From Image to Text in Sentiment Analysis via Regression and Deep Learning

Daniela Onita
Faculty of Mathematics and Computer Science
University of Bucharest
danielaonita25@gmail.com

Liviu P. Dinu
Faculty of Mathematics and Computer Science
University of Bucharest
ldinu@fmi.unibuc.ro

Birlutiu Adriana
Computer Science Department
1 Decembrie 1918 University of Alba Iulia
adriana.birlutiu@uab.ro

Abstract

Images and text represent types of content which are used together for conveying user emotions in online social networks. These contents are usually associated with a sentiment category. In this paper, we investigate an approach for mapping images to text for three types of sentiment categories: positive, neutral and negative. The mapping from images to text is performed using a Kernel Ridge Regression model. We considered two types of image features: *i)* RGB pixel-values features, and *ii)* features extracted with a deep learning approach. The experimental evaluation was performed on a Twitter data set containing both text and images and the sentiment associated with these. The experimental results show a difference in performance for different sentiment categories, in particular the mapping that we propose performs better for the positive sentiment category in comparison with the neutral and negative ones. Furthermore, the experimental results show that the more complex deep learning features perform better than the RGB pixel-value features for all sentiment categories and for larger training sets.

1 Introduction

A quick look at an image is sufficient for a human to say a few words related to that image. However, this very easy task for humans is a very difficult task for the existing computer vision systems. The majority of previous work in computer vision has focused on labeling images with a fixed set of visual categories. However, even though closed vocabularies of visual concepts are a convenient modeling assumption, they are quite restrictive when compared to the vast amount of rich descriptions and impressions that a human can compose.

Some approaches that address the challenge of generating image descriptions have been proposed (Kulkarni et al., 2013; Karpathy and Fei-Fei, 2015). However, these models only rely on objective image descriptors, and do not take into account the subjectivity which appears when describing an image on social networks.

In this work, we want to take a step forward towards the goal of generating subjective descriptions of images that are close to the natural language that is used in social networks. Figure 1 gives a hint to the motivation of our work by showing several samples which were used in the experimental evaluation. Each sample consists of an image and the subjective text associated to it, and has has a sentiment associated to it: negative, neutral or positive.

The goal of our work is to generate subjective descriptions of images. The main challenge towards this goal is in the design of a model that is rich enough to simultaneously reason about contents of images and their representations in natural language domain. Additionally, the model should be free of assumptions about specific templates or categories and instead rely on learning from the training data. The model will go beyond the simple description of an image and give also a subjective impression that the image could make upon a certain person. An example of this is shown in the image from bottom-right of Figure 1, in which we do not have a captioning or description of the animal in the image but the subjective impression that the image makes upon the looker.

Our core insight is that we can map images to subjective natural text by leveraging the image-sentence data set in a supervised learning approach in which the image represents the input and the

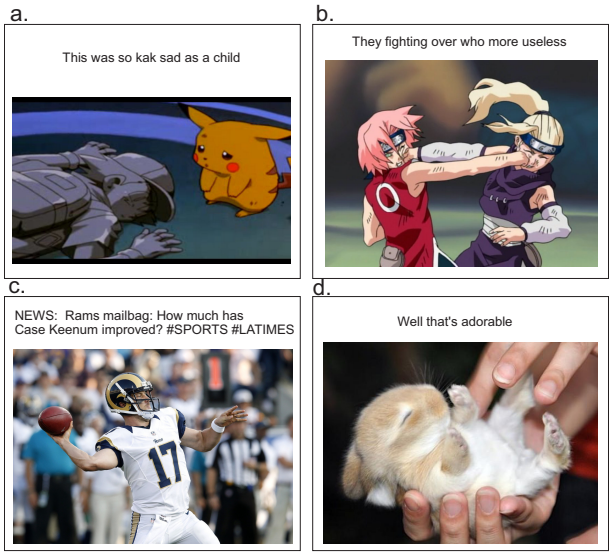

Figure 1: Motivation Figure: Our model treats language as a rich label space and generates subjective descriptions of images. Examples of samples used in the experimental evaluation. Each sample consists of a pair made of an image and the subjective text associated to it. Each sample has a sentiment associated to it: a., b. samples convey a negative sentiment; c. sample conveys a neutral sentiment; d. sample conveys a positive sentiment.

sentence represents the output. We employ a Kernel Ridge Regression for the task of mapping images to text. We considered two types of image features: *i)* RGB pixel-values features, and *ii)* features extracted with a deep learning approach. We used a bag-of-words model to construct the text features. In addition, we consider several sentiment categories associated to each image-text sample, and analyze this mapping in the context of these sentiment categories.

We investigate data from Twitter. These data contain images and text associated to each image. The text is a subjective description or impression of the image, written by a user. Data from social networks, and especially Twitter, is usually associated to a sentiment, which could be a positive, neutral or negative sentiment. We designed a system that automatically associates an image to a set of words from a dictionary, these words being not only descriptors of the content of the image, but also subjective impressions and opinions of the image.

One of the interesting findings of our work is that there is a difference in performance for different sentiment categories, in particular the mapping performs better for the positive sentiment category in comparison with the neutral and negative categories. Furthermore, the experimental results show that the more complex deep learning features perform better than the RGB pixel-value features for all sentiment categories and for larger training sets.

The paper is organized as follows. Section 2 discusses related works. Section 3 describes a Kernel Ridge Regression model for image to text mapping. Section 4 shows the experimental evaluation performed on a real-world data set. Section 5 finishes with conclusions and directions for future research.

2 Related Work

Image captioning. The research presented in this paper is in the direction of image captioning, but goes further to map images to text. The texts that we consider are not only descriptions of the images, which is the task of image captioning, but they contain subjective statements related to the images. Mapping images to text is an extension of the image captioning task, and this mapping allows us to build some dictionaries of words and select from these dictionaries the words which are the most relevant to an image. The learning setting that we investigate in this paper is different to the image captioning setting, because our system automatically associates an image to a set of words from a dictionary, these words being not only descriptors of the content of the image, but also subjective opinions of the image. Image captioning has been actively studied in last years, a recent survey on image captioning is given in (Bai and An, 2018). Several approaches for image captioning are making use of the deep learning techniques (Bai and An, 2018; P. Singam, 2018).

Image description. Several approaches that address the challenge of generating image descriptions have been proposed (Kulkarni et al., 2013; Karpathy and Fei-Fei, 2015; Park et al., 2017; Ling and Fidler, 2017). However, these models only rely on objective image descriptors, and do not take into account the subjectivity which appears when describing an image on social networks.

Sentiment analysis. We investigate mapping images to text in the context of sentiment analysis. Most of the previous research in sentiment analysis is performed on text data. Recent works focus

on sentiment analysis in images and videos (Yu et al., 2016; You et al., 2015; Wang et al., 2016). The research on visual sentiment analysis proceeds along two dimensions: *i)* based on handcrafted features and *ii)* based on features generated automatically. Deep Learning techniques are capable of automatically learning robust features from a large number of images (Jindal and Singh, 2015). An interesting direction for sentiment analysis is related to word representations and capsule networks for NLP applications (Xing et al., 2019; Zhao et al., 2019).

3 Kernel Ridge Regression for Mapping Images to Text

Let $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ be the set of inputs and outputs, respectively, and n represents the number of observations. And let $F_X \in \mathbb{R}^{d_X \times n}$ and $F_Y \in \mathbb{R}^{d_Y \times n}$ denote the input and output feature matrices, where d_X, d_Y represent the dimensions of the input and output features respectively. The inputs represent the images, and the input features can be either simple RGB pixel-values or something more complex, such as features extracted automatically using convolutional neural networks (O'Shea and Nash, 2015). The outputs represent the texts associated to the images and the output features can be extracted using Word2Vec (Ma and Zhang, 2015).

A mapping between the inputs and the outputs can be formulated as a multi-linear regression problem (Cortes et al., 2005, 2007). Combined with Tikhonov regularization, this is also known as Kernel Ridge Regression (KRR). The KRR method is a regularized least squares method that is used for classification and regression tasks. It has the following objective function:

$$\arg_W \min(\frac{1}{2}||WF_X - F_Y^T||_{\mathcal{F}}^2 + \alpha\frac{1}{2}||W||_{\mathcal{F}}^2) \quad (1)$$

where $||\cdot||_{\mathcal{F}}$ is the Frobenius norm, α is a regularization term and the superscript ^{T} signifies the transpose of the matrix.

The solution of the optimization problem from Equation 1 involves the Moore-Penrose pseudo-inverse and has the following closed-form expression:

$$W = F_Y F_X^T (F_X F_X^T + \alpha I_{d_X})^{-1} \in \mathbb{R}^{d_Y \times d_X} \quad (2)$$

which for low-dimensional feature spaces $(d_X, d_Y \leq n)$ can be calculated explicitly (the I_{d_X} in Equation 2 represents the identity matrix of dimension d_X).

For high-dimensional data, an explicit computation of W as presented in Equation 2 without prior dimensionality reduction is computationally expensive. Fortunately, Equation 2 can be rewritten as:

$$\begin{aligned} W &= F_Y F_X^{\mathrm{T}} (F_X F_X^{\mathrm{T}} + \alpha \mathrm{I}_{d_x})^{-1} \\ &= F_Y (F_X^{\mathrm{T}} F_X + \alpha \mathrm{I}_n)^{-1} F_X^{\mathrm{T}} \end{aligned} \quad (3)$$

Making use of the kernel trick, the inputs x_i are implicitly mapped to a high-dimensional Reproducing Kernel Hilbert space (Berlinet and Thomas-Agnan, 2011):

$$\Phi = [\phi(x_1), \ldots, \phi(x_n)]. \quad (4)$$

When predicting a target y_{new} from a new observation x_{new}, explicit access to Φ is never actually needed:

$$\begin{aligned} y_{\text{new}} &= F_Y (\Phi^{\mathrm{T}} \Phi + \alpha \mathrm{I}_n)^{-1} \Phi^{\mathrm{T}} \phi(x_{\text{new}}) \\ &= F_Y (K + \alpha \mathrm{I}_n)^{-1} \kappa(x_{\text{new}}) \end{aligned} \quad (5)$$

With $K_{ij} = \phi(x_i)^{\mathrm{T}} \phi(x_j)$ and $\kappa(x_{\text{new}})_i = \phi(x_i)^{\mathrm{T}} \phi(x_{\text{new}})$, the prediction can be described entirely in terms of inner products in the higher-dimensional space. Not only does this approach work on the original data sets without the need of dimensionality reduction, but it also opens up ways to introduce non-linear mappings into the regression by considering different types of kernels, such as a Gaussian or a polynomial kernel.

4 Experimental Evaluation

4.1 Dataset

We used a data set with images and text that was introduced in (Vadicamo et al., 2017). The data have been collected from Twitter posts over a period of 6 months, and using an LSTM-SVM architecture, the tweets have been divided into three sentiment categories: positive, neutral, and negative. For image labelling the authors have selected data with the most confident textual sentiment predictions and they used these predictions to automatically assign sentiment labels to the corresponding images. In our experimental evaluation we selected 10000 images and the corresponding 10000 tweets from each of the three sentiment categories. Figure 1 shows examples of image and text data used in the experimental evaluation.

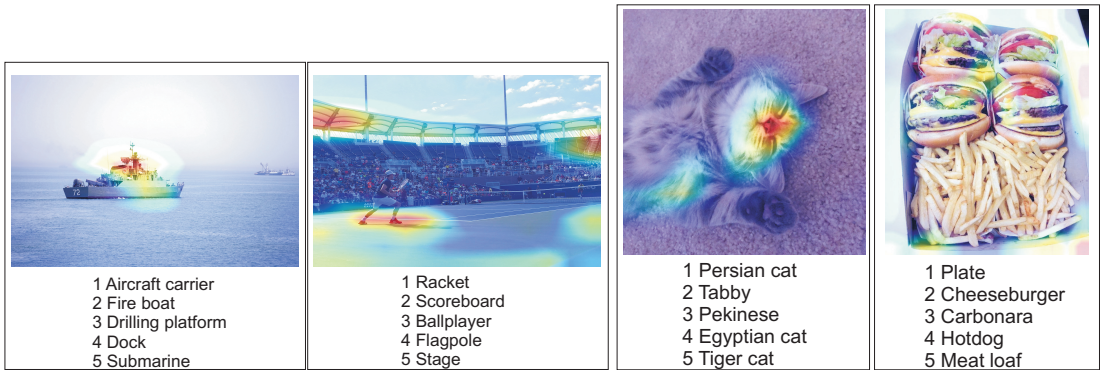

Figure 2: Visualizing heatmaps of class activation in an image.

4.2 Image and Text Features

4.2.1 Image Features

The research on feature extraction from images proceeds along two directions: i) traditional, handcrafted features, and ii) automatically generated features. With the increasing number of images and videos on the web, traditional methods have a hard time handling the scalability and generalization problem. In contrast, automated generated feature-based techniques are capable to automatically learn robust features from a large number of images (Jindal and Singh, 2015). We discuss below how these two directions for extracting features from images apply in our case, in particular, we use RGB pixel-values features for the first direction and Deep Learning based features for the second direction.

RGB pixel-values. In this approach for extracting features from images, we simply convert the images into arrays. Each image was sliced to get the RGB data. The 3-channels RGB image format was preferred instead of using 1-channel image format since we wanted to use all the available information related to an image. Using this approach, each image was described by a 2352 (28 x 28 x 3)-dimensional feature vector.

Deep Learning based features. Deep Learning models use a cascade of layers to discover feature representations from data. Each layer of a convolutional network produces an activation for the given input. Earlier layers capture low-level features of the image like blobs, edges, and colors. This primitive features are abstracted by the high-level layers. Studies from the literature suggest that while using pre-trained networks for feature extraction, the features should be extracted from the layer right before the classification layer (Rajaraman et al., 2018). For this reason, we extracted the features from the last layer before the final classification, so the entire convolutional base was used for this. The features were extracted using the pre-trained convolutional base VGG16 network (Simonyan and Zisserman, 2014). For computational reasons, the images were resampled to a 3232 pixel resolution. The model was initialized by the ImageNet weights. For understanding what part of an image was used to extract the features, visualizing heatmaps of class activation technique was employed. This is a technique which illustrates how intensely the input image activates different channels, how important each channel is with regard to the class and how intensely the input image activates the class. Figure 2 illustrates the heatmaps of class activation for some random images using VGG16 as a pre-trained convolutional base. The VGG16 model makes the final classification decision based on the highlighted parts from each image, and furthermore each image is associated with the five most representative captions.

4.2.2 Text Features

We used a Bag-of-Words (BoW) model (Harris, 1954) for extracting the features from the text samples. The first step in building the BoW model consists of pre-processing the text: removing non-letter characters, removing the html tag from the Twitter posts, converting words to lower cases, removing stop-words and making the split. A vocabulary is built from the words that appear in the text samples. The input of the BoW model is a list of strings and the output is a sparse matrix with the dimension: number of samples x number of words in the vocabulary, having 1 if a given word from the vocabulary is contained in that particular

text sample. We initialized the BoW model with a maximum of 5000 features. We extracted a vocabulary for each sentiment category, and the corresponding 0-1 feature vector for each text sample.

4.3 Experimental Results

Evaluation Measure

Each output of our model represents a very large vector of probabilities, with the dimension equal to the number of words in the dictionary (approximately 5000 components). Each component of the output vector represents the probability of the corresponding word from the vocabulary as being a descriptor of that image. Given this particular form of the output, the evaluation measure was computed using the following algorithm:

1. we sorted in descending order the absolute values of the predicted output vector;
2. we created a new vector containing the first 50 words from the predicted output vector;
3. we computed the Euclidean distance between the predicted output vector values and the actual output vector.

The actual output vector is a sparse vector, a component in this vector is 1 if the corresponding word from the vocabulary is contained in that particular description of the image.

The values computed in step 3) described above were averaged over the entire test data set and the average value obtained was considered as the error.

Experimental Protocol

We designed an experimental protocol, that would help us answer the following questions:

1. Could our proposed Kernel Ridge Regression model map images to natural language descriptors?
2. What is the difference between the two types of image features that we considered? In particular, we are interested whether the more complex deep learning features give a better performance in comparison to the simple RGB pixel-values features.
3. Is there a difference in performance based on the sentiment associated to each image-text sample?

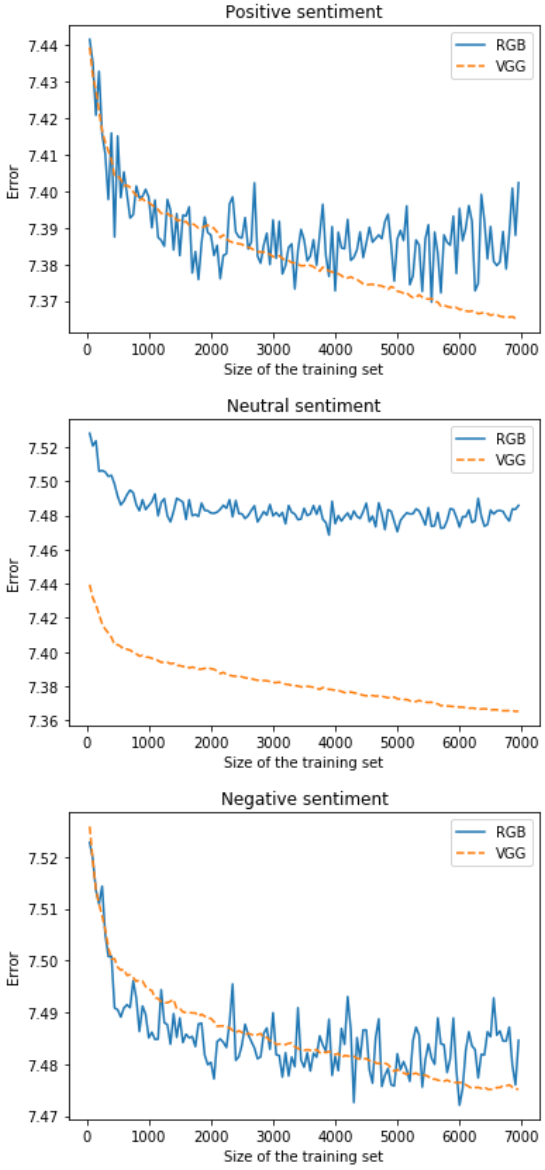

Figure 3: The plots show mean errors and standard deviation for different sizes of the training set. Comparison between RGB pixel-values features and the more complex VGG16 features. The different rows correspond to different sentiment categories: top row - positive sentiment category, middle row - neutral sentiment category, bottom row - negative sentiment category.

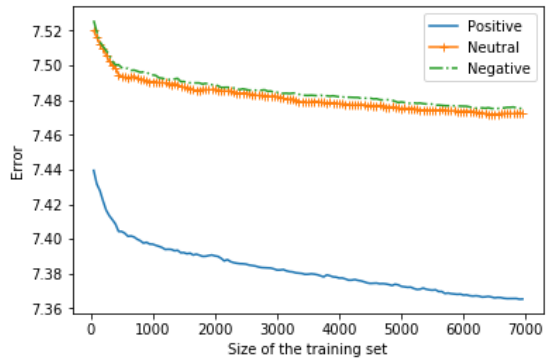

Figure 4: Comparison of the learning performance based on the type of sentiment using the VGG16 image features.

We designed the following experimental protocol. For each of the three sentiment categories, we randomly split the data 5 times into training and testing, taking 70% for training and the rest for testing. For training the model, we considered different sizes of the training set: from 50 to 7000 observations with a step size of 50. For a correct evaluation, the models built on these different training sets, were evaluated on the same test set. The error was averaged over the 5 random splits of the data into training and testing.

Results

The first two questions raised above can be answered by analyzing the experimental results shown in Figure 3. The plots show the learning curve (mean errors and standard deviations) for different sizes of the training set and for different sentiment categories. Since the error decreases as the training size increases, we can say that there is a learning involved, thus our proposed model can map images to natural language descriptors.

The plots from Figure 3 also show the comparison between the RGB pixel-values and VGG16 features for the three categories of sentiments considered. Overall, the more complex deep learning features give a better performance in comparison to the simple RGB pixel-values features.

To answer the third question, we analyzed the experimental results shown in Figure 4. There is a significant difference in learning performance for the positive sentiment category in comparison with the other two categories, both using RGB pixel-values features and VGG16 features. The positive category is simpler to be learned because of the subjective part from images: a positive feeling can be interpreted as positive for the majority of the people, but a neutral or a negative sentiment can be interpreted as having a different meaning depending on the people.

Furthermore, analyzing again Figure 3, we see that the neutral sentiment category has a different behaviour in comparison with the positive and negative sentiment categories, with respect to the image features used. In the case of neutral sentiment, the more complex VGG16 features appear to have a better performance than the simpler RGB pixel-values features as the size of the data increases. For positive and negative sentiment categories the simpler RGB pixel-values features lead to an error which varies a lot, while using the VGG16 features, the error is more stable.

5 Conclusions and Future Work

In this work, we investigated a method for image to text mapping in the context of sentiment analysis. The mapping from images to text was performed using a Kernel Ridge Regression model. We considered two types of image features: *i)* the simple RGB pixel-values features, and *ii)* a more complex set of image features extracted with a deep learning approach. Furthermore, in this paper we took a step forward form the image captioning task, which allows us to build some dictionaries of words and select from these dictionaries the words which are the most relevant to an image. We performed the experimental evaluation on a Twitter data set containing both text and images and the sentiment associated with these. We found that there is a difference in performance for different sentiment categories, in particular the mapping performs better for the positive sentiment category in comparison with the neutral and negative ones for both features extraction techniques.

We plan to further extend our approach by investigating the input-output kernel regression type of learning (Brouard et al., 2016). The output kernel would allow us to take into account the structure in the output space and benefit from the use of kernels. We also plan to integrate in our model textual captions of images obtained using a pre-trained network (Simonyan and Zisserman, 2014). The textual captions could be used as a new type of features and can be compared and integrated with the other two types of image features considered.

References

Bai, S. and An, S. (2018). A survey on automatic image caption generation. *Neurocomputing*, 311.

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science Business Media.

Brouard, C., Szafranski, M., and D'Alché-Buc, F. (2016). Input output kernel regression: supervised and semi-supervised structured output prediction with operator-valued kernels. *The Journal of Machine Learning Research*, 17(1):6105–6152.

Cortes, C., Mohri, M., and Weston, J. (2005). A general regression technique for learning transductions. In *Proceedings of ICML 2005*, pages 153–160. ACM Press.

Cortes, C., Mohri, M., and Weston, J. (2007). A general regression framework for learning stringto-string mappings. In *Predicting Structured Data*. MIT Press.

Harris, Z. (1954). Distributional structure. pages 146–62.

Jindal, S. and Singh, S. (2015). Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning. In *2015 International Conference on Information Processing (ICIP)*, pages 447–451. IEEE.

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. (2013). Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903.

Ling, H. and Fidler, S. (2017). Teaching machines to describe images via natural language feedback. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5075–5085. Curran Associates Inc.

Ma, L. and Zhang, Y. (2015). Using word2vec to process big text data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2895–2897. IEEE.

O'Shea, K. and Nash, R. (2015). An introduction to convolutional neural networks. *ArXiv e-prints*.

P. Singam, Ashvini Bhandarkar, B. M. C. T. K. K. (2018). Automated image captioning using convnets and recurrent neural network. *International Journal for Research in Applied Science and Engineering Technology*.

Park, C., Kim, B., and Kim, G. (2017). Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 895–903.

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., and Thoma, G. R. (2018). Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, 6:e4568.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Vadicamo, L., Carrara, F., Cimino, A., Cresci, S., Dell'Orletta, F., Falchi, F., and Tesconi, M. (2017). Cross-media learning for image sentiment analysis in the wild. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.

Wang, J., Fu, J., Xu, Y., and Mei, T. (2016). Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 3484–3490. AAAI Press.

Xing, F. Z., Pallucchini, F., and Cambria, E. (2019). Cognitive-inspired domain adaptation of sentiment lexicons. *Information Processing & Management*, 56(3):554–564.

You, Q., Luo, J., Jin, H., and Yang, J. (2015). Joint visual-textual sentiment analysis with deep neural networks. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 1071–1074, New York, NY, USA. ACM.

Yu, Y., Lin, H., Meng, J., and Zhao, Z. (2016). Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. *Algorithms*, 9(2):41.

Zhao, W., Peng, H., Eger, S., Cambria, E., and Yang, M. (2019). Towards scalable and reliable capsule networks for challenging nlp applications. *arXiv preprint arXiv:1906.02829*.