Visual Sentiment Analysis Based on Objective Text Description of Images

Alessandro Ortis
University of Catania
Viale A. Doria 6 - 95125
Catania, Italy
ortis@dmi.unict.it

Giovanni M. Farinella
University of Catania
Viale A. Doria 6 - 95125
Catania, Italy
gfarinella@dmi.unict.it

Giovanni Torrisi
JOL Catania - Telecom Italia
Viale A. Doria 6 - 95125
Catania, Italy
giovanni.torrisi@telecomitalia.it

Sebastiano Battiato
University of Catania
Viale A. Doria 6 - 95125
Catania, Italy
battiato@dmi.unict.it

***Abstract*—Visual Sentiment Analysis aims to estimate the polarity of the sentiment evoked by images in terms of positive or negative sentiment. To this aim, most of the state of the art works exploit the text associated to a social post provided by the user. However, such textual data is typically noisy due to the subjectivity of the user which usually includes text useful to maximize the diffusion of the social post. In this paper we extract and employ an *Objective Text* description of images automatically extracted from the visual content rather than the classic *Subjective Text* provided by the users. The proposed method defines a multimodal embedding space based on the contribute of both visual and textual features. The sentiment polarity is then inferred by a supervised Support Vector Machine trained on the representations of the obtained embedding space. Experiments performed on a representative dataset of 47235 labelled samples demonstrate that the exploitation of the proposed *Objective Text* helps to outperform state-of-the-art for sentiment polarity estimation.**

***Index Terms*—Visual Sentiment Analysis, Social Media Analysis, Objective Text Description, Multimodal Embedding**

I. Introduction

Social media users continuously post images together with their opinions and share their emotions. This trend has supported the growing of new application areas, such as semantic-based image selection from crowdsourced collections [1], [2], Social Event Analysis [3] and Sentiment Analysis on Visual Contents [4]. 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 [5], [6] or have employed text to define a sentiment ground truth [7], [8]. More recent approaches combine visual and text feaures by exploiting well-known semantic and sentiment lexicons [9], [10].

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 [11], Visual Sentiment Analysis [12], Image Retrieval [13]–[20], and Event Classification [21] 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 by using scene understanding methods [22]–[24]. Since the text describing the images is automatically extracted, in our approach, we denote it as "objective" emphasizing the fact that it is different to the "subjective" text written by the user for an image of a post.

In [12] two different datasets are considered by crawling public images from Instagram and Flickr respectively. Three types of features, called views, are combined to form an embedding space. 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., the sentiment polarity). The work in [12] achieved significant improvements with respect to other Visual Sentiment Analysis methods [7], [8], [25], [26].

In order to perform a fair comparison with respect to the state of the art, we considered the dataset used in [12] as well as the same evaluation protocol. Differently than [12], we defined three text-based views by exploiting the proposed objective text as input instead of the subjective text provided by users. 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 to define embedding spaces to be used for sentiment polarity estimation. We also employ a dimensionality reduction strategy for the learned embedded representations.

The feature evaluation performed in this paper focuses on the task of Visual Sentiment Analysis, however the observations and the achieved insights are useful also to other systems which exploit the text associated to social images.

II. Related Works

Several papers investigated the problem of joint modelling the representation of images and associated text or tags for different tasks, such as image retrieval [16], [18], [27], social images understanding [1], image annotation [28] and visual sentiment analysis [7], [8], [12], [29]. The authors of [7] studied the correlations between the sentiment evoked by images and their visual content with the aim to classify

978-1-5386-7021-7/18/$31.00 2018 IEEE

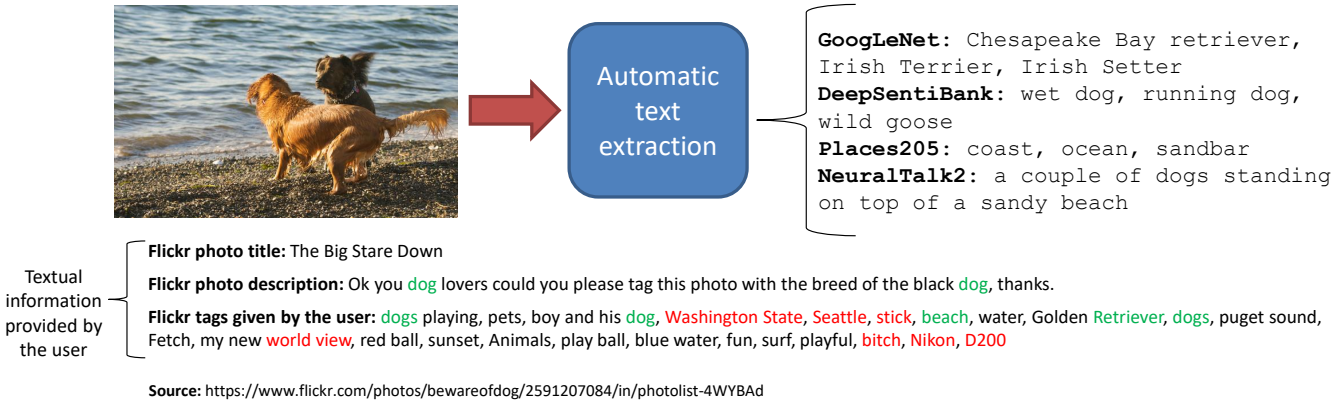

Fig. 1. Given an image, the proposed pipeline extracts Objective Text by exploiting four different deep learning architectures. The considered architectures are used to extract text related to objects, scene and image description. The subjective text provided by the user is also provided, it presents very noisy words which are highlighted in red. The words that appear either in the subjective and objective texts are highlighted in green.

images as positive or negative. They used the SentiWordNet [10] thesaurus to extract numerical sentiment values from Flickr metadata. 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 [8] 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.

A model that combines textual and visual information is presented in [25]. The subjective textual data such as comments and captions on the images are considered as contextual information. In [29] 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. The approach proposed in [12] combines visual features with text-based features extracted from the text subjectively associated by the users to images (i.e., descriptions and tags). 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 [16], [17], [30], 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 [10] are larger than a threshold, and applying the Bag of Words on this restricted 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 [7], [8], [25], [26]. 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 a text which is not related to the content. In previous approaches, the authors face several issues related to the subjective text associated to images. For instance, the framework presented in [25] implements an unsupervised approach aimed to address the lack of proper annotations/labels in the majority of social media images. In [30], 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 [31] the authors exploit large noisily annotated image collections to improve image classification.

III. Proposed Approach

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

A. 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. Most of the existing works exploit social subjective textual information associated to images either to define the ground truth [8] (i.e., by performing textual Sentiment Analysis on the text) or as an additional data modality (i.e., views) [12], [29]. 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 methods, it can be a very noisy source because it is provided by the users. There is no guarantee that the subjective text accompanying an image is useful for the sentiment analysis task. In addition, the tags associated to social images are often

selected by users with the purpose to maximize the visibility of such images by the platforms 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. Those tags are independent from the sentiment evoked by the images.

The semantic of an image may be given by a single object category [16], while the user-provided tags may include a number of additional terms correlated with the object which could be related to a larger vocabulary. These information are hence not always useful for sentiment analysis. Alternatively, the semantic might be given by the contribute of multiple keywords corresponding to objects, scene types, or attributes. Figure 1 shows an example image taken from the Flickr dataset used in [12]. The textual information below the image is the subjective text provided by the user. 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. For example, 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). 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 this paper proposes 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 obtained by employing four deep learning models trained to accomplish different visual inference tasks on the input image. In particular, the objective text associated to an image is obtained by considering the output text labels of *GoogLeNet* [23] (Object Classification), *Places205* [24] (Scene Recognition), *DeepSentiBank* [1] [4] (Adjective-Noun Pair) and the image description generated by *NeuralTalk2* [22] (Image Captioning). At the top right part of Figure 1 the objective text automatically extracted from the image is shown. The inferred text is very descriptive and each model provides distinctive information. The objective text extracted by the different scene understanding methods has a pre-defined structure, therefore all the images have the same quantity of textual objective information. For each considered classifier (i.e., GoogLeNet [23], DeepSentiBank [4] and Places205 [24]) the results are ranked by the output probability of the prediction and only the top three labels are considered. In particular, we considered the minimum number of labels that guarantee (in a probabilistic sense) a total classification probability close to 1. To this aim, we analysed the distribution over the output classification probabilities and observed that the first three labels allow to achieve a total classification probability very close to 1, avoiding to include noisy labels with respect to the visual content. Finally, the employed image captioning method (NeuralTalk2 [22]) provides an overall description of the scene which is not constrained to objects/places categories (e.g., actions).

[1]Our implementation exploits the *MVSO English* model provided by [32], that corresponds to the *DeepSentiBank* CNN fine-tuned to predict 4342 English Adjective Noun Pairs.

B. Features Extraction

The proposed approach exploits one visual view and three textual features based on the objective text extracted from the images, namely Objective Textual (OT), Objective Sentiment (OS) and Objective Revisited (OR) features. The following subsections details the feature extraction process.

1) Visual View: As in [12] we consider five image descriptors used in various Computer Vision tasks. In particular, we extracted 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 [33] and the SentiBank 1200 mid-level visual representation presented in [8]. Then, all the obtained representations have been reduced to 500 dimensions using Principal Component Analysis (PCA).

2) 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 [12]. 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. As shown in Figure 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. This allows to obtain a wide objective description of the image which takes into account different semantic aspects of the visual content. Redundant terms are not a drawback for the proposed approach, indeed the presence of more occurrences of similar or related terms (e.g., dog, dogs, retriever, setter, etc.) enhances the weight of these correct terms in the representation extracted by our framework, and reduces the effect of noisy results such as the third result extracted with DeepSentiBank in Figure 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 as in [12], [16], [34].

To further compare the considered Objective Textual representation with respect to other state of the art solutions, we implemented the feature extraction process described in [35]. 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 [10], 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. All the text-based features considered in the proposed approach 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, lemmatizing and stop words removal. 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. Starting from the pre-processed Objective Text, we propose to extract the following text-based features:

- **Objective Textual (OT):** we obtained this feature by computing a 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 [35] highlights an issue related to the use of SentiWordNet scores for sentiment analysis. Indeed, most of the existing sentiment feature extraction methods (including [12]) ignore words which neutral sentiment is higher than either positive and negative ones, albeit they comprise the 93.75% of SentiWordNet entries. In [35] the sentiment scores associated to the neutral words are modulated 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 word counts are weighted with the predominant sentiment score after the scores revisiting (positive, negative or zero if neutral). We use this process on the proposed Objective Text. The OR feature that 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 the 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.

C. Embedding Different Views

Recently, several papers for jointly modelling images and associated text with Canonical Correlation Analysis (CCA) have been proposed [12]–[17]. 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. In the CCA embedding space, projections of different views are directly comparable by a similarity function defined over the elements of the embedding space [16].

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:

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

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 [36], this optimization problem can be formulated as a standard eigenproblem.

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

IV. Experimental Settings And Results

A. Dataset

In [12] the authors performed the experiments with two different datasets crawled from Instagram and Flickr. 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). Following the experimental protocol, we discarded the images labelled as neutral. The final dataset used in the experiments has a total of 47235 images.

B. Embedded Vectors

In Section III-C 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. In this paper we exploited a weighted embedding transformation which emphasize the most significant projection dimensions [17]. The final representation of the data from the *i-th* view into the weighted embedding space is defined as:

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

TABLE I

PERFORMANCE EVALUATION OF THE PROPOSED METHOD WITH RESPECT TO THE BASELINE METHOD PRESENTED IN [12]. FOR EACH SETTING, THE AVERAGE AND STANDARD DEVIATION OF TEST CLASSIFICATION ACCURACY OVER 10 RUNS IS REPORTED. THE BEST RESULT IS HIGHLIGHTED IN BOLD, WHEREAS THE SECOND BEST RESULT IS UNDERLINED. SEE TEXT FOR DETAILS.

	Experiment ID	Embedded Views	Full Feature	Truncated Features (99%)
Subjective Features Proposed in [12]	K1	V+T+S	66.56 ±0.43 %	66.11 ±0.45 %
	K2	V+T	71.67 ±0.36 %	71.55 ±0.57 %
	K3	V+S	62.19 ±0.63 %	62.89 ±0.45
Considering Subjective and/or Objective Features	O1	V+T+OS	68.88 ±0.49 %	69.23 ±0.38 %
	O2	V+OT+S	66.97 ±0.57 %	66.34 ±0.68 %
	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 %

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], [16], [17].

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 (i.e., the minimum number of eigenvalues which normalized sum is greater or equal than 0.99). We call these representations *Truncated Features*.

C. 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. A linear SVM has been used to establish the sentiment polarity over the different compared representations. The parameter C of the linear SVM was determined by 10-fold cross validation.

Table I shows the obtained results. Each row describes a different experimental setting, corresponding to a specific combination of the input features described in Section III-B 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 Equation (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 I 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 [12] (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.

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 I). In particular, using the feature OT instead of T provides a mean accuracy improvement of 1.81% (compare O3 and K2 in Table I). Adding the view T to the experimental setting O3 yields an increment of 0.48% (O10 in Table I). Note that, adding the proposed OT feature to the experimental setting K2 (i.e., [12]) 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. 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 [12] is very limited. Furthermore, the sentiment score of most of the SentiWordNet words is often neutral. Therefore, most of the words of the sentiment vocabulary are still neutral after the re-ranking.

V. CONCLUSION AND FUTURE WORKS

This paper addresses the challenge of image sentiment polarity estimation by proposing a novel source of text for this task. The aim is to deal with the issue related to the text provided by users which is commonly used in most of the previous works. We presented a study in which Objective Text extracted considering the visual content of images is compared with respect to the Subjective Text provided by users. This study first identifies several drawbacks brought by the Subjective Text due its intrinsic nature, and then demonstrates experimentally that the exploitation of Objective Text associated to images provides better results than the use of the Subjective Text provided by the user. The Objective Text exploited by our approach does not present the highlighted

issues, and it is automatically extracted from the image. 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.

REFERENCES

[1] S. Battiato, G. M. Farinella, F. L. Milotta, A. Ortis, L. Addesso, A. Casella, V. D'Amico, and G. Torrisi, "The social picture," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 2016, pp. 397–400.

[2] S. Rudinac, M. Larson, and A. Hanjalic, "Learning crowdsourced user preferences for visual summarization of image collections," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1231–1243, 2013.

[3] S. Qian, T. Zhang, C. Xu, and J. Shao, "Multi-modal event topic model for social event analysis," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 233–246, 2016.

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

[5] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 83–92.

[6] J. Yuan, S. Mcdonough, Q. You, and J. Luo, "Sentribute: image sentiment analysis from a mid-level perspective," in *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*. ACM, 2013.

[7] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *Proceedings of the 18th ACM International Conference on Multimedia*. ACM, 2010, pp. 715–718.

[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 *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, 2013, pp. 223–232.

[9] G. A. Miller, "Wordnet: a lexical database for english," in *Communications of the ACM*, vol. 38, no. 11. ACM, 1995, pp. 39–41.

[10] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of The European Language Resources Association*, vol. 6. Citeseer, 2006, pp. 417–422.

[11] Z. Yuan, J. Sang, and C. Xu, "Tag-aware image classification via nested deep belief nets," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–6.

[12] M. Katsurai and S. Satoh, "Image sentiment analysis using latent correlations among visual, textual, and sentiment views," in *In proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2016, pp. 2837–2841.

[13] S. J. Hwang and K. Grauman, "Accounting for the relative importance of objects in image retrieval." in *Proceedings of British Machine Vision Conference, Vol. 1*, no. 2, 2010.

[14] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM International Conference on Multimedia*. ACM, 2010, pp. 251–260.

[15] S. J. Hwang and K. Grauman, "Learning the relative importance of objects from tagged images for retrieval and cross-modal search," *International Journal of Computer Vision*, vol. 100, no. 2, pp. 134–153, 2012.

[16] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210–233, 2014.

[17] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong, "Transductive multi-view embedding for zero-shot recognition and annotation," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 584–599.

[18] L. Pang, S. Zhu, and C.-W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2008–2020, 2015.

[19] P. Cui, S. Liu, and W. Zhu, "General knowledge embedded image representation learning," *IEEE Transactions on Multimedia*, 2017.

[20] Z. Yuan, J. Sang, C. Xu, and Y. Liu, "A unified framework of latent feature learning in social media," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1624–1635, 2014.

[21] X. Yang, T. Zhang, and C. Xu, "Cross-domain feature learning in multimedia," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 64–78, 2015.

[22] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[24] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems*, 2014, pp. 487–495.

[25] Y. Wang, S. Wang, J. Tang, H. Liu, and B. Li, "Unsupervised sentiment analysis for social media images," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina*, 2015, pp. 2378–2379.

[26] T. Mike, B. Kevan, P. Georgios, C. Di, and K. Arvid, "Sentiment in short strength detection informal text," *Journal of the Association for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, 2010.

[27] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. Snoek, and A. D. Bimbo, "Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval," *ACM Computing Surveys (CSUR)*, vol. 49, no. 1, p. 14, 2016.

[28] J. Johnson, L. Ballan, and L. Fei-Fei, "Love thy neighbors: Image annotation by exploiting image metadata," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4624–4632.

[29] 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*, vol. 75, no. 5, pp. 2507–2525, 2016.

[30] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 529–545.

[31] G. Wang, D. Hoiem, and D. Forsyth, "Building text features for object image classification," in *In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1367–1374.

[32] 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 *Proceedings of the 23rd ACM International Conference on Multimedia*. ACM, 2015, pp. 159–168.

[33] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang, "Designing category-level attributes for discriminative visual recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 771–778.

[34] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 902–909.

[35] C. Hung and H.-K. Lin, "Using objective words in sentiwordnet to improve sentiment classification for word of mouth," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 47–54, 2013.

[36] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.