Image Captioning with Sentiment Terms via Weakly-Supervised Sentiment Dataset

Andrew Shin
andrew@mi.t.u-tokyo.ac.jp
Yoshitaka Ushiku
ushiku@mi.t.u-tokyo.ac.jp
Tatsuya Harada
harada@mi.t.u-tokyo.ac.jp

Graduate School of
Information Science and Technology,
The University of Tokyo
Tokyo, Japan

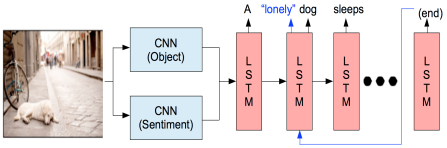

Figure 1: Overall workflow of our model

Image captioning task has become a highly competitive research area with application of convolutional and recurrent neural networks, especially with the advent of long short-term memory (LSTM) architecture. However, its primary focus has been a factual description of the images, mostly objects and their actions. While such focus has demonstrated competence, describing the images with non-factual elements, namely sentiments of the images expressed via adjectives, has mostly been neglected. We attempt to address this issue by fine-tuning an additional convolutional neural network solely devoted to sentiments, where dataset on sentiment is built from a data-driven, multi-label approach.

Building a dataset on sentiments accompanies a number of challenges. First, because sentiments are subjective by nature, it is difficult to label the images in a reliable way. We handle this problem by treating the images as having multiple labels. We utilize *Binary Relevance* [1] in which m training examples x_i whose associated labels form a set Y are viewed as following:

$$D_j = \{(x_i, \phi(Y_i, y_j)) | 1 \leq i \leq m\}$$

$$\text{where } \phi(Y_i, y_j) = \begin{cases} 1, & \text{if } y_j \in Y_i \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Then, the set of labels for unseen example is determined by the obtained binary classifiers g_j for q classes:

$$Y = \{y_j | g_j(x) > 0, 1 \leq j \leq q\} \quad (2)$$

Second issue is the financial cost of building such dataset. We remark that the viewers' comments associated with the images frequently reflect the dominant sentiments of the images, and exploit them with natural language processing techniques and SentiWordNet [2] to automatically label the images, building a large, weakly-supervised sentiment dataset at zero cost.

Model	Sentiment Term
ImNet+	Young man
Bigram	Poor man
Style	Surgical shirt
Ours	Confused man

Figure 2: Example of generated modifying terms from each model.

We fine-tune a separate convolutional neural network on our sentiment dataset. Roughly inspired by the mechanism in which two hemispheres of human brain perform separate functions of logical and emotional perception, we juxtapose two separate convolutional neural networks, for object and sentiment classification, respectively. We train the obtained representation using two networks with captions, and compare the results with various baseline models. Since automatic evaluation metrics are not designed to handle sentiment terms, we mainly resort to human evaluation as our primary metric. Although ground truth captions contain only a limited amount of sentiment terms, the results demonstrate that our features were able to learn better mapping between the images and sentiment terms than baseline models.

[1] M. Boutell, J. Luo, X. Shen, and C. Brown. *Learning Multi-label Scene Classification*. *Pattern Recognition*, Vol. 37, pp.1757-1771, 2004.

[2] S. Baccianella, A. Esuli, F. Sebastiani. *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*. In *Language Resources and Evaluation*, 2010.