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

Daniela Onita[1], **Liviu P. Dinu**[1], **Adriana Birlutiu**[2]

[1] University of Bucharest, Faculty of Mathematics and Computer Science

[2] 1 Decembrie 1918 University of Alba Iulia, Computer Science Department

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. There is a difference in performance for different sentiment categories, the mapping performs better for the positive sentiment category. Furthermore, the more complex deep learning features perform better than the RGB pixel-value features for all sentiment categories.

Motivation

Giving a subjective description of an image is a very easy task for humans, but very difficult for the existing computer vision systems [3, 2]. The majority of previous work in computer vision has focused on labeling images with a fixed set of visual categories [1, 4]. In this work, we 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. Fig. 1 gives a hint to the motivation of our work:

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.

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 features extracted automatically using convolutional neural networks. The outputs represent the texts associated to the images and the output features.
- A mapping between the inputs and the outputs can be formulated as a multi-linear regression problem. 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:

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

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

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

- Predicting a target y_{new} from a new observation x_{new}:

$$\begin{aligned} y_{\text{new}} &= F_Y(\Phi^{\mathrm{T}}\Phi + \alpha \mathrm{I}_n)^{-1}\Phi^{\mathrm{T}}\phi(x_{\text{new}}) \\ &= F_Y(K + \alpha \mathrm{I}_n)^{-1}\kappa(x_{\text{new}}) \end{aligned} \quad (3)$$

where ϕ is the high-dimensional mapping of the inputs x_i to a Reproducing Kernel Hilbert space and

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

- 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.
- This formulation supports non-linear mappings by a Gaussian or a polynomial kernel.

Experimental Evaluation

Dataset

- We used a data set with images and text that introduced in [5].
- The data have been divided into three sentiment categories: positive, neutral, and negative using LSTM-SVM.
- 10000 images and the corresponding 10000 tweets from each of the three sentiment categories were selected.

Image Features

- **RGB pixel-values** - each image was described by a 2352 dimensional feature vector.
- **Deep Learning based features** - the features were extracted using the pre-trained convolutional base VGG16 network.

Text Features

- Bag-of-Words (BoW) model was used for extracting the features from the text samples.
- A vocabulary was built for each sentiment category.
- Features: 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.

Experimental Protocol

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

Q1 Could our proposed Kernel Ridge Regression model map images to natural language descriptors?

Q2 What is the difference between the two types of image features that we considered?

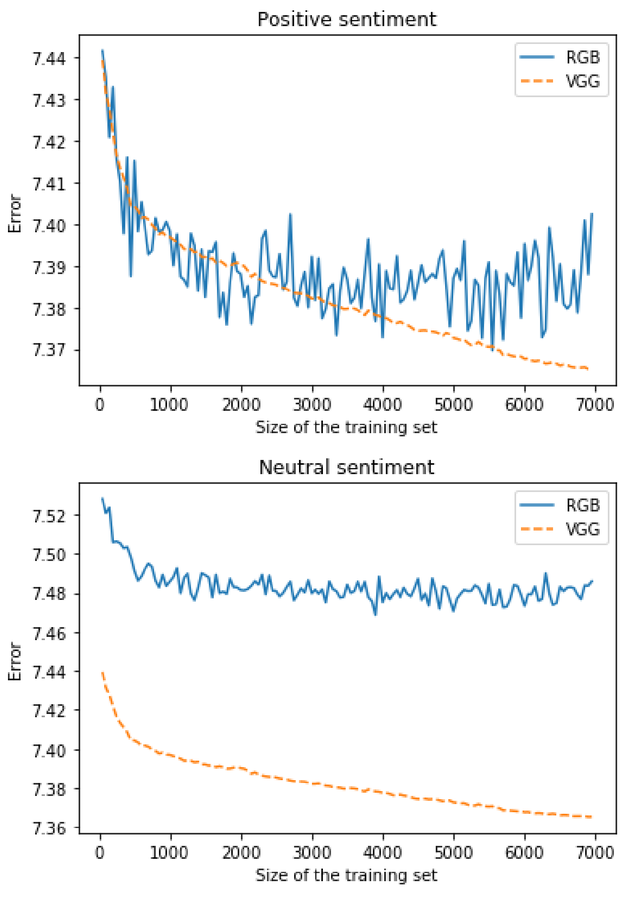
Q3 Is there a difference in performance based on the sentiment associated to each image-text sample?

For each of the three sentiment categories:

- we randomly split the data 5 times into training and testing;
- 70% data 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.

Evaluation Measure

Given the particular form of the outputs, we used an atypical evaluation measure which we designed to take into account the specific form of these outputs. We sorted in descending order the absolute values of the predicted output vector. A new vector containing the first 50 words from the predicted output vector was created. We computed the Euclidean distance between the predicted output vector values and the actual output vector. The values were averaged over the entire test data set and the average value obtained was considered as the error.



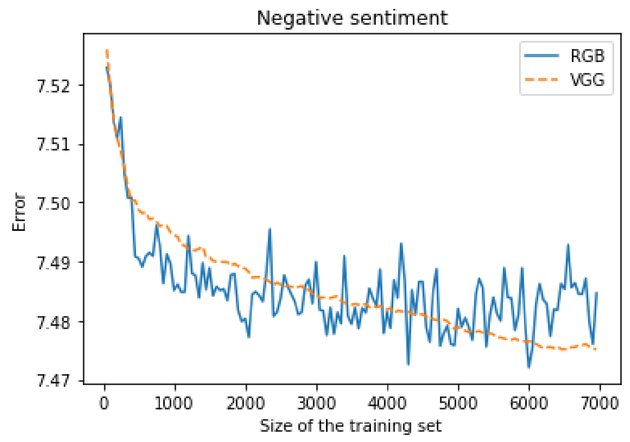


Figure 2: The plots show mean errors and standard deviation for different sizes of the training set. Comparison between RGB pixel-values features and the more complex VGG16 features. The different rows correspond to different sentiment categories: top row - positive sentiment category, middle row - neutral sentiment category, bottom row - negative sentiment category.

Results

- The answer for Q1 and Q2: analyzing the experimental results shown in Fig. 2. The plots show the learning curve (mean errors and standard deviations) for different sizes of the training set and for different sentiment categories.
- The plots from Fig. 2 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.
- The answer for Q3: analyzing the experimental results shown in Fig. 3. 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.

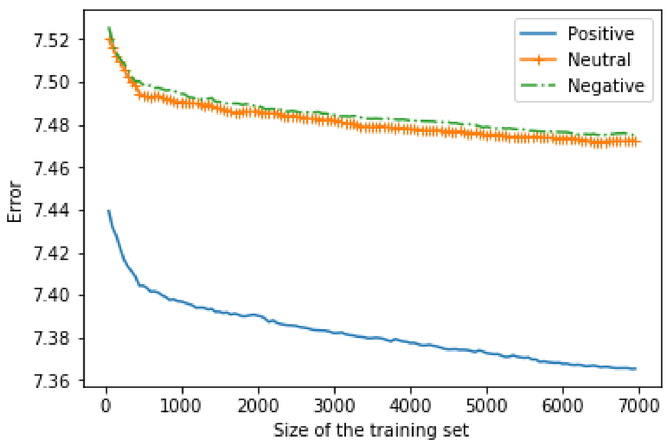


Figure 3: Comparison of the learning performance based on the type of sentiment using the VGG16 image features.

Conclusions and Future Work

- The mapping performs better for the positive sentiment category in comparison with the neutral and negative ones for both features extraction techniques.
- The more complex deep learning features perform better than the RGB pixel-value features for all sentiment categories and for larger training sets.
- Future plans: investigating the input-output kernel regression type of learning and integrating to our models textual captions of images obtained using a pre-trained network.

References

[1] Shuang Bai and Shan An. A survey on automatic image caption generation. *Neurocomputing*, 311, 05 2018.

[2] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[3] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013.

[4] Huan Ling and Sanja Fidler. 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., 2017.

[5] Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. Cross-media learning for image sentiment analysis in the wild. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

RANLP 2019 International Conference "Recent Advances in NLP", Varna, Bulgaria, September 2019