

# Data Augmentation Method for the Image Sentiment Analysis

Alexander Rakovsky<sup>1</sup>, Arseny Moskvichev<sup>2</sup>, Andrey Filchenkov<sup>1</sup>

<sup>1</sup>ITMO University

Saint Petersburg, Russia

<sup>2</sup>Saint Petersburg State University

Saint Petersburg, Russia

{str13r, arseny.moskvichev}@gmail.com, afilchenkov@corp.mail.ru

**Abstract**— Training convolutional neural networks, which are the most successful models in the field of image sentiment analysis, requires a massive dataset. Acquiring such a dataset usually implies manual labeling of a large collection of images, which is slow and expensive. In this paper, we report the preliminary results of the automated data collection and labeling method, which is based on the use of hashtags provided by the FLICKR image sharing social network users, and word2vec word embeddings. At first, a set of images was acquired, each image's emotional coloring was estimated by human assessors. The hashtags of every particular image were converted to the vector representation and then averaged. The supervised learning algorithm was then trained to predict the level of image positiveness, based on the hashtag vector representation. After that, the algorithm can be used to predict the labels of previously unseen images, thus substantially broadening the dataset. The proposed approach offers noticeable benefits when compared to the alternatives, presenting an optimal balance of simplicity and efficacy. Additionally, we present the results, provided by a convolutional neural network regressor trained on the acquired data, and discuss the directions for further research.

## I. INTRODUCTION

While the task of text sentiment analysis had for a long time been attracting researchers' attention, the problem of image sentiment analysis was not addressed as often. In recent years, however, more articles devoted to this task began to emerge [1]. One possible explanation for this trend is that the increased popularity of convolutional networks had spurred the search for new applications. Indeed, as for today, convolutional networks provide the best results in this field [2]. One of the obstacles for achieving even better results is that the large dataset is required for successful training of convolutional neural networks. Acquiring such a dataset implies manual labeling of the images, which is slow and expensive. At the same time, millions of images accompanied by hashtags or descriptions are being uploaded to the numerous social network websites. It is a huge source of information, the vast amount of labeling work is done by users absolutely for free. If we are to find a simple and reliable way of extracting the desired labels from the provided hashtags, it would dramatically simplify the data collection step. Presenting and testing one such method is the main impact of this paper.

It should be noted, however, that the goals of our project are twofold: firstly, we want to develop a method of automatically acquiring estimations of the image's emotional content, based on the hashtags; secondly, we want to use this method to acquire a large dataset of labeled images, so that we can train a convolutional network to predict the emotional content of the image. The accuracy of such a classifier would serve as a best estimation of the usefulness of the data augmentation solution. This is an ongoing project, and in this paper, we present the preliminary results on each of these two endeavors.

## II. METHOD

### A. Acquiring the data

In collaboration with an expert psychologist, we had manually formed a short list of positive and negative hashtags. There were only six negative and six positive words in this list. Positive part included such words as "happiness", "joy", "gladness", "fun", "delight" and "pleasure", while words like "sadness", "sorrow", "anxiety", "fear", "worry" and "stress" exemplify the negative part of the list. We shall use the term keywords to refer to the words in this initial list.

Using the FLICKR API, 1900 images and their corresponding hashtags were downloaded. Among these, a half of the images had at least one positive hashtag, and the other half had at least one negative. The image crawler was so constrained as to avoid downloading multiple images uploaded to the FLICKR website by the same user. We thus excluded the possibility of learning the individual author's style and preferences in image and hashtag choice.

Each image emotional coloring was rated by 3 assessors, on a 0–10 scale, 0 being "very negative", and 10 being "very positive". The assessors were working independently, and were paid \$0.01 per image. The final estimation of each image's emotional coloring was acquired by averaging the assessors' answers.

### B. Hashtag preprocessing

The keywords and hashtags were converted to the

multidimensional vector representation using the word2vec word embedding technique [3], [4]. These hashtag representations were averaged for each picture to obtain a single vector intended to represent the general semantic content of the picture's hashtags.

C. Estimating the image emotional coloring

Estimation of image emotional coloring is the crucial part of our work, since the high quality estimations are necessary for the augmented data to be useful in the subsequent training of the image sentiment analysis classifiers or regressors. The goal is to generate labels in such a way that they would match as closely as possible to those, provided by the human assessors. Four approaches for the label generation were tested.

- In the baseline approach, we simply assigned a label that corresponds to the coloring of the initial hashtag, by which the picture was found.
- In the naïve approach, we calculated the cosine distances from the average vectorized hashtags to each of the vectorized keywords. After that the distances corresponding to the same class (positive or negative) were averaged, and the binary predictions were made depending on which of the two resulting average distances was smaller. It should be noted that several variations of this naïve approach were tested, but they all produced similar low-quality results, and we only report one, to provide a comparison for other methods.
- In the dictionary approach, we used the AFINN library [5] to get the estimation of the hashtag emotional coloring (positive or negative).
- In the machine learning approach, we treat the obtained 12 distances as input features for the classifier. The binary outputs for the classifier were obtained by setting the 0.5 threshold on the average human assessor's image estimations. Thus, the classifier was trained to predict the true image sentiment labels provided by the manual estimation. Three standard machine learning algorithms were tested, and their performance was evaluated using 10-fold cross-validation. Another approach would be to treat the problem as a regression task, using averaged human assessor estimations as a continuous target variable and setting the threshold on the model's predictions to generate the label. This approach yielded almost identical results in terms of classification accuracy, therefore we do not list it as a separate method.

D. Image sentiment analysis model

The acquired dataset was then used to train the convolutional neural network regressor. The network implementation from the Keras package [6] was used; the architecture of the network is described on Fig. 1.

III. RESULTS

A. Data augmentation

As can be seen from Table I, the labels generated by the machine learning model closely match those provided by the human assessors. Indubitably, the extremely high numbers achieved by the machine learning approach can only be interpreted in comparison with the results of the baseline, naïve and dictionary methods. It is the fact that the accuracy in the machine learning case is much higher than in all other cases. This allows us to draw optimistic conclusions about the viability of the method.

When we treat the problem as a regression task, it is also informative to look at the pairwise correlations among the assessors' and the algorithm's estimations of the emotional coloring of the images. The average correlation between different assessors' estimations was 0.861853, while the average correlation between the KNN regression algorithm's prediction and the different assessors' estimations was 0.831628.

However, the acquired results cannot be directly compared to those of any other previous works, because the sample is highly specific (this issue is discussed in the Conclusion section) as well as the task itself.

B. Image sentiment analysis regression

In the Table II we present the results of the convolutional neural network image sentiment analysis regressor. The Table III serves to provide a few examples of the neural networks' predictions on the test set, and the human assessors' average positiveness estimations for these images.

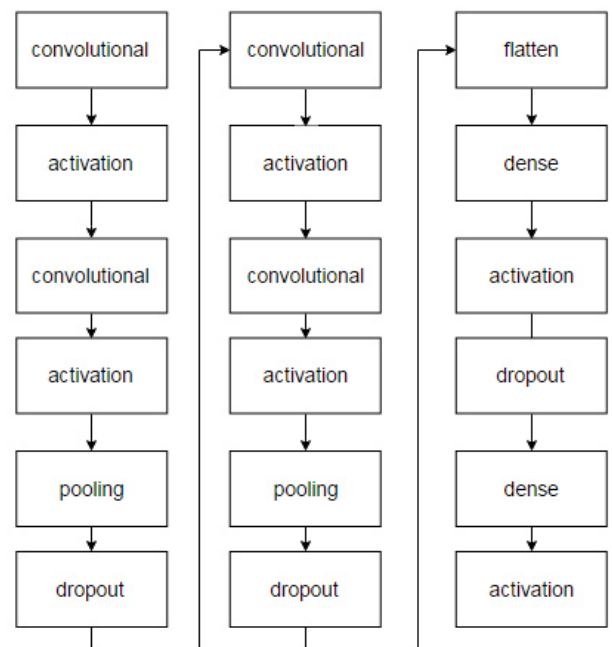


Fig. 1. Convolutional neural network layer structure. The rectified linear unit (ReLU) activation function was used in all layers except for the last one, where the linear activation was employed instead.

The mean squared error measure can be hard to interpret as it is, so in the Table II we provide the results for the random prediction, along with the network's results. The target variable was linearly scaled to the range from zero to one and the random predictions were generated by sampling from a continuous uniform distribution with the same range. It can clearly be seen that the network outperforms the baseline.

TABLE I. DATA AUGMENTATION METHODS COMPARISON

Method	Accuracy
Baseline approach	0.78
Naive approach	0.67
Dictionary usage (AFINN)	0.83
Naive Bayes classifier	0.81 ± 0.08
SVM	0.91 ± 0.08
kNN	0.95 ± 0.03

VI. CONCLUSION






A method for the automated hashtag-based image labeling is proposed and tested. The preliminary results suggest that thus acquired image labels may serve as a measure of the image emotional coloring, giving a quality nearly equal to that of manual labeling. Therefore, these labeled images can be used as a mean of acquiring additional data for training the image sentiment analysis classifier. The method does not require large amounts of data for the pre-training phase and does not require the semantic or syntactic preprocessing of the hashtags, which is fortunate, considering their irregular and ever-changing nature. There is still room for improvement; for example, using the embedding specifically trained on the hashtag corpora (like in [7]) may help to achieve even better results.

TABLE II. IMAGE SENTIMENT REGRESSION RESULTS

Method	MSE
Random prediction	0.14±0.04
Convolutional neural network	0.08±0.03

There are, however, some limitations. An image has to have hashtags to be thus labeled, and there are no guarantees that the set of images with no hashtags has the same emotional coloring as the set of images that have at least one hashtag. Moreover, an additional investigation is required to estimate the method performance in the case of labeling a random image from the set of hashtagged images, as opposed to labeling the images that are guaranteed to have at least one emotionally colored hashtag. In other words, the acquired sample of images is not fully representative of the set of all images that one might want to classify as positive or negative.

TABLE III. IMAGE POSITIVENESS ESTIMATION EXAMPLES

Image	Model	Human assessors
	0.62	0.65
	0.57	0.27
	0.49	0.54
	0.4	0.36
	0.31	0.61

Estimation of the degree to which the acquired emotionally hashtagged images are representative of the general set of images, and to what extent its representativeness is affecting the acquired results is the main goal of our continuing work. The new images are being collected and their emotional coloring is being estimated. This time, the images are selected completely at random, thus forming a more representative sample. The overall procedure will be approximately the same, but we will not keep the set of keywords fixed anymore, so that it may be optimized as well.

Overall, the described data augmentation method is easy to use and it outperforms the other methods of similar complexity on the acquired data. Moreover, thus acquired labels can be used to train image sentiment analysis models. However, more experiments are to be done to provide a more accurate estimate of the method's performance.

#### ACKNOWLEDGMENT

The authors acknowledge Saint-Petersburg State

University for a research grant 8.38.351.2015 and the the Government of the Russian Federation for grant 074-U01.

#### REFERENCES

- [1] Prabowo R, Thelwall M. "Sentiment analysis: A combined approach." in *Journal of Informetrics* 3, no.2, Apr. 2009, pp. 143-157.
- [2] You Q, Luo J, Jin H, Yang J. "Robust image sentiment analysis using progressively trained and domain transferred deep networks" in *arXiv preprint arXiv:1509.06041*, 2015
- [3] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. "Distributed representations of words and phrases and their compositionality" in *Advances in neural information processing systems*, 2013. pp. 3111-3119.
- [4] Mikolov T, Chen K, Corrado G, Dean J. "Efficient estimation of word representations in vector space." in *arXiv preprint arXiv:1301.3781*, 2013.
- [5] GitHub Repository with implementation of AFINN dictionary, Web: <https://github.com/fnielsen/afinn>.
- [6] Official website of Keras deep learning library, Web: <https://keras.io>.
- [7] Weston J, Chopra S, Adams K. "#TagSpace: Semantic embeddings from hashtags" in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1822-1827.