31] [36], is able to set the state-of-the-art with minimal feature engineering, which makes such architectures more robust and flexible than previous ones.

While there is a large literature on extracting emotional information from textual content, research on image-based sentiment analysis is still in its early stages. First approaches to infer affects from images were based on supervised or semi-supervised frameworks that map low-level features of images into human emotions [29, 42]. To overcome the *affective gap* between low-level features and emotional content of an image, Borth *et al*. [8], Yuan *et al*. [51] and Jou *et al*. [22] employed visual entities or attributes to extract mid-level visual representations. The main contribution of [8] and [22] was building a large scale visual sentiment ontology, called VSO and Multilingual VSO, respectively, which consist of thousands of semantic objects (Adjective Noun Pairs – ANP) strongly linked to emotions. Using their ontologies they also proposed a bank of detectors, namely *SentiBank* and *MVSO*, that can automatically extract these mid-level representations. For the sentiment prediction task, a classifier is learned on the top of these mid-level representations that are used as image features.

[1] http://www.t4sa.it

This means that the ANP's textual sentiments are not explicitly used for the sentiment prediction. Just recently, Li *et al*. [27] investigated the fusion of the sentiment prediction obtained using *ANP responses as image features* with the sentiment prediction obtained using the *ANP's textual sentiment values*. The approaches used in [22, 27] relied on deep learning algorithms, which recently have significantly changed the research landscape in a broad set of domains, such as visual object recognition and image classification. Following the success of deep learning, many other approaches based on deep neural networks have been proposed for visual sentiment prediction and image emotion classification [9, 11, 20, 37, 50]. Most of these have in common the use of CNNs trained, or fine-tuned, on sentiment-related and labeled data. In [9, 11, 20, 22], state-of-the-art CNN architectures (*e.g*. AlexNet [23], PlaceCNN [52], and GoogleNet [44]) were exploited. In [50], You *et al*. proposed a custom CNN architecture specifically designed for visual sentiment prediction. Since their network was trained on weakly labeled images, they also proposed an approach, called PCNN, to reduce the impact of noisy training images. There are also several publications on analyzing sentiments using multi-modalities, such as text and image. For example, Cao *et al*. [10] proposed a late fusion is used to combine the predictions obtained from text and image, while You *et al*. [49] proposed a cross-modality consistent regression (CCR) scheme for joint textual and visual analysis. Recently, multimodal learning approaches have been proposed for joint textual and visual understanding. Baecchi *et al*. [5] proposed a multimodal feature learning schema based on CBOW and denoising autoencoders to perform sentiment analysis. In [30], the authors proposed a deep end-to-end architecture where each modality is encoded is an appropriate sub-network (an RNN for text and a CNN for images) and then fused in a multimodal layer. Similarly, Ma *et al*. [28] encodes both modalities with convolutional networks.

Differently from all the approaches previously presented, in our work we are interested in (i) understanding human sentiments by exploiting a large-scale dataset of unlabeled images collected from social-media, (ii) training sentiment classifiers without any prior knowledge of the collected data. Although textual information accompanying social media images is often incomplete and noisy, it can be effectively exploited to enable unsupervised sentiment analysis. A first step in this direction was taken in [47] where an Unsupervised SEntiment Analysis (USEA) for social-media images, based on nonnegative matrix factorization, was proposed. Our work differs from that of Wang *et al*. [47] in the way of exploiting the textual information and in the final classifier. In fact, we train a deep neural network that can be used to predict the sentiment of any new image *without* using any textual information for those images, while USEA infers sentiments for images by *jointly* considering visual and textual information.

3. Data Collection

Both the textual and the multimedia data used in this work have been collected from Twitter, by means of a streaming crawler. The data collection process took place from July to December 2016, lasting around 6 months in total. During this time span we exploited Twitter's Sample API[2] to access a random 1% sample of the stream of all globally produced tweets. All the tweets collected in this way have undergone a filtering step where we have applied a set of simple rules in order to retain only data that could be useful for our sentiment analysis task. Specifically, we discarded:

1. retweets;
2. tweets not containing any static image (i.e., we also discarded tweets containing only videos and/or animated GIFs);
3. tweets not written in the English language;
4. tweets whose text was less than 5 words long.

Rules 1 and 2 help increasing the quality of collected multimedia data. In detail, enforcing Rule 1 avoids collecting large numbers of duplicated images, while Rule 2 ensures that each collected tweet have at least a static image for the visual sentiment classification task. Rules 3 and 4 are instead aimed at guaranteeing that we have enough textual data for the textual sentiment classification task.

The above set of rules filtered as much as 98.7% of all the tweets collected from the Twitter stream. Anyway, the huge volume of tweets produced globally still allowed to collect a stream of more than 43 useful (i.e., not filtered out) tweets per minute, on average. At the end of the data collection process, the total number of tweets in our dataset is $\sim$ 3.4M, corresponding to $\sim$ 4M images. Then, we classified the sentiment polarity of the *texts* (as described in Section 4) and we selected the tweets having the most confident textual sentiment predictions to build our *Twitter for Sentiment Analysis* (T4SA) dataset. T4SA contains little less than a million tweets, corresponding to $\sim$ 1.5M images. We publicly release all the collected tweets and the T4SA dataset.

4. From Textual to Visual Sentiment Analysis

The text extracted from the collected tweets has been classified according to the sentiment polarity using an adapted version for the English language of the ItaliaNLP Sentiment Polarity Classifier [13]. This system was successfully employed in the SENTIment POLarity Classification task [6], which was organized within Evalita 2016,

[2] https://dev.twitter.com/streaming/reference/get/statuses/sample

the 5th evaluation campaign of Natural Language Processing and Speech tools for Italian. This classifier is based on a tandem LSTM-SVM architecture. SVM classification algorithms use “sparse” and “discrete” features in document classification tasks, making really hard the detection of relationships in sentences, which is often the key factor in detecting the overall sentiment polarity in documents [45]. On the contrary, LSTM networks are a specialization of Recurrent Neural Networks (RNN) which are able to capture long-term dependencies in a sentence. This type of neural network was recently tested on Sentiment Analysis and proved to outperform previous systems [34]. In this work, the tandem system uses LSTM to learn the feature space and to capture temporal dependencies, while the SVMs are used for classification. SVMs combine the document embedding produced by the LSTM in conjunction with a wide set of general-purpose features qualifying the lexical and grammatical structure of the text.

We employed a bidirectional LSTM (bi-LSTM) architecture since these kind of architecture allows to capture long-range dependencies from both directions of a document by constructing bidirectional links in the network [41].

In the training phase, the bi-LSTM network is trained considering the training documents and the corresponding gold labels. Once the statistical model of the bi-LSTM neural network is computed, for each document of the training set, a document vector (document embedding) is computed exploiting the weights that can be obtained from the penultimate network layer (the layer before the SoftMax classifier) by giving in input the considered document to the LSTM network. The document embeddings are used as features during the training phase of the SVM classifier in conjunction with a set of widely used document classification features. Once the training phase of the SVM classifier is completed the tandem architecture is considered trained. The same stages are involved in the classification phase: for each document an embedding vector is obtained exploiting the previously trained LSTM network. Finally the embedding is used jointly with other document classification features (see Section 5.2 for further details) by the SVM classifier which outputs the predicted class.

In order to evaluate the performance of our Sentiment Classifier, we performed a 10-fold cross validation over 6,293 tweets belonging to the dataset distributed for the SemEval-2013 Task on Sentiment Analysis in Twitter. We used a subset (approximately the 60%) of the original dataset since at the time we downloaded the tweets through the scripts provided by the task organizers, just a subset of all the dataset was still available in Twitter. Our classifier reported an average F1-score of 66.15. This result is particularly good considering that the winner of the competition, the NRC-Canada team [33], achieved a F1-score of 69.02 using *all* the available training data.

Our Sentiment Classifier was used to analyze the text of a large set of user-generated multimedia contents (containing both text and images). We selected data with the most confident textual sentiment predictions and we used these predictions to automatically assign sentiment labels to the corresponding images. The aim was to automatically build a training set for learning a *visual* classifier able to discover the sentiment polarity of a given image. We modeled this task as a three-way image classification problem in which each image can be classified as either *positive*, *neutral*, or *negative*. We exploited deep Convolutional Neural Networks (CNNs) as trainable classifiers due to their effectiveness in numerous vision tasks [1, 2, 23, 38, 43]. Deep CNNs allow a machine to automatically learn representations of data with multiple levels of abstraction that can be used for detection or classification tasks. A deep CNN is a feed-forward neural network composed of a possibly large number of convolutional layers with learnable filter banks that can be seen as a trainable stack of feature extractors. Each layer of a deep network extracts useful knowledge from its input to generate a feature with a higher level of abstraction, and all the layers are jointly optimized using backpropagation in order to predict difficult high level concepts directly from pixels. Convolutions are particularly suitable for visual data, since they are able to model the spatial correlation of neighboring pixels better than normal fully connected layers. For a classification problem, the final outputs of the CNN are the confidences for each class the network has been trained on.

To build our visual classifier, we leveraged on transfer learning. We used a balanced subset of our dataset T4SA to fine-tune known and successful deep CNNs architectures pretrained on generic datasets of images. Doing so, we are able to exploit additional knowledge already stored in the trained network while training for sentiment prediction. In particular, we used *HybridNet* [52] and *VGG-19* [43] models. In Section 5.2, we describe the fine-tuning process used for building our visual sentiment classifiers.

5. Experimental Evaluation

5.1. Dataset Preparation

All the ∼3.4M tweets collected as reported in Section 3 were analyzed by the textual sentiment polarity classifier described in Section 4. In order to produce a reliable dataset for learning a visual sentiment classifier, we selected only the tweets classified with a confidence $\geqslant 0.85$[3]. The resulting dataset contains $371,341$ Positive, $629,566$ Neutral, and $31,327$ Negative tweets. As expected, the dataset proved to be very imbalanced, a frequent and known issue in social media sentiment data [25]. In order to increase

[3] Using this threshold, the classifier achieves state of the art accuracy of 0.71 in terms of F-score.

Sentiment	**T4SA**		**T4SA** w/o near-duplicates	**B-T4SA**
	(tweets)	*(images)*	*(images)*	*(images)*
Positive	371,341	501,037	372,904	156,862
Neutral	629,566	757,895	444,287	156,862
Negative	179,050	214,462	156,862	156,862
Sum	904,395	1,473,394	974,053	470,586

Table 1. Our Twitter for Sentiment Analysis (T4SA) dataset and its subsets used for learning our visual classifiers. Each tweet (text and associated images) is labeled according to the sentiment polarity of the text, predicted by our tandem LSTM-SVM architecture.

the number of Negative tweets we selected a lower filtering threshold for this class, obtaining $179,050$ examples. Notably, in our experiments conducted on T4SA, the difference in precision on the classification of positive, neutral, and negative never exceeds 1% and thus, the lower threshold used for selecting negative examples did not impact the quality of the learning. Starting from this dataset, we selected a balanced subset of images to train visual sentiment classifiers. To do so, we performed the following steps:

- We labeled each image of T4SA on the basis of the corresponding textual sentiment classification.
- We removed corrupted and near-duplicate images resulting in $\sim$ 974K unique images.
- We selected a *balanced* subset composed by $156,862$ images for each class, resulting in $470,586$ images. We call this subset B-T4SA.
- We split B-T4SA in training, validation and test subsets, corresponding approximately to 80%, 10%, and 10% of the images.

Details on T4SA and its subsets are summarized in Table 1. Notice that the size of the balanced subset (i.e., B-T4SA) was highly influenced by the low number of tweets classified as negative. Moreover, these negatives contain many artificial images (*e.g*. screenshots and memes), which made our analysis more challenging. In fact, we encountered difficulties in automatically collecting negative tweets that also contain natural images by using only a random sample of all globally produced tweets. However, in order to avoid possible biases in our analyses, we deliberately avoided to use any keyword during the data collection process.

5.2. Experimental Settings

Text Analysis Settings As described in Section 4, we employed a bidirectional LSTM to learn a document embedding into a feature space. We applied a dropout factor of 0.45 to both input gates and to the recurrent connections in order to prevent overfitting, a typical issue in neural networks [17]. For what concerns the optimization process, categorical cross-entropy is used as a loss function and optimization is performed by the *rmsprop* optimizer [46]. We used the Keras [12] deep learning framework to develop the LSTM network.

Each input word to the LSTM is represented by a low dimensional, continuous and real-valued vector, also known as word embedding [31], and all the word vectors are stacked in a word embedding matrix. For this work, we used GloVe [36] pre-trained vectors since these are computed considering the word context information. GloVe website provides freely available pre-trained vectors computed from a 2B English tweets corpus.

The document embedding produced by the LSTM is used in conjunction with other document features by the SVM classifier. The other document features focused on a wide set of features ranging across different levels of linguistic description. The features are organised into three main categories: raw and lexical text features, morpho-syntactic features and lexicon features. With the exception of the lexicon features, these features were already tested and described in [13]. To extract the lexicon features we exploited three freely available resources: The Bing Liu Lexicon [19], which includes approximately 6,000 English words, the MultiPerspective Question Answering Subjectivity Lexicon [48], which consists of approximately 8,200 English words, and the SentiWordNet 3.0 Lexicon [4] that consists of more than 117,000 words. For each word in these lexicons the associated polarity is provided. In addition, we manually developed a lexicon of positive and negative emoticons, which are usually a strong indicator of tweet polarity. By exploiting the described resources, the following features were extracted: positive/negative emoticon distribution, sentiment polarity n-grams, sentiment polarity modifiers, the distribution of sentiment polarity, the most frequent sentiment polarity and changes of polarity in tweet sections. The last lexicon feature is calculated using the word embedding produced by Glove and it is obtained by computing separately the average of the word embeddings of the nouns, adjectives, and verbs of the tweet.

Image Analysis Settings We used B-T4SA training subset to fine-tune two different pretrained networks, namely: AlexNet pretrained on ILSVRC2012 [40] + Places205 [52] (also called HybridNet [52]) and VGG-19 [43] pretrained on ILSVRC2012 [4]. We replaced the last fully connected layer *fc8* with a new one having 3 outputs, and we experimented two different fine-tuning strategies we named

[3] http://nlp.stanford.edu/projects/glove/

[4] Both the pre-trained models can be downloaded from the *Caffe Mode Zoo* (`http://caffe.berkeleyvision.org/model_zoo.html`)

Sentiment	**Twitter Testing Dataset**		
	5 agree	**$\geq$ 4 agree**	**$\geq$ 3 agree**
Positive	581	689	769
Negative	301	427	500
Sum	882	1,116	1,269

Table 2. Twitter Testing Dataset [50].

Model	**Twitter Testing Dat.** *(pos, neg)*			**B-T4SA** test set *(pos, neu, neg)*
	5 agree	**$\geq$ 4 agree**	**$\geq$ 3 agree**	
Random Classifier	0.500	0.500	0.500	0.333
CNN [50]	0.722	0.686	0.667	-
PCNN [50]	0.747	0.714	0.687	-
Hybrid-T4SA FT-F	0.766	0.748	0.723	0.499
Hybrid-T4SA FT-A	0.741	0.709	0.686	0.491
VGG-T4SA FT-F	0.768	0.737	0.715	0.506
VGG-T4SA FT-A	**0.785**	**0.755**	**0.725**	**0.513**

Table 3. Prediction accuracy on the different test sets.

FT-F and *FT-A*. In *FT-A* we fine-tune all the trainable layers, while in *FT-F* the parameters of convolutional layers are fixed and only the last fully connected layers *fc6-8* are trained.

We trained both networks with both the fine-tune strategies using Caffe [21] for 15 epochs using SGD with momentum $\mu = 0.9$, a learning rate of 0.001 divided by 10 every 5 epochs, and L2 regularization with a weight decay of 10^{-5}. For HybridNet we used a batch size of 128, while for VGG-19 we used a batch size of 32 and batch accumulation of 2 to lower the GPU memory footprint.

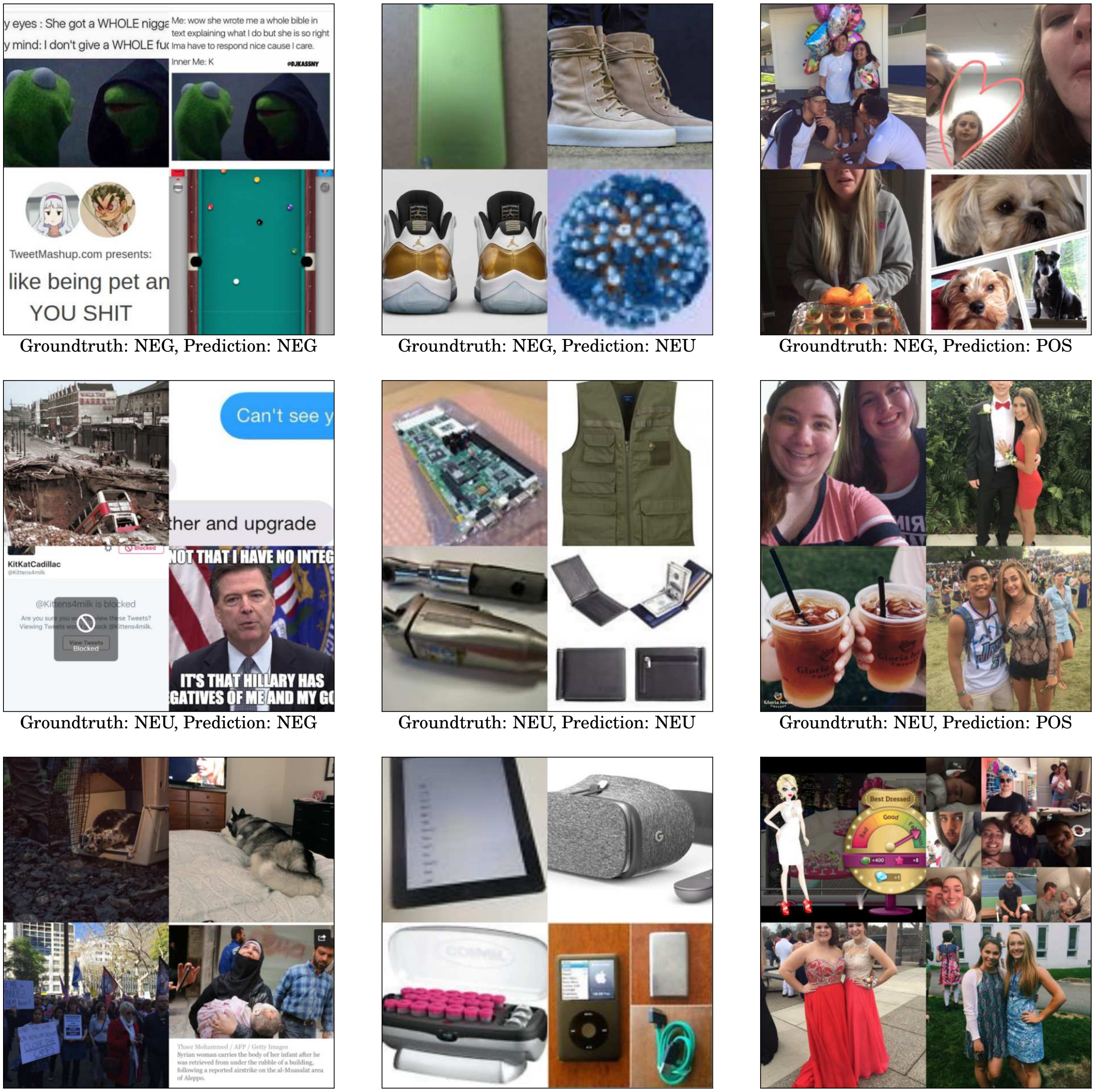
We assessed the performance of our models using our B-T4SA test set and the so-called Twitter Testing Dataset presented in [50]. The latter contains a total of $1,269$ images having a positive or negative sentiment that have been manually labeled by five Amazon Mechanical Turk (AMT) workers. The images are partitioned into three subsets, namely “Five agree”, “At least four agree”, and “At least three agree”, on the basis of the labeling results from the five AMT workers. The details of this dataset are reported in Table 2. Twitter Testing Dataset was also used in [9, 20, 27, 50], allowing for a through comparison of our systems with the state-of-the-art.

5.3. Results

In Table 3, we reported the accuracy obtained by the fine-tuned models in our B-T4SA test set and in all the subsets provided by the Twitter Testing Dataset. As baseline references, we report the results obtained by a random classifier and by the CNN and PCNN models [50]. The results show that the models trained with our cross-media approach effectively classify the images of Twitter Testing Dataset which were manually annotated. In particular, our best model (VGG-T4SA FT-A) correctly classifies 78.5% of the *five agree* testing images, outperforming similar models trained on high-quality sentiment-related hand-labeled data [50]. Since the Twitter Testing Dataset only provides binary labels (positive/negative) as groundtruth, we obtained a binary classification from our three-way model taking the maximum confidence between the positive and the negative confidences. We also performed experiments using two-way fine-tuned nets trained only on positives and negatives images provided by our training set, but we observed that there is no significant difference in performance with respect to predictions derived from three-way models.

Notice that the range of the accuracy values obtained on our B-T4SA is lower because those experiments concern a three-way classification. Moreover, the groundtruth for the evaluation is not hand-labeled since it is derived from the analysis of the textual information collected “in the wild”, thus it contains noisy labels that inevitably lower the accuracy upper bound. Figure 1 reports the most confident classifications on this test set. From a qualitative point of view, in several cases the sentiment polarity of the images is better represented by the prediction of our model than the sentiment polarity inferred from the text associated to the images. Since the textual sentiment analysis was used to label both testing and training images, this, on one hand, explains the relatively lower accuracy obtained by our models on the B-T4SA test set and, on the other hand, confirms that we used sufficiently large set of training data to allow the deep neural network to handle noisy labels. Moreover, we observed that the use of the *neutral* class is particularly suitable for analyzing web images, which often depict simple objects or commercial products. For this reason we used a three-way model, unlike other state-of-the-art approaches which use a binary classification model [9, 20, 50].

Many papers on visual sentiment analysis report the 5-fold cross-validation accuracy on the Twitter Testing Dataset [9, 20, 27, 50], in which the prediction of each fold is computed with a model fine-tuned on the other four folds. In order to compare to those approaches, we also tested our models in this setting, and we reported the 5-fold accuracy obtained in Table 4. However, we think that this measure is inappropriate for our cross-media approach since it tends to highlight how well a pretrained model adapts to a specific task or dataset. In fact, as evidenced by the results, models trained on generic (not sentiment-related) datasets, like AlexNet and VGG-19, not necessarily perform worse than the same models previously finetuned on a sentiment-related dataset. For example our best approach (VGG-T4SA FT-A), achieves a 5-fold accuracy of 0.896 on the five agree subset, that corresponds only to an improvement of 1.5% with respect to VGG-19 trained on a generic


y eyes : She got a WHOLE nigg
y mind: I don't give a WHOLE fu
Me: wow she wrote me a whole bible in text explaining what I do but she is so right Ima have to respond nice cause I care.
Inner Me: K
@DJKASSNY
TweetMashup.com presents:
like being pet an
YOU SHIT
Can't see y
ther and upgrade
KitKatCadillac
@Kittens4milk is blocked
NOT THAT I HAVE NO INTEG
IT'S THAT HILLARY HAS
GATIVES OF ME AND MY G
Best Dressed
Good
Thaer Mohammed / AFP / Getty Images
Syrian woman carries the body of her infant after he was retrieved from under the rubble of a building, following a reported airstrike on the al-Muasalat area of Aleppo.

Groundtruth: NEG, Prediction: NEG

Groundtruth: NEG, Prediction: NEU

Groundtruth: NEG, Prediction: POS

Groundtruth: NEU, Prediction: NEG

Groundtruth: NEU, Prediction: NEU

Groundtruth: NEU, Prediction: POS

Groundtruth: POS, Prediction: NEG

Groundtruth: POS, Prediction: NEU

Groundtruth: POS, Prediction: POS

Figure 1. The most confident classifications of our model on our B-T4SA test set, grouped by all possible (groundtruth, predicted class) couples. Rows (from top to bottom) contains images labeled respectively *negative*, *neutral* and *positive* on the basis of textual sentiment analysis. Columns (from left to right) contain images visually classified respectively as *negative*, *neutral* and *positive* by our model.

dataset. Moreover, our technique is based on a cross-media approach, i.e. it relies on labels not coming from a visual inspection of the images. Thus, we think that a fine-tuning of our models on manually labeled images is inappropriate for our goal. In any case, also in this test setting, our models outperform other state-of-the-art visual classifiers, such as PCNN, DeepSentiBank and MVSO, which were trained on high-quality sentiment-related dataset.

Taken together, these results indicate that our cross-media learning approach is a first, important step towards building systems able to learn the sentiment polarity of images autonomously from the Web.

Method	**Training details**	**Twitter Testing Dataset**		
		5 agree	**≥ 4 agree**	**≥ 3 agree**
Approaches without intermediate fine-tuning				
GCH [42] (res from [50]) *	-	0.684	0.665	0.66
SentiBank [8] (res from [50]) ○	-	0.709	0.675	0.662
LCH [42] (res from [50]) *	-	0.710	0.671	0.664
GCH+ BoW [42] (res from [50]) *	-	0.710	0.685	0.665
LCH+ BoW [42] (res from [50]) *	-	0.717	0.697	0.664
Sentribute [51] (res from [50]) ○	-	0.738	0.709	0.696
CNN [50] •	Custom architecture *tr* on Flickr (VSO) [8]	0.783	0.755	0.715
AlexNet [23] (res from [9]) •	AlexNet [23] *tr* on ILSVRC2012 [40]	0.817	0.782	0.739
PlaceCNN [52] (res from [9]) •	AlexNet [23] *tr* on Places205 [52]	0.830	-	-
GoogleNet [44] (res from [20]) •	GoogleNet [44] *tr* on ILSVRC2012 [40]	0.861	0.807	0.787
HybridNet •	AlexNet [23] *tr* on (ILSVRC2012 [40] + Places205 [52])	0.867	0.814	0.781
VGG-19 •	VGG-19 [43] *tr* on ILSVRC2012 [40]	0.881	0.835	0.800
Approaches using an intermediate fine-tuning				
PCNN [50] •	Custom architecture *tr* on Flickr (VSO) [8] + *ft* on Flickr (VSO) [8]	0.773	0.759	0.723
DeepSentiBank [11] (res from [9]) ○•	AlexNet [23] *tr* on ILSVRC2012 [40] + *ft* on Flickr (VSO) [8]	0.804	-	-
MVSO [EN] [22] (res from [9]) ○•	DeepSentiBank [11] *ft* on MVSO-EN [22]	0.839	-	-
Hybrid-T4SA FT-A (Ours) •	AlexNet [23] *tr* on (ILSVRC2012 [40] + Places205 [52]) + *ft* on B-T4SA	0.864	0.830	0.800
Hybrid-T4SA FT-F (Ours) •	AlexNet [23] *tr* on (ILSVRC2012 [40] + Places205 [52]) + *ft* on B-T4SA	0.873	0.832	0.810
VGG-T4SA FT-F (Ours) •	VGG-19 [43] *tr* on ILSVRC2012 [40] + *ft* on B-T4SA	0.889	0.857	0.815
VGG-T4SA FT-A (Ours) •	VGG-19 [43] *tr* on ILSVRC2012 [40] + *ft* on B-T4SA	**0.896**	**0.866**	**0.820**

* Approch based on low-level features
○ Approch based on mid-level features
• Approch based on deep learning

Table 4. 5-Fold Cross-Validation Accuracy of different methods on Twitter Testing Dataset. *tr* stands for 'trained'; *ft* stands for 'fine-tuned'. Note that in these experiments *all* the deep models are again fine-tuned on four folds of the Twitter Testing Dataset. During cross-validation we fine-tuned all the weights of our FT models.

6. Conclusions

This application paper deals with the problem of training a visual sentiment classifier from a large set of multimedia data, without the need of human annotators. We leveraged on a cross-media learning approach showing that even if the textual information associated to Web images is often noisy and ambiguous, it is still useful for learning robust visual classifiers. To this scope, we collected and used more than 3 million tweets, and we experimentally shown that our approach is effective for learning visual sentiment classifier in the wild. We publicly released all the collected data and our trained models for future research and applications.

Acknowledgments

This work was partially founded by *Smart News, Social sensing for breakingnews*, co-founded by the Tuscany region under the FAR-FAS 2014 program, CUP CIPE D58C15000270008. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

[1] G. Amato, F. Carrara, F. Falchi, C. Gennaro, C. Meghini, and C. Vairo. Deep learning for decentralized parking lot occupancy detection. *Expert Syst. Appl.*, 72:327 – 334, 2017. 4

[2] G. Amato, F. Falchi, and L. Vadicamo. Visual recognition of ancient inscriptions Using Convolutional Neural Network and Fisher Vector. *JOCCH*, 9(4):21:1–21:24, Dec. 2016. 4

[3] M. Avvenuti, S. Cresci, F. Del Vigna, and M. Tesconi. Impromptu crisis mapping to prioritize emergency response. *Computer*, 49(5):28–37, 2016. 1

[4] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC 2010*. 5

[5] C. Baecchi, T. Uricchio, M. Bertini, and A. Del Bimbo. A multimodal feature learning approach for sentiment analysis of social network multimedia. *Multimedia Tools and Applications*, 75(5):2507–2525, 2016. 3

[6] F. Barbieri, V. Basile, D. Croce, M. Nissim, N. Novielli, and V. Patti. Overview of the evalita 2016 sentiment polarity classification task. In *EVALITA 2016*. 3

[7] A. Bermingham and A. F. Smeaton. Classifying sentiment in microblogs: is brevity an advantage? In *CIKM 2010*. ACM. 2

[8] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Multimedia 2013*. ACM. 1, 2, 8

[9] V. Campos, B. Jou, and X. Giró i Nieto. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. *Image and Vision Computing*, 2017. 3, 6, 8

[10] D. Cao, R. Ji, D. Lin, and S. Li. A cross-media public sentiment analysis system for microblog. *Multimedia Systems*, 22(4):479–486, 2016. 1, 3

[11] T. Chen, D. Borth, T. Darrell, and S. Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *CoRR*, abs/1410.8586, 2014. 3, 8

[12] F. Chollet. Keras. https://github.com/fchollet/keras, 2015. 5

[13] A. Cimino and F. Dell'Orletta. Tandem LSTM-SVM approach for sentiment analysis. In *EVALITA 2016*. 3, 5

[14] S. Cresci, M. Tesconi, A. Cimino, and F. Dell'Orletta. A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages. In *WWW Companion 2015*. ACM. 1

[15] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):5:1–5:60, May 2008. 1

[16] J. Deriu, M. Gonzenbach, F. Uzdilli, A. Lucchi, V. D. Luca, and M. Jaggi. Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In *SemEval @ NAACL-HLT 2016*. 2

[17] Y. Gal and Z. Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *NIPS 2016*. 5

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

[19] M. Hu and B. Liu. Mining and summarizing customer reviews. In *SIGKDD 2004*. ACM. 5

[20] J. Islam and Y. Zhang. Visual sentiment analysis for social images using transfer learning approach. In *BDCloud-SocialCom-SustainCom 2016*. IEEE. 3, 6, 8

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 6

[22] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S.-F. Chang. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Multimedia 2015*. ACM. 1, 2, 3, 8

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS 2012*. 3, 4, 8

[24] M. Laver, K. Benoit, and J. Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331, 005 2003. 1

[25] S. Li, Z. Wang, G. Zhou, and S. Y. M. Lee. Semi-supervised learning for imbalanced sentiment classification. In *IJCAI 2011*. 4

[26] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. M. Snoek, and A. D. Bimbo. Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *ACM Comput. Surv.*, 49(1):14:1–14:39, June 2016. 1

[27] Z. Li, Y. Fan, W. Liu, and F. Wang. Image sentiment prediction based on textual descriptions with adjective noun pairs. *Multimedia Tools and Applications*, pages 1–18, 2017. 3, 6

[28] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *ICCV 2015*, pages 2623–2631. IEEE, 2015. 3

[29] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Multimedia 2010*. ACM. 2

[30] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014. 3

[31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS 2013*. 2, 5

[32] G. Mishne, N. S. Glance, et al. Predicting movie sales from blogger sentiment. In *Computational Approaches to Analyzing Weblogs 2006*. AAAI. 1

[33] S. M. Mohammad, S. Kiritchenko, and X. Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *SemEval @ NAACL-HLT 2013*. 4

[34] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov. Semeval-2016 task 4: Sentiment analysis in twitter. In *SemEval @ NAACL-HLT 2016*. 2, 4

[35] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM 2010*. AAAI. 1

[36] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP 2014*. 2, 5

[37] T. Rao, M. Xu, and D. Xu. Learning multi-level deep representations for image emotion classification. *arXiv preprint arXiv:1611.07145*, 2016. 3

[38] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *CVPRW 2014*, pages 512–519. IEEE, 2014. 4

[39] M. Rouvier and B. Favre. SENSEI-LIF at semeval-2016 task 4: Polarity embedding fusion for robust sentiment analysis. In *SemEval @ NAACL-HLT 2016*. 2

[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 5, 8

[41] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. 4

[42] S. Siersdorfer, E. Minack, F. Deng, and J. Hare. Analyzing and predicting sentiment of images on the social web. In *Multimedia 2010*. ACM. 1, 2, 8

[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 4, 5, 8

[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR 2015*. IEEE. 3, 8

[45] D. Tang, B. Qin, and T. Liu. Document modeling with gated recurrent neural network for sentiment classification. In *EMNLP 2015*. 4

[46] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012. 5

[47] Y. Wang, S. Wang, J. Tang, H. Liu, and B. Li. Unsupervised sentiment analysis for social media images. In *IJCAI 2015*. 3

[48] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT-EMNLP 2005*. 5

[49] Q. You, J. Luo, H. Jin, and J. Yang. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *WSDM 2016*. ACM. 1, 3

[50] Q. You, J. Luo, H. Jin, and J. Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. *CoRR*, abs/1509.06041, 2015. 1, 3, 6, 8

[51] J. Yuan, S. Mcdonough, Q. You, and J. Luo. Sentribute: Image sentiment analysis from a mid-level perspective. In *WISDOM @ SIGKDD 2013*. ACM. 2, 8

[52] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS 2014*. 3, 4, 5, 8