Multimedia Tools and Applications
https://doi.org/10.1007/s11042-019-08312-7

Exploiting objective text description of images for visual sentiment analysis

Alessandro Ortis[1] · **Giovanni Maria Farinella**[1] · **Giovanni Torrisi**[2] · **Sebastiano Battiato**[1]

Received: 23 July 2018 / Revised: 18 July 2019 / Accepted: 1 October 2019 /
Published online: 07 January 2020

Abstract

This paper addresses the problem of Visual Sentiment Analysis focusing on the estimation of the polarity of the sentiment evoked by an image. Starting from an embedding approach which exploits both visual and textual features, we attempt to boost the contribution of each input view. We propose to extract and employ an *Objective Text* description of images rather than the classic *Subjective Text* provided by the users (i.e., title, tags and image description) which is extensively exploited in the state of the art to infer the sentiment associated to social images. *Objective Text* is obtained from the visual content of the images through recent deep learning architectures which are used to classify object, scene and to perform image captioning. *Objective Text* features are then combined with visual features in an embedding space obtained with Canonical Correlation Analysis. The sentiment polarity is then inferred by a supervised Support Vector Machine. During the evaluation, we compared an extensive number of text and visual features combinations and baselines obtained by considering the state of the art methods. Experiments performed on a representative dataset of 47235 labelled samples demonstrate that the exploitation of *Objective Text* helps to outperform state-of-the-art for sentiment polarity estimation.

Keywords Visual sentiment analysis · Objective text features · Embedding spaces · Social media

✉ Alessandro Ortis
ortis@dmi.unict.it

Giovanni Maria Farinella
gfarinella@dmi.unict.it

Giovanni Torrisi
giovanni.torrisi@telecomitalia.it

Sebastiano Battiato
battiato@dmi.unict.it

[1] University of Catania, Viale A. Doria 6, Catania 95125, Italy

[2] JOL Catania - Telecom Italia, Viale A. Doria 6, Catania 95125, Italy

1 Introduction

The rise of social media has opened new opportunities to better understand people's interests towards topics, brands or products. Social media users continuously post images together with their opinions and share their emotions. This trend has supported the growing of new Machine Learning application areas, such as semantic-based image selection from crowd-sourced collections [3, 37], Social Event Analysis [34] and Sentiment Analysis on Visual Contents [8]. As well as the definition of new approaches to address classic tasks such as products rating prediction [26] and election forecasting [45], based on the Web contents publicly shared by users. Visual Sentiment Analysis aims to infer the sentiment evoked by images in terms of positive or negative polarity. Early methods in this field focused only on visual features [28, 48] (ignoring the text associated to the images) or have employed text to define a sentiment ground truth [5, 38]. More recent approaches exploit a combination of visual and text features in different ways. Most of them, consider SentiWordNet [11] and the WordNet [30] lexicons as external knowledge to extract useful semantic information from textual data. In particular, SentiWordNet provides three types of sentiment polarity scores for each word defined in WordNet [30], and describes how much positive or negative are the terms.

In this paper we propose to exploit the text automatically extracted from images to build an embedding space where the correlation among visual and textual features is maximized. Several previous works define models which learn a joint representation over multimodal inputs (i.e., text, image, video, and audio) to perform Image Classification [49], Visual Sentiment Analysis [25], image retrieval [9, 12, 13, 19, 20, 32, 36, 50], and event classification [44] by exploiting social media contents. The text associated to images is typically obtained by considering the meta-data provided by the user (e.g., image title, tags and description). Differently than previous approaches, our framework describes images in an "objective" way through state-of-the-art scene understanding methods [24, 39, 51]. Since the text describing the images is automatically inferred, in our approach, we denote it as "objective" emphasizing the fact that it is different to the "subjective" text written by the user for a visual content (i.e., image) of a social media post.

In [25] two different datasets are considered, by crawling public images from Instagram and Flickr respectively. To represent contents for sentiment analysis estimation, the authors proposed three different type of features extracted considering pairs of images and the related subjective texts: a visual feature defined by combining different visual descriptors usually used for visual classification [12–14], a feature obtained by using the traditional Bag of Words approach on the subjective text, and a sentiment feature obtained by selecting the words of the subjective text whose sentiment scores (positive or negative) reported in SentiWordNet [11] are larger than a threshold, and applying the Bag of Words on this restricted vocabulary. These three types of features, called views, are then combined to form an embedding space by using multi-view Canonical Correlation Analysis (CCA) [16]. The aforementioned features projected to the computed embedding space are then exploited to train a binary classifier which is used to infer the final positive or negative sentiment (i.e., sentiment polarity).

Although the subjective text associated to social images can be exploited as additional source of input to infer the sentiment polarity of an image, two different users could associate very different texts to the same image. This makes the features extracted from the subjective text prone to be noisy. In the definition of the dataset used in [25] the authors observed that the tags associated to social images can be very noisy, and for this reason they avoided to exploit the textual data for the definition of the sentiment ground

truth of their experimental dataset (i.e., they decided to build the ground truth by manual labelling).

The authors of [25] compared a pool of representations usually used for the task of Visual Sentiment Analysis [5, 29, 38, 42] mainly based on hand-crafted visual features and textual information provided by users (i.e., subjective), using the same evaluation protocol and the same dataset. In order to perform a fair comparison with respect to the state of the art, we considered the same dataset used in [25] as well as the same evaluation protocol. Differently than [25], we built the textual and sentiment views by exploiting the objective text as input instead of the subjective text provided by users, with the aim to assess the benefits of using the proposed source of text in lieu of the text commonly used in previous approaches. To this aim, we exploited four state of the art deep learning architectures to automatically extract the objective text from the input images. To further assess the effectiveness of our approach, we have also considered different combinations of subjective, objective text and visual features for the definition of the embedding space to be used for sentiment polarity estimation. Since the visual representations used in the state of the art are based on the combination of hand crafted features (e.g., GIST, color histograms, etc.), in the second part of the experimental evaluation we consider the possibility to further boost the performances of the system by exploiting a deep visual representation. To this aim, for each considered deep architecture (*GoogLeNet* [39], *Places205* [51] and *DeepSentiBank* [8][1]) we extracted an internal representation of the input image and trained an SVM for the task of polarity prediction (i.e., binary classification). The results of this experiment provide another strong baseline for the performance evaluation. Deep visual features have been combined with objective text features for comparative evaluation with respect to the baseline (i.e., deep feature alone). We selected the best performing deep visual feature and performed the evaluation pipeline of the proposed approach considering this stronger visual feature.

Differently than common methods, in the proposed system the input text is automatically extracted from the images, by exploiting several deep learning architectures. Based on the extracted text, we built three textual features that are combined in different ways with the visual and subjective text descriptors to obtain different embedding spaces. The contributions of this paper are the following

- we first highlight and then experimentally demonstrate the weakness of the subjective text associated to images usually provided by users for the sentiment prediction task;
- we propose an alternative source of text, which is user-independent. For this reason we refer to it as *Objective Text*. Experimental results demonstrate the effectiveness of such textual data, which allows to obtain the best results compared with several baseline and state-of-the-art approaches;
- considering the proposed source of text, we evaluate several number of combinations of textual and visual features. Furthermore, for each experimental setting we evaluated the possibility to reduce the dimensionality of the exploited features by a truncation strategy which keeps the 99% of the original information;
- in the second part of the evaluation, we attempt to further improve the performances of the proposed system by employing a deep based visual feature together with objective text. To properly select the deep representation, we performed a comparative evaluation of three state-of-the-art deep architectures.

[1] Our implementation exploits the *MVSO English* model provided by [23], that corresponds to the *DeepSentiBank* CNN fine-tuned to predict 4342 English Adjective Noun Pairs.

This work extends our previous work in [31], by including a more detailed analysis description of the experiments performed in [31], as well as extended experiments that evaluate the combination of the Objective Features with deep based visual features. The feature evaluation performed in this paper focuses on the task of Visual Sentiment Analysis, however the observations and the achieved insights result useful also to other systems which exploit the text associated to social images. The paper is organized as follows. In Section 2 a brief review of the state of the art in this context is presented. Section 3 describes the features we have used to infer the sentiment polarity. Section 4 details the experimental settings and discusses the results. Finally, Section 5 concludes the paper.

2 Related works

The aim of Visual Sentiment Analysis is to infer the sentiment polarity associated to images in terms of positive or negative engagement. Most of the early works in this field try to associate low-level visual features to sentiments. These works have been influenced by empirical studies in the context of psychology, art and image aesthetics [4, 10, 21, 40].

Recently, the rise of social media provides huge amount of pictures with user generated accompanying text, such as title, description, tags and comments. This allows the analysis, by Machine Learning approaches, of huge amount of real-word images published on social media by users. Several papers investigated the problem of joint modelling the representation of Internet images and associated text or tags for different tasks, such as image retrieval [13, 27, 32], social images understanding [3], image annotation [22] and visual sentiment analysis [2, 5, 25, 38].

The authors of [38] studied the correlations between the sentiment evoked by images and their visual content with the aim to classify images as positive or negative. They used the thesaurus SentiWordNet [11] to extract numerical sentiment values from Flickr metadata (e.g., title, description and tags). This study demonstrated that there are strong dependencies between sentiment values and visual features (i.e., SIFT based bag-of-visual words, and local/global RGB histograms).

In [5] the authors built a large scale visual sentiment ontology of semantic Adjective-Noun Pairs (ANPs) based on psychological theories and web mining (SentiBank). After building the ontology, the authors trained a set of visual concept detectors providing a mid-level representation of sentiment for a given image. There are also approaches that try to predict sentiment directly from pixels by exploiting Convolutional Neural Networks (CNNs) trained on large scale datasets [43, 46] or by properly fine-tuning pre-trained models [6, 7].

A model that combines textual and visual information is presented in [42]. The subjective textual data such as comments and captions on the images are considered as contextual information. In [2] the authors presented different learning architectures for sentiment analysis of social posts containing short text messages and an image (i.e., Tweets). They exploited a representation learning architecture that combines the input text with the polarity ground truth. This model is further extended with a Denoising Autoencoder when the visual information is present. The approach proposed in [25] combines visual features with text-based features extracted from the text subjectively associated to images (i.e., descriptions and tags). Specifically, the authors exploited a feature obtained by using the traditional Bag of Words approach on the subjective text, and a sentiment feature obtained by selecting the words of the subjective text whose sentiment scores (positive or negative) reported in SentiWordNet [11] are larger than a threshold, and applying the Bag of Words on this

revised vocabulary. The considered features are exploited to define an embedding space in which the correlation among the projected features is maximized. Then a sentiment classifier is trained on the features projected in the embedding space. This approach outperformed other state-of-the-art methods [5, 29, 38, 42].

The authors of [52] proposed a joint visual-textual sentiment analysis system trying to exploit more than one modality. In particular, the authors considered a cross-modality attention mechanism and semantic embedding learning based on Bidirectional Recurrent Neural Networks (BRNN) with the aim to design a model able to focus on the visual and textual features that mostly contribute to the sentiment classification. Huang et al. [17] propose an approach that defines three attention models aimed to learn effective sentiment classifiers for visual and textual inputs and the correlation between the two modalities. Then, a late fusion approach is used to combine the three attention models. Is important to notice that the text sources associated to images exploited in the aforementioned works can be very noisy due the subjectivity of such text. Different users can describe and tag the same image in different ways, including also text which is not related to the content. Considering that several visual sentiment analysis methods rely on the text provided by users [2, 17, 25, 42, 52], the proposed paper presents a study on the effect of the bias contained in this text toward the task of visual sentiment prediction. In particular, we investigated the use of an alternative objective text source. In previous approaches, the authors face different issues related to the subjective text associated to images. For instance, the framework presented in [42] implements an unsupervised approach aimed to address the lack of proper annotations/labels in the majority of social media images. In [14], the authors tried to learn an efficient image-sentence embedding by combining a large amount of weakly annotated images (where the text is obtained by considering title, descriptions and tags) with a smaller amount of fully annotated ones. In [41] the authors exploit large noisily annotated image collections to improve image classification.

To the best of our knowledge, this is the first work that propose the exploitation of objective text automatically extracted from images to deal with the issues related to the subjectivity nature of the text provided by users for Visual Sentiment Analysis purposes.

3 Proposed approach

In this Section we highlight the main differences between subjective and objective text, present the features extraction process and detail how to build the embedding space in order to exploit jointly different kind of features (views).

3.1 Subjective vs objective text

Analysing social pictures for Sentiment Analysis brings several advantages. Indeed, pictures published through social platforms are usually accompanied by additional information that can be considered. Several meta-data are available, depending on the specific platform, but in general all the pictures published through a social platform have at least a title, a description and a number of "significant" tags. Most of the existing works in the field exploit social subjective textual information associated to images either to define the ground truth [5] (i.e., by performing textual Sentiment Analysis on the text) or as an additional data modality (i.e., views) [2, 25]. In the latter case, both the visual and the textual information are used as input to establish the sentiment polarity of a post.

Although the text associated to social images is widely exploited in the state-of-the-art to address different tasks and to improve the semantics inferred from images, it can be a very noisy source because it is provided by the users; the reliability of such input is often based on the capability and the intent of the users to provide textual data that are coherent with respect to the visual content of the image. There is no guarantee that the subjective text accompanying an image is useful for the sentiment analysis task. It is usually related to a specific purpose or intention of the user that published the picture on the platform. Often, the subjective user description and tags are related to the semantic of the images or to the context of acquisition rather than sentiment. In addition, the tags associated to social images are often selected by users with the purpose to maximize the retrieval and/or the visibility of such images by the platform search engine. In Flickr, for instance, a good selection of tags helps to augment the number of views of an image, hence its popularity in the social platform. These information are hence not always useful for sentiment analysis.

As discussed in [13], the semantic of an image can be defined by a single object category, while the user-provided tags may include a number of additional terms correlated with the object coming from a larger vocabulary. Alternatively, the semantic might be given by multiple keywords corresponding to objects, scene types, or attributes. In the context of image retrieval, the authors of [13] exploited three views to build the embedding space with a Canonical Correlation Analysis approach (CCA). The first and the second views were related to visual and textual features respectively, whereas the third view was obtained considering the ground truth annotations (i.e., category) and the search keywords used to download the images. When these information were missing, the authors obtained the third view by clustering the tags, aiming to reduce the overall noise.

To better explain the problem, is useful to reason on a real case. Figure 1 shows an example image taken from the Flickr dataset used in [25]. The textual information below the image is the subjective text provided by the Flickr's user. Namely the photo title, the description and the tags are usually the text that can be exploited to make inferences on the image. As shown by this example, the text can be very noisy with respect to any task aimed to understand the sentiment that can be evoked by the picture. Indeed the title is used to describe the tension between the depicted dogs, whereas the photo description is used to ask a question to the community. Furthermore, most of the provided tags include misleading text such as geographical information (i.e., Washington State, Seattle), information related to the camera (i.e., Nikon, D200), objects that are not present in the picture (i.e., boy, red ball, stick) or personal considerations of the user (i.e., my new word view). Moreover, in the subjective text there are many redundant terms (e.g., dog). Another drawback of the text associated to social images is that two users can provide rather different information about the same picture, either in quality and in quantity. Finally, there is not guarantee that such text is present; this is an intrinsic limit of all Visual Sentiment Analysis approaches exploiting subjective text.

Starting from the aforementioned observations about the subjective text associated to social images, in this work we propose to exploit an objective aspect of the textual source that comes directly from the understanding of the visual content of the images. This text is achieved by employing four deep learning models trained to accomplish different visual inference tasks on the input image. At the top right part of Fig. 1 the objective text automatically extracted with different scene understanding methods is shown. In this case, the inferred text is very descriptive and each model provides distinctive information related to objects, scene, context, etc. The objective text extracted by the three different scene understanding methods has a pre-defined structure, therefore all the images have the same

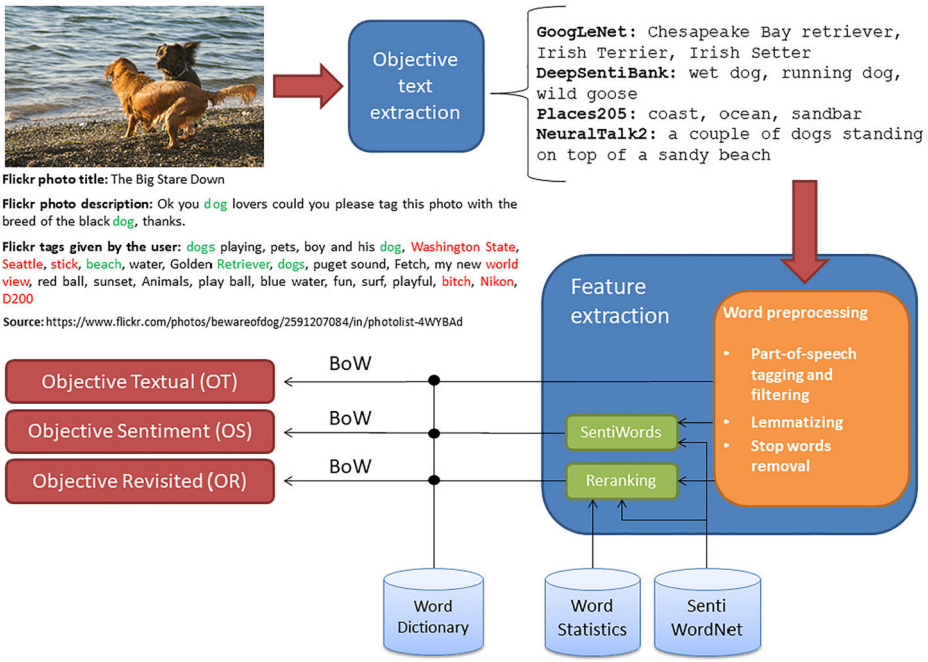


Fig. 1 Given an image, the proposed pipeline extracts Objective Text by exploiting four different deep learning architectures. The considered architectures are used to extract text related to objects, scene and image description. The obtained Objective Text is processed to produce three different features: the Objective Textual feature (OT) which is the BoW representation of the extracted text based on the whole dictionary, the Objective Sentiment feature (OS) which is the BoW representation obtained considering only the words with strong sentiment scores according to SentiWordNet, and the Objective Revisited feature (OR) which is the weighted BoW representation of the extracted text, in which the weight of each word is given by its statistics and sentiment scores according to the SentiWordNet lexicon. The figure shows also the subjective text associated to the image by the user (i.e., title, description and tags) at top left. The subjective text presents very noisy words which are highlighted in red. The words that appears either in the subjective and objective texts are highlighted in green

quantity of textual objective information. For each considered scene understanding method (i.e., GoogLeNet [39], DeepSentiBank [8] and Places205 [51]) the classification results are ranked by the output probability of the classifier and only the first three labels related to the classification results are considered in our framework. Augmenting the number of the classification results leads to the inclusion of wrong categories. Therefore, we considered the minimum number of labels that guarantee (in a probabilistic sense) a total classification probability close to 1 (i.e., the minimum number of outputs which probabilities sum distribution has a tendency near the value 1). To this aim, we analysed the distribution over the output classification probabilities (see Fig. 2) to understand the number of labels to be considered to describe the images. In our experiments, we observed that considering only the first three labels is a reliable approach to achieve a total classification probability very close to 1, avoiding to include noisy labels with respect to the visual content.

Finally, we used also a method able to produce an image caption (NeuralTalk2 [24]). That provides one more objective description (i.e., one more view) which we consider as objective text feature for sentiment purposes.

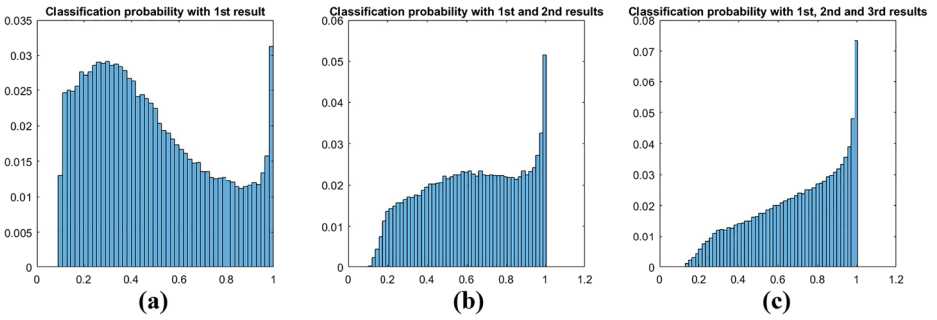


Fig. 2 In order to choose the number of classification labels to be taken into account for the objective text, we performed a statistical analysis of the output probabilities. The histograms show the probability distributions computed over a set of classification probability outputs of the exploited Deep Learning Architectures. The first histogram (**a**) is the distribution of the output probabilities obtained by considering only the first label in the classification ranking (i.e., corresponding to the highest classification output probability), whereas the second and the third histograms (**b**) and (**c**) show the distributions of the cumulative probability obtained by summing the first two and the first three probabilities of the obtained labels respectively. The distributions (**a**) and (**b**) have a strong tendency around the values 0.3 and 0.6. The histogram (**c**) instead shows a monotonic increasing shaped distribution with only one peak close to the value 1; indicating that the correct prediction is mostly within the first three labels

3.2 Features extraction

The proposed approach exploits one hand-crafted feature and three deep visual representations as visual views, and three text features to represent the objective text extracted from the images, namely Objective Textual (OT) and Objective Sentiment features (OS) and Objective Revisited (OR). As mentioned above, we use scene understanding approaches to extract objective text for the images.

3.2.1 Classic visual view

As in [25] we consider five image descriptors used in various Computer Vision tasks, such as object and scene classification. In particular, the extracted visual features are a 3×256 RGB histogram, a 512 dimensional GIST descriptor, a Bag of Words image descriptor using a dictionary with 1000 words with a 2-layer spatial pyramid and max-pooling, the 2000 dimensional attribute features presented in [47] and the SentiBank 1200 mid-level visual representation presented in [5]. Each image descriptor has been mapped by using the random Fourier feature mapping [35] or the Bhattacharyya kernel mapping [33]. Then, all the obtained representations have been reduced to 500 dimensions using Principal Component Analysis (PCA). The final visual feature vector associated to each image has a length of 2500 and is obtained as concatenation of all the PCA projected visual features. In our experiments we used the above pipeline to represent visual content of images in order to perform a fair comparison with respect to [25].

3.2.2 Deep visual view

In the last few years Convolutional Neural Networks (CNNs) have been showing outstanding performances in many Computer Vision challenges. Furthermore, CNNs have proved

to be very effective for transfer learning problems. In order to improve the contribution given by the visual view in the computed embedding space, in our experiments we exploited three state-of-the-art deep architectures to extract their inner visual representations (i.e., GoogLeNet [39], DeepSentiBank [8] and Places205 [51]). We first performed a set of baseline experiments based only on the deep visual representation. Then, we evaluated the contribute of the deep visual view in the embedding space, combined with the other extracted views.

3.2.3 Text views

Five text-based features are used in our experiments. Two of them are the same textual (T) and sentiment (S) views used in [25]. These features reflect the subjective text information provided by the users. Moreover, we built three textual features based on the Objective Text obtained through deep learning architectures. The overall pipeline for Objective Textual based features extraction is sketched in Fig. 1. Each exploited deep learning architecture provides a description, in some sense objective, of the input image from a different point of view, as each architecture has been trained for a different task (e.g., object recognition, place recognition, etc.). For instance, the deep architecture specialized for object classification (i.e., GoogLeNet [39]) finds the principal objects within the picture (e.g., dog), providing information about dog breeds in Fig. 1. The Adjective-Noun Pair classifier (i.e., DeepSentiBank [8]) agree with the previous result that the main object is a dog and provides other information in form of Adjectives-Noun Pairs (i.e., "wet dog" and "running dog"). The network devoted to place classification (i.e., Places205 [51]) gives further information about the location and the depicted environment (i.e., "coast", "ocean" and "sandbar"). Furthermore, the caption generated by NeuralTalk2 [24] provides a confirmation of all the previous inferences putting them in context through a description. The use of different architectures allows to obtain a wide objective description of the image content which consider different semantic aspect of the visual content. Although the exploited deep learning architectures are different, they all describe the same image, and it implies the generation of some redundant terms. This has not been considered as a drawback, indeed the presence of more occurrences of similar or related terms (e.g., dog, dogs, retriever, terrier, setter, etc.) enhance the weight of these correct terms in the representation extracted by our framework. On the other hand, this redundancy reduces the effect of noisy results such as the third result extracted with DeepSentiBank in Fig. 1 (i.e., "wild goose"). For these reasons, in the Bag of Words text representation exploited in the proposed paper, we considered the number of occurrences of each word of the vocabulary in the text associated to the image, instead of considering a binary vector representation which encodes the presence or the absence of each word [13, 15, 25].

To further compare the considered Objective Textual representation with other state of the art solutions, we implemented the feature extraction process described in [18]. According to this approach, a given text is represented as a feature vector which elements are obtained by multiplying the sentiment scores of the contained words by their frequencies. The sentiment scores are taken form SentiWordNet [11], and a re-ranking of such scores is performed for the words whose neutral score is higher than either the negative and the positive ones. In our experiments, we implemented both the re-ranking procedure and the feature extraction process of [18] for comparison purposes.

In this paper all the text-based features are obtained through a Bag of Words (BoW) representation of the objective text extracted from the input picture. These representations share

the same pre-processing stage of the text extracted with the deep learning architectures. This includes the procedures commonly applied in text mining:

- **Part of speech tagging and filtering:** this step choose a proper part of speech tag of each word, to solve its ambiguity. This step is needed since in SentiWordNet a word with a different part of speech tag might have a different sentiment value and, hence, dominant polarity. Considering our input source, we already know that the two words resulting from DeepSentiBank corresponds to an adjective-noun pair, and the most of the Places205 and GoogLeNet outputs are nouns. Therefore, this preprocessing mainly contributes on the text obtained with NeuralTalk2;
- **Lemmatizing:** since only base form of words are stored in SentiWordNet, we performed a lemmatizing step;
- **Stop words removal:** this step removes words that contain no semantic concepts, such as articles and prepositions.

The above pre-processing steps allow to obtain co-occurrences of the words describing the image from different semantic aspects of the visual content. Indeed, the proposed approach benefits from the inferences coming from architectures trained for different tasks: object classification, places classification, Adjective-Noun Pair classification and image description. Starting from the pre-processed Objective Text, we propose to extract the following text-based features:

- **Objective Text (OT):** we obtained this feature by computing a classic Bag of Words representation followed by a SVD dimensionality reduction. The final feature has dimension 1500.
- **Objective Sentiment (OS):** we computed the Bag of Words representation by using a reduced dictionary of sentiment related words (called sentiment vocabulary), followed by a SVD feature dimensionality reduction to obtain 20 dimensional vectors. We considered only the words which either positive or negative sentiment score in SentiWordNet, is higher than 0.15.
- **Objective Revisited (OR):** the paper described in [18] proposed an interesting text representation for the task of sentiment analysis. Furthermore, it highlights an issue related to the use of SentiWordNet scores for sentiment analysis. Indeed, most of the existing sentiment feature extraction methods (including [25]) ignore words which neutral sentiment is higher than either positive and negative ones, albeit they comprise the 93.75% of SentiWordNet entries. The authors of [18] proposed a revisiting procedure of the sentiment scores associated to the neutral words that modules the sentiment scores according to the probability of a word to appear in a positive or a negative sentence. Then, the representation of a given text is a weighted BoW vector which elements are obtained by weighting the word counts with the predominant sentiment score (positive, negative or zero if the neutral score remains the higher even after the scores revisiting). We use this process on the proposed Objective Text. The OR feature we compute is hence a vector W in which each W_i element is defined as follows:

$$W_i = \begin{cases} TF_i \times posW_i, & where\ W_i \in [pos\ words] \\ TF_i \times negW_i, & where\ W_i \in [neg\ words] \\ 0 & otherwise \end{cases} \quad (1)$$

where $posW_i$ and $negW_i$ denote the positive and negative sentiment scores of the *i-th* word, and TF_i is the number of occurrences of the *i-th* word in the Objective Text extracted from the considered image.

All the process described above for word dictionaries definitions, SVD computation and OT, OS and OR parameter settings have been done considering only the Objective Text associated to the training set images of the dataset used for our experiments. The methods are then evaluated on a different test set.

3.3 Embedding different views

Recently, several papers for jointly modelling images and associated text with Canonical Correlation Analysis (CCA) have been proposed [12, 13, 19, 20, 25, 36]. CCA is a technique that maps two or more views into a common embedding space. The CCA is used to find the projections of multivariate input data such that the correlation between the projected features is maximized. This space is cross-modal, therefore the embedded vectors representing the projections of the original views are treated as the same type of data. Thus, in the CCA embedding space, projections of different views are directly comparable by a similarity function defined over the elements of the embedding space [13].

Let ϕ^i be the data representation matrix in the *i-th* view. The n_v projection matrices W^i are learned solving the following minimization problem:

$$\begin{aligned} \min_{\{W^i\}_i^{n_v}} &= \sum_{i,j=1}^{n_v} Trace\left(W^i \Sigma_{ij} W^j\right) \\ &= \sum_{i,j=1}^{n_v} \left\| \phi^i W^i - \phi^j W^j \right\|_F^2 \\ s.t. \left[W^i\right]^T \Sigma_{ii} W^i &= I \quad \left[w_k^i\right]^T \Sigma_{ij} w_l^j = 0 \\ i \neq j, k \neq l &\quad i, j = 1, \ldots, n_v \quad k, l = 1, \ldots, n \end{aligned} \tag{2}$$

where W^i is the projection matrix which maps the *i-th* view matrix $\phi^i \in \Re^{n \times m_i}$ into the embedding space, w_k^i is the *k-th* column of W^i and Σ_{ij} is the covariance matrix between ϕ^i and ϕ^j. The dimensionality of the embedding space m_e is the sum of the input view dimensions $m_e = \sum_i^{n_v} m_i$. Therefore $W^i \in \Re^{m_i \times m_e}$ transforms the m_i dimensional vectors of the *i-th* view into the embedding space with dimension m_e. As demonstrated in [16], this optimization problem can be formulated as a standard eigenproblem. In the proposed work we exploited the multi-view CCA implementation provided by [12]. The same code has been used by the state of the art Visual Sentiment Analysis method proposed in [25] which we used as baseline to compare our method.

In Section 4.2, we describe how to use the embedding space learned from multiple views to obtain the features used in the proposed approach.

4 Experimental settings and results

4.1 Dataset

In [25] the authors performed the experiments with two different datasets crawled from Instagram and Flickr. For each image in the dataset, the image description and tags have been taken into account to obtain the text on which build the text based features. The images are not available for download, but the authors published the list of images' ids to allow performing comparison with other approaches. Due to the recently changes in Instagram

policies, we were unable to download images from this platform. Therefore we used only the dataset obtained downloading Flickr images. Some of the pictures were missing at the moment of crawling (e.g., removed by the users). Only 69893 Flickr images were available at the time of our analysis. Following the experimental protocol, we considered the images with positive or negative ground truth, discarding the images labelled as neutral. The final dataset used in the experiments has a total of 47235 images. Although the dataset used in our experiments is a subset of the Flickr images used in [25] due to aforementioned reasons, the number of either positive and negative images is comparable with the number of positive and negative images of the original dataset (see Table 1).

To evaluate the performances of the different compared sentiment classification approaches, we considered the original sentiment labels which have been obtained via crowdsourcing [25]. During the labelling process, each image has been presented to three people who were asked to provide a five-point scale sentiment score. The final ground truth has been defined by considering the majority votes of polarity for each image. The images labelled as neutral, as well as the images resulting in disagreement among people, have been discarded.

4.2 Embedded vectors

In Section 3.3 we described the CCA technique, and defined how to obtain the projection matrices W_i, related to each view i, by solving an optimization problem.

We exploited a weighted embedding transformation which emphasize the most significant projection dimensions [12]. The final representation of the data from the *i-th* view into the weighted embedding space is defined as:

$$\Psi^i = \phi^i W^i \left[D^i \right]^{\lambda} = \phi^i W^i \tilde{D}^i \tag{3}$$

where D^i is a diagonal matrix which diagonal elements are the eigenvalues in the embedding space, λ is a power weighting parameter, which is set to 4 as suggested in [12].

In our experiments we further considered a reduced projection obtained by taking only the first components of W^i encoding the 99% of the original information. The number of components to keep is obtained by considering the minimum number of eigenvalues (i.e., the diagonal elements of D) which normalized sum is greater or equal than 0.99. We call these representations *Truncated Features* in our experiments.

4.3 Performance evaluation

The dataset has been randomly separated into a training set and test set, considering a proportion of 1:9 between the number of test and training images, and including a balanced number of positive and negative examples. As a performance evaluation metric, we computed the average and standard deviation of test classification accuracy over 10 runs,

Table 1 Number of positive and negative images in our dataset and in the original dataset used in [25]

	Positive	Negative
Dataset in [25]	48139	12606
Our Dataset	37622	9613

Table 2 Performance Evaluation of the proposed method with respect to the baseline method presented in [25]

	Experiment ID	Embedded views	Full feature	Truncated features (99%)
Subjective F	K1	V+T+S	66.56 ±0.43 %	66.11 ±0.45 %
eatures Proposed	K2	V+T	71.67 ±0.36 %	71.55 ±0.57 %
in [25]	K3	V+S	62.19 ±0.63 %	62.89 ±0.45
Considering Subjective	O1	V+T+OS	68.88 ±0.49 %	69.23 ±0.38 %
and/or Objective	O2	V+OT+S	66.97 ±0.57 %	66.34 ±0.68 %
Features	O3	V+OT	<u>73.48</u> ±0.54%	<u>72.54</u> ±0.65 %
	O4	V+OS	66.58 ±0.70 %	66.41 ±0.53 %
	O5	V+OT+OS	69.83 ±0.58 %	69.62 ±0.53 %
	O6	V+T+S OT+OS	68.04 ±0.55 %	67.39 ±0.19 %
	O7	V+T+OR	66.04 ±0.54 %	66.74 ±0.45 %
	O8	V+OT+OR	68.29 ±0.54 %	67.84 ±0.68 %
	O9	V+OR	64.60 ±0.70 %	63.08 ±0.82 %
	O10	V+T+OT	**73.96** ±0.39 %	**72.66** ±0.70 %

The best result is highlighted in bold, whereas the second best result is underlined. See text for details

repeating the data shuffling at each run.[2] A linear SVM has been used to establish the sentiment polarity over the different compared representations. For each experimental setting we used LibLinear[3] to train a linear SVM classifier. The parameter C of the linear SVM was determined by 10-fold cross validation.

Table 2 shows the obtained results. Each row describes a different experimental setting, corresponding to a specific combination of the input features described in Section 3.2 used to build the embedding space. The column "Full Feature" reports the results obtained by considering the full-size representation in the embedding space obtained by applying (3), whereas the results of the experiments performed with the truncated feature representations are reported in the last column (i.e., "Truncated Features"). In Table 2 all the tests with prefix "O" (Objective) are related to the exploitation of features extracted with the proposed method, whereas the features V, T and S refer to the features extracted with the method presented in [25] (Visual, Textual and Sentiment respectively). The third column lists the views used for the computation of the embedding space. For instance, V+T refers to the two-view embedding based on Visual and Textual features, V+OT+OS is related to the three-view embedding based on Visual, Objective Textual, and Objective Sentiment features, and so on.

As first, it is simple to note that all the tests where the Objective Text description is used achieve better results with respect to the experimental settings in which the corresponding Subjective Text features are exploited (see Table 2 and Fig. 3). Figure 3 compares the experimental settings which differ by the exploitation of one or more subjective or objective features. This allows the comparison between the exploitation of the proposed "Objective

[2]The code to repeat the performance evaluation is available at the URL: http://iplab.dmi.unict.it/sentimentembedding/

[3]http://www.csie.ntu.edu.tw/~cjlin/liblinear/.

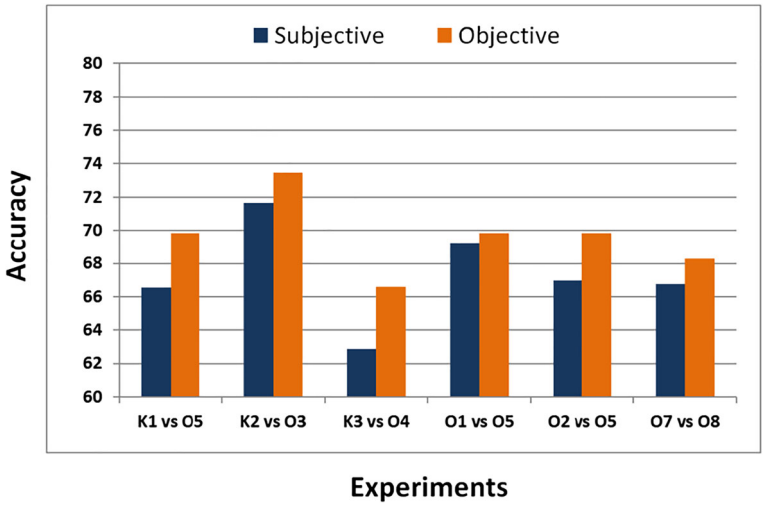


Fig. 3 Explicit comparison between the experimental settings that differ by the use of one or more particular objective/subjective feature (e.g., O5 differs from K1 by the exploitation of OT and OS instead of T and S, respectively). In all the experiments, the objective features (orange bins) perform better than the corresponding subjective ones (blue bins). For each experimental setting, the accuracy value reported in this histogram is the best achieved result between the experiment performed by using the full feature and the truncated one (see Table 2)

Text" with respect to the classic "Subjective Text". Among this subset of experimental settings, the good results are obtained by exploiting Visual feature and Objective Text (O3 in Table 2). In particular, using the feature OT instead of T provides a mean accuracy improvement of 1.81% (compare O3 and K2 in Table 2). Adding the view T to the experimental setting O3 yields an increment of 0.48% (O10 in Table 2). Note that, adding the proposed OT feature to the experimental setting K2 provides an improvement of 2.29% (compare K2 with respect to O10). These observations highlight the effectiveness of the features extracted from Objective Text with respect to the features extracted from the subjective one. Finally, when the proposed truncated features are employed the classification accuracy has a mean decrease of 0.31%, which is lower than the standard deviation of the accuracy values computed over 10 runs in all the performed experiments. This means that, even if the exploitation of the truncated features causes a decrease of the mean accuracy, the range of the values obtained in the two cases are comparable. To better assess this observation, Fig. 4 shows the box-and-whisker plots (henceforth, referred to simply as boxplots) obtained from the distributions of the accuracy values reported in Table 2.

The fact that the sentiment features (i.e., S and OS) do not achieve good results is probably due to the fact that the number of sentiment words considered to build the Bag of Words representations according to the method proposed in [25] is very limited. Furthermore, the sentiment score of most of the SentiWordNet words is often neutral, indeed the 97.75% of SentiWordNet words are words which neutral score is higher than either the negative and the positive ones. Therefore, most of the words of the sentiment vocabulary are still neutral. In particular, we observed that about 61% of the words in the sentiment vocabulary used in this paper are neutral for SentiWordNet, the 24% are negative and 15% are positive. As a result, the feature extraction process for sentiment features produces very sparse and rather uninformative Bag of Words sentiment representations. This observation is confirmed by

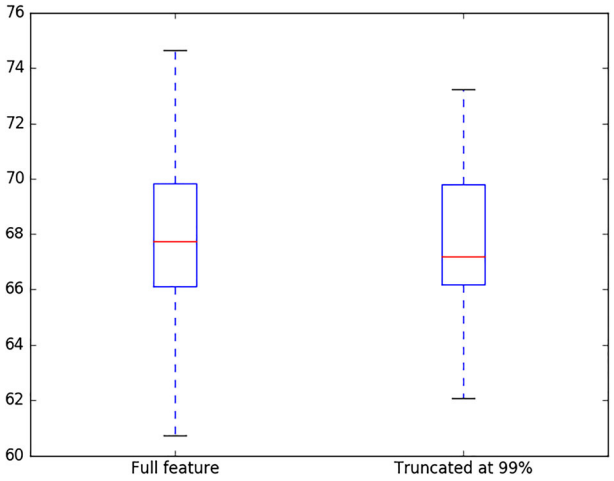


Fig. 4 Comparison between the distributions of the achieved accuracy values between the experimental settings which exploit the full size features and the ones which exploit the truncated features. The two distributions have been obtained by considering the accuracy and standard deviation values reported in Table 2

the fact that the best results are obtained considering Visual, Objective Textual and Subjective Textual features (O10 in Table 2). To summarize, from this first set of experiments we observed that:

- the Objective text features (i.e., OT and OS) performs better than the Subjective text features (i.e., T and S);
- the features obtained by exploiting the full vocabulary (i.e., T and OT) performs better than the ones obtained by the usage of sentiment biased textual features (i.e., S, OS and OR);
- the truncation of the projected features performed by exploiting the principal eigenvalues causes a slightly reduction of the mean accuracy. However, taking into account the variability of the achieved results, we observed that the statistical distribution of the accuracy values of the experiments performed with the truncated features is comparable with the one related to the experiments performed with the full features;
- the best result are obtained by the experimental setting O10 (i.e., V+T+OT, 73.96%). The second best result is obtained by the experimental setting O3 (i.e., V+OT, 73.48%). The comparison between several experimental settings demonstrate that the contribute given by the exploitation of the proposed Objective Textual feature (OT) is significantly higher than the contribute given by the Subjective Textual feature (T).

4.3.1 Improving the visual representation

The experiments of the previous section are needed to perform a fair comparison with the method presented in [25]. Indeed, the first part of the experimental evaluation performed in this paper is aimed to compare the exploitation of the Objective Text with the Subjective one, keeping the same Visual View (V) in all the embeddings. However, recent works in Computer Vision provide several stronger deep learning based visual representations. Some of them can be easily extracted from the deep architectures to be exploited in this paper jointly with the Objective Text.

Since, from a computational point of view, the effort to extract the Objective Text and the corresponding deep features is similar (i.e., a single feed forward step), it worth to consider such deep visual representations as alternative features for the addressed semantic classification task. Therefore, we extracted the deep features representations of images by using the considered CNNs (i.e., GoogLeNet [39], DeepSentiBank [8] and Places205 [51]). Instead of focusing on the final output (i.e., classification), we extracted the activations of the earlier layer and trained an SVM classifier, according to the above described evaluation protocol. Since the achieved representations are based on stronger visual features than the one which are included in the visual view (V), the results of this procedure provide a strong and challenging additional baseline for our evaluation experiments. Table 3 shows the classification results obtained by training an SVM for the task of sentiment polarity prediction when only the aforementioned deep visual features are employed. As we expected, the deep feature extracted from DeepSentiBank outperforms the others, as this CNN has been trained for the task of Adjective Noun Pair (ANP) prediction, which is strongly related to the task of sentiment polarity classification. The boxplot diagram shown in Fig. 5 is useful to compare the results reported in Table 3 achieved by the different deep visual representations on the different runs. As described in Section 3.2, the Visual View (V) exploited by the proposed approach includes the SentiBank 1200 mid-level visual representation. This feature can be considered an earlier version of the one provided by DeepSentiBank and it takes into account only 1200 ANPs. The DeepSentiBank CNN is trained to predict 4342 different ANPs.

Considering that the performances achieved by only using the deep feature representation extracted with DeepSentiBank are better than the two best results obtained by the proposed method (O3 and O10 in Table 2), we repeated the performance evaluation of the CCA embedding based representations considering the DeepSentiBank visual feature (DS) instead of the visual feature (V). The results are reported in Table 4. The exploitation of the deep visual representation (DS) produced a further improvement of the performances in all the experimental settings (compare Tables 3 and 4). In particular, the combination of the DeepSentiBank visual feature (DS) and the Objective Text (OT) feature provides a mean improvement of 2.82% in accuracy with respect to the best two results (DO3 versus O3 and DO10 versus O10). Also, the results obtained exploiting jointly DeepSentiBank features and the Objective Text (i.e., 76.78% of DO3 in Table 4) are better than the results obtained when only DeepSentiBank features are used (i.e., 75.92% in Table 3). Considering the results detailed in Table 4, we observe that in this case the truncation procedure of the projected features produces a more significant decrease of the accuracy score (1.02% in mean). An interpretation of such results is that in the case of the projected representations obtained by considering hand crafted visual features (V), there are some components that can be truncated without a significant decrease in performance. Whereas when the projected representations are obtained by considering the deep visual features (DS), almost all the representation components (i.e., including the ones not in the 99% most informative

Table 3 Results obtained by training an SVM on the deep features extracted from GoogLeNet [39], DeepSentiBank [8] and Places205 [51] (pool5/7x7_s1, fc7 and fc7 respectively)

Architecture	Feature dimension	Results
DeepSentiBank [8]	4096	**75.92** ±0.65
GoogLeNet [39]	1024	75.14 ±0.46
Places205 [51]	4096	73.83 ±0.65

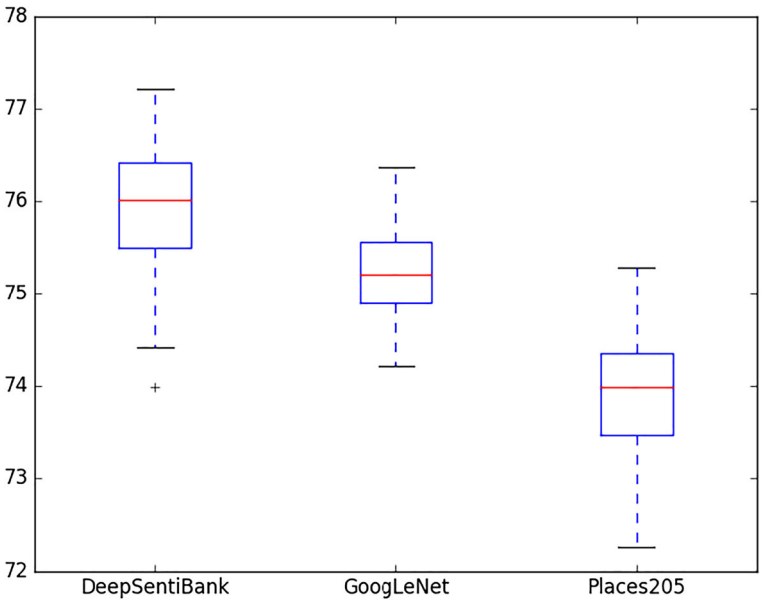


Fig. 5 Comparison between the distributions of the accuracy values reported in Table 3 achieved by exploiting only a deep visual image representation to train the SVM classifier

components) provide an non negligible contribute for the classification task. For a better and complete comparison of the achieved results, Fig. 6 shows the accuracy values obtained with all the different experimental settings detailed in Tables 2 and 4, considering either the full feature and the truncated feature experiments.

Table 4 Performance Evaluation considering Deep Visual Representations

	Experiment ID	Embedded views	Full feature	Truncated features (99%)
Deep Visual and	DK1	DS+T+S	69.19 ±0.52%	67.36 ±0.64 %
Subjective Features	DK2	DS+T	74.87 ±0.52 %	73.74 ±0.75 %
[25]	DK3	DS+S	64.70 ±0.68 %	64.29 ±0.79
Deep Visual and	DO1	DS+T+OS	71.30 ±0.25 %	70.34 ±0.34 %
Objective Features	DO2	DS+OT+S	69.42 ±0.44 %	68.29 ±0.68 %
	DO3	DS+OT	**76.78** ±0.42 %	<u>74.46</u> ±0.67 %
	DO4	DS+OS	69.01 ±0.88 %	68.90 ±0.49 %
	DO5	DS+OT+OS	72.00 ±0.37 %	71.16 ±0.86 %
	DO6	DS+T+S OT+OS	69.77 ±0.31 %	68.58 ±0.34 %
	DO7	DS+T+OR	69.59 ±0.55 %	68.36 ±0.55 %
	DO8	DS+OT+OR	70.61 ±0.65 %	69.14 ±0.52 %
	DO9	DS+OR	66.43 ±0.61 %	66.58 ±0.73 %
	DO10	DS+T+OT	<u>76.31</u> ±0.55 %	**74.52** ±0.45 %

The best result is highlighted in bold, whereas the second best result is underlined. See text for details

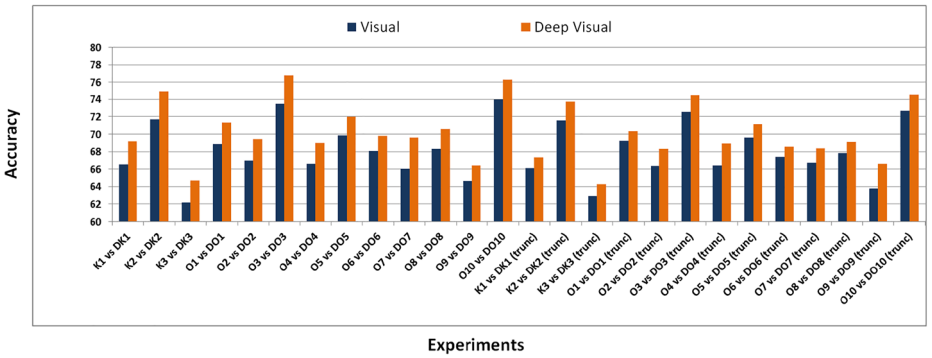


Fig. 6 Comparison of all the considered experimental settings, directly comparing the settings in which the same textual features are combined with the visual or the deep visual one. In all the experiments, the settings that exploit the deep visual features in the embeddings (orange bins) perform better than the corresponding settings that exploit the visual features (blue bins)

5 Conclusion and future works

This paper addresses the challenge of image sentiment polarity classification. To this aim, we presented an approach which exploits the correlations among visual and textual features associated to images. The considered features are exploited to build an embedding space in which correlations among the different features are maximized with CCA. Then, an SVM classifier is trained on the features of the built embedding space for prediction purposes. In the work presented in this paper the contribute of either visual and textual features in the CCA embedding is deeply investigated and assessed. The first part of the work presents a study in which Objective Text extracted considering the visual content of images is compared with respect to the Subjective Text provided by users. The best results have been obtained by exploiting the proposed Objective Text based features. This study demonstrates that the exploitation of Objective Text associated to images provides better results than the use of the Subjective Text provided by the user. Furthermore, it brings several advantages. Indeed it has a pre-defined structure, providing the same quantity of textual objective information for all the images. Each exploited deep learning architecture used to extract the objective text contributes to the description of the image from a different perspective.

On the other hand, we identified several drawbacks brought by the Subjective Text due its intrinsic nature. Indeed, the subjective text associated to images by users presents very noisy terms. It does not respect a pre-defined scheme or length constraints, therefore the text sources associated to different images may have very different structures. Moreover, there is no guarantee of the presence of such text for all the considered images. The Objective Text exploited by our approach doesn't present the aforementioned issues, and it is automatically extracted from the image. Two similar images are likely to have very similar Objective Text, whereas we cannot say anything about their subjective text provided by the users. These observations and our experimental results support the use of Objective Text automatically extracted from images for the task of Visual Sentiment Analysis in lieu of the Subjective Text provided by users, and suggest the investigation of the exploitation of such text also for other task related to the association between text and images. Finally, the obtained results show that all the text features based on the SentiWordNet scores do not achieve good results. Mainly due to the lack of strong positive or negative terms in the analysed text. An in-depth investigation on this aspect is needed: future works will be devoted to the exploitation of

more sentiment oriented information to build features that reflects the emotions evoked by images. Social platforms provides interesting data to infer people reactions toward images (i.e., users social engagement). They include comments, likes and shares of users that see the image in the platform which can be used to extend the Visual Sentiment Analysis methods to improve the overall accuracy for sentiment polarity estimation.

With the aim to further boost the performances of the proposed system, we considered different visual features based on deep architectures. This evaluation demonstrated that deep based visual representations perform better than the hand crafted visual features proposed in previous works for the task of sentiment polarity classification. Furthermore, these experiments confirmed that the contribution of Objective Text based features is higher than the one provided from Subjective Text ones. Our performance evaluation considers 52 different combinations of features to build CCA embedding spaces, obtained by considering different textual and visual features, and different strong baselines based on the exploitation of deep based visual features. Experiments confirmed that the textual features extracted from the proposed Objective Text outperform the ones based on the Subjective Text provided by users by considering different combinations of features.

Considering the high contribute given by deep visual representations, in future works we will further investigate the task of image sentiment prediction by taking into account also models and techniques that try to predict sentiment directly from pixels [43, 46]. The work in [7] and further extended in [6], presents a large study on the suitability of pre-trained CNN for the task of visual sentiment prediction. Several aspects and their effect on the classification accuracy have been investigated, including data augmentation, ambiguous annotations, weight initialization, etc. In future works, we will also consider alternative methods to combine multiple classifiers, such as the ensemble of deep models presented in [1].

Acknowledgments This work has been partially supported by Telecom Italia TIM - Joint Open Lab.

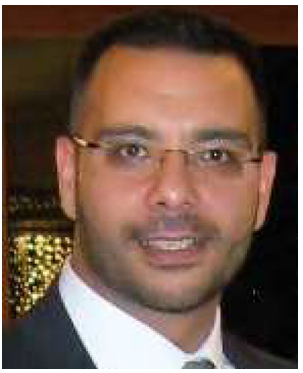
References

1. Ahmad K, Mekhalfi ML, Conci N, Melgani F, Natale FD (2018) Ensemble of deep models for event recognition. ACM Transactions on Multimedia Computing Communications, and Applications (TOMM) 14(2):51
2. Baecchi C, Uricchio T, Bertini M, Del Bimbo A (2016) A multimodal feature learning approach for sentiment analysis of social network multimedia. Multimed Tools Appl 75(5):2507–2525
3. Battiato S, Farinella GM, Milotta FL, Ortis A, Addesso L, Casella A, D'Amico V, Torrisi G (2016) The social picture. In: Proceedings of the 2016 ACM on international conference on multimedia retrieval, pp 397–400. ACM
4. Battiato S, Moltisanti M, Ravì F, Bruna AR, Naccari F (2013) Aesthetic scoring of digital portraits for consumer applications. In: IS&T/SPIE electronic imaging, pp 866008–866008. International Society for Optics and Photonics
5. Borth D, Ji R, Chen T, Breuel T, Chang SF (2013) Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of the 21st ACM international conference on multimedia, pp 223–232. ACM
6. Campos V, Jou B, i Nieto XG (2017) From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction. Image and Vision Computing 65:15–22. https://doi.org/10.1016/j.imavis.2017.01.011. http://www.sciencedirect.com/science/article/pii/S0262885617300355. Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing
7. Campos V, Salvador A, Giró-i Nieto X, Jou B (2015) Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction. In: Proceedings of the 1st international

workshop on affect & sentiment in multimedia, ASM '15. ACM, New York, pp 57–62. https://doi.org/10.1145/2813524.2813530
8. Chen T, Borth D, Darrell T, Chang SF (2014) Deepsentibank:, Visual sentiment concept classification with deep convolutional neural networks. arXiv:1410.8586
9. Cui P, Liu S, Zhu W (2017) General knowledge embedded image representation learning. IEEE Transactions on Multimedia
10. Datta R, Joshi D, Li J, Wang JZ (2006) Studying aesthetics in photographic images using a computational approach. In: European conference on computer vision, pp 288–301. Springer
11. Esuli A, Sebastiani F (2006) Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of The European language resources association, vol 6, pp 417–422. Citeseer
12. Fu Y, Hospedales TM, Xiang T, Fu Z, Gong S (2014) Transductive multi-view embedding for zero-shot recognition and annotation. In: Proceedings of the European conference on computer vision, pp 584–599. Springer
13. Gong Y, Ke Q, Isard M, Lazebnik S (2014) A multi-view embedding space for modeling internet images, tags, and their semantics. Int J Comput Vis 106(2):210–233
14. Gong Y, Wang L, Hodosh M, Hockenmaier J, Lazebnik S (2014) Improving image-sentence embeddings using large weakly annotated photo collections. In: Proceedings of the European conference on computer vision, pp 529–545. Springer
15. Guillaumin M, Verbeek J, Schmid C (2010) Multimodal semi-supervised learning for image classification. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 902–909. IEEE
16. Hardoon DR, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: An overview with application to learning methods. Neural Comput 16(12):2639–2664
17. Huang F, Zhang X, Zhao Z, Xu J, Li Z (2019) Image–text sentiment analysis via deep multimodal attentive fusion. Knowl-Based Syst 167:26–37
18. Hung C, Lin HK (2013) Using objective words in sentiwordnet to improve sentiment classification for word of mouth. IEEE Intell Syst 28(2):47–54
19. Hwang SJ, Grauman K (2010) Accounting for the relative importance of objects in image retrieval. In: Proceedings of British machine vision conference, vol 1, 2
20. Hwang SJ, Grauman K (2012) Learning the relative importance of objects from tagged images for retrieval and cross-modal search. Int J Comput Vis 100(2):134–153
21. Itten J (1962) The art of color; the subjective experience and objective rationale of colour
22. Johnson J, Ballan L, Fei-Fei L (2015) Love thy neighbors: Image annotation by exploiting image metadata. In: Proceedings of the IEEE international conference on computer vision, pp 4624–4632
23. Jou B, Chen T, Pappas N, Redi M, Topkara M, Chang SF (2015) Visual affect around the world: A large-scale multilingual visual sentiment ontology. In: Proceedings of the 23rd ACM international conference on multimedia, pp 159–168. ACM
24. Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3128–3137
25. Katsurai M, Satoh S (2016) Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In: Inproceedings of the IEEE international conference on acoustics, speech and signal processing, pp 2837–2841. IEEE
26. Lei X, Qian X, Zhao G (2016) Rating prediction based on social sentiment from textual reviews. IEEE Trans Multimed 18(9):1910–1921
27. Li X, Uricchio T, Ballan L, Bertini M, Snoek CG, Bimbo AD (2016) Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. ACM Comput Surveys (CSUR) 49(1):14
28. Machajdik J, Hanbury A (2010) Affective image classification using features inspired by psychology and art theory. In: Proceedings of the 18th ACM international conference on multimedia, pp 83–92. ACM
29. Mike T, Kevan B, Georgios P, Di C, Arvid K (2010) Sentiment in short strength detection informal text. Journal of the Association for Information Science and Technology 61(12):2544–2558
30. Miller GA (1995) Wordnet: a lexical database for english. In: Communications of the ACM, vol 38, pp 39–41. ACM
31. Ortis A, Farinella GM, Torrisi G, Battiato S (2018) Visual sentiment analysis based on on objective text description of images. In: 2018 International conference on content-based multimedia indexing (CBMI), pp 1–6. IEEE
32. Pang L, Zhu S, Ngo CW (2015) Deep multimodal learning for affective analysis and retrieval. IEEE Trans Multimed 17(11):2008–2020
33. Perronnin F, Sénchez J, Xerox YL (2010) Large-scale image categorization with explicit data embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2297–2304

34. Qian S, Zhang T, Xu C, Shao J (2016) Multi-modal event topic model for social event analysis. IEEE Trans Multimed 18(2):233–246
35. Rahimi A, Recht B et al (2007) Random features for large-scale kernel machines. In: Inproceedings of the neural information processing systems, vol 3, pp 5
36. Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet GR, Levy R, Vasconcelos N (2010) A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM international conference on multimedia, pp 251–260. ACM
37. Rudinac S, Larson M, Hanjalic A (2013) Learning crowdsourced user preferences for visual summarization of image collections. IEEE Trans Multimed 15(6):1231–1243
38. Siersdorfer S, Minack E, Deng F, Hare J (2010) Analyzing and predicting sentiment of images on the social web. In: Proceedings of the 18th ACM international conference on multimedia, pp 715–718. ACM
39. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition
40. Valdez P, Mehrabian A (1994) Effects of color on emotions. In: Journal of experimental psychology: General, vol. 123, p. 394. American Psychological Association
41. Wang G, Hoiem D, Forsyth D (2009) Building text features for object image classification. In: Inproceedings of the IEEE conference on computer vision and pattern recognition, pp 1367–1374
42. Wang Y, Wang S, Tang J, Liu H, Li B (2015) Unsupervised sentiment analysis for social media images. In: Proceedings of the 24th international joint conference on artificial intelligence, Buenos Aires, Argentina, pp 2378–2379
43. Xu C, Cetintas S, Lee K, Li L (2014) Visual sentiment prediction with deep convolutional neural networks. arXiv:1411.5731
44. Yang X, Zhang T, Xu C (2015) Cross-domain feature learning in multimedia. IEEE Trans Multimed 17(1):64–78
45. You Q, Cao L, Cong Y, Zhang X, Luo J (2015) A multifaceted approach to social multimedia-based prediction of elections. IEEE Trans Multimed 17(12):2271–2280
46. You Q, Luo J, Jin H, Yang J (2015) Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: 29th AAAI conference on artificial intelligence
47. Yu FX, Cao L, Feris RS, Smith JR, Chang SF (2013) Designing category-level attributes for discriminative visual recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 771–778
48. Yuan J, Mcdonough S, You Q, Luo J (2013) Sentribute: image sentiment analysis from a mid-level perspective. In: Proceedings of the 2nd international workshop on issues of sentiment discovery and opinion mining. ACM
49. Yuan Z, Sang J, Xu C (2013) Tag-aware image classification via nested deep belief nets. In: 2013 IEEE international conference on multimedia and expo (ICME), pp 1–6. IEEE
50. Yuan Z, Sang J, Xu C, Liu Y (2014) A unified framework of latent feature learning in social media. IEEE Trans Multimed 16(6):1624–1635
51. Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A (2014) Learning deep features for scene recognition using places database. In: Advances in neural information processing systems, pp 487–495
52. Zhu X, Cao B, Xu S, Liu B, Cao J (2019) Joint visual-textual sentiment analysis based on cross-modality attention mechanism. In: International conference on multimedia modeling, pp 264–276. Springer

Alessandro Ortis is a Ph.D Student in Computer Science at the University of Catania. He has been working in the field of Computer Vision research since 2012, when he joined to the IPLab (Image Processing Laboratory). He obtained the Master Degree in Computer Science (summa cum laude) from the University of Catania in March 2015. Alessandro was awarded with the Archimede Prize for the excellence of academic career conferred by the University of Catania in 2015. Along the way he has done two research internships at STMicroelectronics in 2011/2012 and at Telecom Italia in 2015. His research interests lie in the fields of Computer Vision, Machine Learning and Multimedia. Alessandro is reviewer for several international conferences and journals. He is co-author of 10 papers published in international conferences, 1 journal and co-inventor of 1 International Patent.



Giovanni Maria Farinella is Assistant Professor at the Department of Mathematics and Computer Science, University of Catania, Italy. He received the (egregia cum laude) Master of Science degree in Computer Science from the University of Catania in April 2004. He was awarded the Ph.D. in Computer Science from the University of Catania in October 2008. From 2008 he serves as Professor of Computer Science for undergraduate courses at the University of Catania. He is also an Adjunct Professor at the School of the Art of Catania in the field of Computer Vision for Artists and Designers (Since 2004). His research interests lie in the field of Computer Vision, Pattern Recognition and Machine Learning. He is author of one book (monograph), editor of 5 international volumes, editor of 4 international journals, author or co-author of more than 100 papers in international book chapters, international journals and international conference proceedings, and of 18 papers in national book chapters, national journals and national conference proceedings. He is co-inventor of 4 patents involving industrial partners. Dr. Farinella serves as a reviewer and on the board programme committee for major international journals and international conferences (CVPR, ICCV, ECCV, BMVC). He has been Video Proceedings Chair for the International Conferences ECCV 2012 and ACM MM 2013, General Chair of the International Workshop on Assistive Computer Vision and Robotics (ACVR - held in conjunction with ECCV 2014, ICCV 2015 and ECCV 2016), and chair of the International Workshop on Multimedia Assisted Dietary Management (MADiMa) 2015/2017. He has been Speaker at international events, as well as invited lecturer at industrial institutions. Giovanni Maria Farinella founded (in 2006) and currently directs the International Computer Vision Summer School (ICVSS). He also founded (in 2014) and currently directs the Medical Imaging Summer School (MISS). Dr. Farinella is an IEEE Senior Member and a CVF/IAPR/GIRPR/AIxIA/BMVA member.

Giovanni Torrisi is a Telecom Italian researcher at the Joint Open Lab of Catania. His research interests include Internet of Things, Wearable Devices, Mobile Design & Development, Big Data and Data Visualization. He has a Degree in Computer Science (summa cum laude) received in 2012 from the University of Catania. Giovanni is co-author of 4 papers published in international conferences and 2 journals.

Sebastiano Battiato is Full Professor of Computer Science at University of Catania. He received his degree in computer science (summa cum laude) in 1995 from University of Catania and his Ph.D. in Computer Science and Applied Mathematics from University of Naples in 1999. From 1999 to 2003 he was the leader of the ?Imaging? team at STMicroelectronics in Catania. He joined the Department of Mathematics and Computer Science at the University of Catania in 2004 (respectively as assistant professor, associate professor in 2011 and full professor in 2016). He is currently Chairman of the undergraduate program in Computer Science, and Rector's delegate for Education (postgraduates and Phd). He is involved in research and directorship of the IPLab research lab (http://iplab.dmi.unict.it). He coordinates IPLab participation to large scale projects funded by national and international funding bodies, as well as by private companies. Prof. Battiato has participated as principal investigator in many international and national research projects. His research interests include image enhancement and processing, image coding, camera imaging technology and multimedia forensics. He has edited 6 books and coauthored about 200 papers in international journals, conference proceedings and book chapters. Guest editor of several special issues published on International Journals. He is also co-inventor of 22 international patents, reviewer for several international journals, and he has been regularly a member of numerous international conference committees. Chair of several international events (ICIAP 2017, VINEPA 2016, ACIVS 2015, VAAM2014-2015-2016, VISAPP2012-2015, IWCV2012, ECCV2012, ICIAP 2011, ACM MiFor 2010-2011, SPIE EI Digital Photography 2011-2012-2013, etc.). He is an associate editor of the SPIE Journal of Electronic Imaging and at the IET Image Processing Journal. He is the recipient of the 2011 Best Associate Editor Award of the IEEE Transactions on Circuits and Systems for Video Technology. He is director (and co-founder) of the International Computer Vision Summer School (ICVSS), Sicily, Italy. He is a senior member of the IEEE.