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

Abstract

Image captioning task has become a highly competitive research area with successful 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, including the objects, movements, and their relations. While such focus has demonstrated competence, describing the images along 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. Our experimental results show that our method can generate image captions with sentiment terms that are more compatible with the images than solely relying on features devoted to object classification, while capable of preserving the semantics.

1 Introduction

Image captioning task bridges the gap between two of the most fundamental artificial intelligence domains, namely language and vision. Recent surge of deep learning approaches has escalated the task to an unprecedented stage, where generated captions can nearly rival those by humans [8][11][12][17][27][28]. However, the objective of image captioning task has revolved around the factual description of the images, such as the objects, their motions, and their relations. On the contrary, non-factual components subject to viewers' interpretation of the images, mostly appearing in a form of adjective or adverb, have been missing. We define such subjective elements as the *sentiment* of the image, and modifying terms describing it as *sentiment terms*. Such non-factual sentiment terms broaden the expressibility, enrich the aesthetics of the language, and are more human-like.

The reason that research on image captioning with sentiment terms has stagnated is partly due to lack of dataset specialized in sentiments, and the difficulty of building such dataset, which inevitably poses several conundrums. First, there is no clear boundary between classes. An image labeled as 'happy' may also be labeled as 'cute,' 'beautiful,' etc., and the same holds true in the opposite sentiment polarity. One way to deal with this issue may be to have a highly limited number of inclusive classes, as is often done in facial

Pages 53.1-53.12
DOI: https://dx.doi.org/10.5244/C.30.53

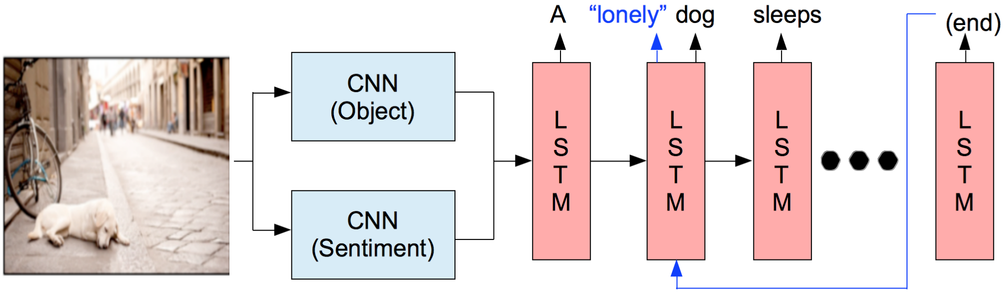


Figure 1: Overall workflow of our model

expression classification task [2]. While this has an advantage that the distinction between classes is comparatively clear, it is at the cost of losing the subtle nuances apparent within the inclusive classes. For example, a non-negligible discrepancy lies in between 'hilarious' and 'peaceful,' both of which belong to the inclusive positive sentiment polarity. Also, it is difficult to port such limited number of classes to the images of a broader domain, in which the range of possible subject matters is extremely wide and humans are frequently not present. Unlike facial expressions, sentiments from the images of general domain can be interpreted with a great variety, often accompanying disagreements among the viewers. Furthermore, certain images may permit labels from opposite polarities to be attached (*e.g.*, 'friendly' and 'eerie' for a smiling pierrot). In fact, the results from our human evaluation in Section 4.2 testify that humans indeed find it very difficult to agree on a single label for given images, even when the number of classes is relatively few. We thus conclude that the sentiments should be represented with multiple labels, as there is no single 'correct' label, but only an indefinite set of acceptable, appropriate labels.

Another practical issue has to do with the financial cost of building such dataset. If we were to rely on crowd sourcing services to have 1 million images manually labeled, as was the case for ImageNet [9], the cost would easily skyrocket up to tens of thousands of dollars. Even so, due to the subjective nature of sentiments, it is not guaranteed that the results will be reliable. As an alternative to manual labelling, we note that the viewers' comments towards the images on social network frequently reflect the sentiments of the images. We exploit this characteristic of the comments in order to inexpensively label the images. As we will see in Section 4.1, it requires attentive filtering processes and is only weakly supervised, but is capable of building a fairly agreeable dataset at virtually zero financial cost.

In this paper, we tackle a novel problem of image captioning with sentiment terms. We build sentiment dataset in a data-driven, multi-label setting, from which an additional convolutional neural network (CNN) learns sentiment features. Since our work is fundamentally an incremental work built on top of conventional image captioning task, we generally follow the approach of CNN-LSTM pipeline for the most part, except features for object classification and sentiment classification are obtained separately, and the LSTM unit with highest probability is revisited after sentence is complete, in order to produce the sentiment term. Figure 1 presents a diagram of the workflow of our model. Throughout the paper, we refer to 'sentiment terms' as the words whose positive or negative score on SentiWordNet [1] is 0.5 or higher.

Our main contributions can be summarized as following: 1) proposal of a novel task of image captioning with sentiment terms, 2) utilization of multi-label learning to deal with subjective nature of sentiments, and 3) introduction of a data-driven approach to inexpensively build a dataset on sentiments and its public release.

2 Related Work

Traditionally, sentiment classification of images has been carried out mostly with hand-crafted features. For example, Siersdorfer et al. [22] suggested that SIFT combined with global color histogram can be a good indicator of the sentiments of the images, although dealing only with positive/negative binary classification. Borth et al. [4] represented images with adjective-noun pairs collected from web mining and analyzing tags associated with the images. More recently, Katayev et al. [16] demonstrated that neural networks can be fine-tuned to distinguish between different styles and atmospheres, and that it outperforms other hand-crafted features, such as GIST or color histogram. This led to an idea that we may also be able to fine-tune neural networks to determine the appropriate sentiment of a given image.

A majority of recent work on image captioning task have been dominated by the usage of convolutional and recurrent neural networks for feature extraction and caption generation respectively, although with substantial variations. Karpathy et al. [17] exploited multimodal RNN to generate descriptions of image regions, aided by the alignment model of CNN over image regions and bidirectional RNN over sentences, which are intermingled via a multi-modal embedding. Inspired by statistical machine translation, Vinyals et al. [27] built a model in which the encoder RNN for source sentences is replaced by CNN features of images. Long short-term memory (LSTM) was employed as a generative RNN of non-linear function. Xu et al. [28] took a similar workflow, but introduced attention-based model, which learns to update the saliency while generating corresponding words. Donahue et al. [11] expanded the CNN-LSTM architecture to activity recognition and video recognition by building long-term recurrent convolutional networks (LRCNs). Time-varying inputs are processed by CNN whose outputs are fed to a stack of LSTMs.

However, these works have mostly overlooked the inclusion of sentiment terms in their captions. In this regard, most intimate to the nature of our work is by Mathews et al [20]. They proposed a switching RNN model, consisting of two parallel RNNs for factual and sentiment description respectively. However, they built separate models for positive and negative terms and applied it to the same set of images under premise that any image can be interpreted from either sentiment polarity. While it is true for certain images as was discussed in Section 1, there are a substantial amount of images that hardly permit an interpretation from both sentiment polarities (for example, it is rare to see a description with a negative term given a close-up of a toddler's smiling face or a blooming flower). We thus believe that the polarity of the sentiment terms in the description should be determined automatically, unaided by manual choice of polarity. It consequently follows that a single RNN suffices for us, although we still need two CNNs for separate feature extractions.

3 Model

3.1 Multi-label Learning

Note that, although we utilize multi-label setting, our objective deviates from that of traditional multi-label learning in that we do not necessarily aim to predict identical set of labels as ground truths, as there exists no definitive set of labels. In fact, prediction of a single appropriate label suffices since there is usually only one modifying term for an object at a time. Thus, multi-label setting in our case is for representing the images and projecting them in a sensible space, rather than replicating identical set of labels.

Multi-label classification itself is an active research area with a variety of approaches. The bottom-line for us is that the approach should be implementable with ease in standard deep learning frameworks, Caffe [15] in our case. One possibility is to utilize the approach known as *Binary Relevance* [5][30] which decomposes the multi-label learning into a set of independent binary classification problems. Thus, m training examples x_i whose associated labels form a set Y are viewed as following:

$$D_j = \{(x_i, \phi(Y_i, y_j)) | 1 \leq i \leq m\}$$
$$\text{where } \phi(Y_i, y_j) = \begin{cases} 1, & \text{if } y_j \in Y_i \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

In our case, x_i corresponds to CNN features, extracted from 2nd fully-connected layer (fc7) of VGG [23]. Then, the set of labels for unseen example is determined by the obtained binary classifiers g_j for q classes:

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

While simplistic, it has proven to generalize well in various domains, and has become a foundation for more sophisticated multi-label learning techniques [30]. Another feasible approach is *Random k-Labelsets* [25], in which every unique set of labels is considered a distinct class. It has two obvious downsides that the number of classes exponentially grows, and that there may be classes in test set that are unseen in training set. We thus opt to proceed with the mechanism of binary relevance.

While multi-label setting can be implemented with slice layers in deep learning frameworks, its setup can be highly tricky. A much simpler method that essentially performs the same task is to simply duplicate the images and assign them different labels. The benefit is its simplicity, while the downside is that the size of dataset grows, which in our case approximately doubled. Note that, since the predicted label for a given image will always be the same, we limit the images in the test set to those containing only one label. Otherwise, the accuracy will never be able to go beyond 100/(average number of labels)% at best.

3.2 Caption Generation

Encouraged by its recent successes in image captioning task [11][17][20][27][28], we employ LSTM [13] as our caption generator, and follow its conventional setting for the most part. The input to LSTM are the features extracted from the second fully-connected layer (fc7) of CNN, although our model necessitates additional CNN features as will be discussed in Section 5. Word vectors are trained with random initialization, and sigmoid function is used for non-linearity throughout all gates except along with hyperbolic tangent for memory cell update as follows:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\ g_t &= tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \end{aligned} \tag{3}$$

The unique part of our LSTM is that we force it to contain at least one sentiment term in its prediction. We hypothesize that sentiment terms are most likely to modify the nouns most characteristic of the sentence, hence most characteristic of the image. We thus keep track of the probability every time a noun is predicted, and once the prediction of the entire sentence is complete, return to the LSTM unit which predicted the noun with highest probability, and feed it again with the output from the previous LSTM unit. We keep a separate vocabulary V_{senti} consisting of sentiment terms only, which is a subset of V consisting of all terms, and predict a word again at the LSTM unit we return to, but this time only from V_{senti}. Thus, we are essentially forcing an insertion of a modifying sentiment term that may have been skipped in favour of the characteristic noun due to smaller likelihood. In summary, sentiment term w_{senti} is the term in V_{senti} which maximizes

$$p_{t_{\text{return}}+1}(w_{\text{senti}}) \; = \text{LSTM}(x_{t_{\text{return}}})(w_{\text{senti}}), \tag{4}$$

where $x_{\text{t}_{\text{return}}}$ is the input at $t = t_{\text{return}}$ determined by the learned parameters and word vectors up to that state. Also, originally predicted word $w_{\text{t}_{\text{return}}+1}$ at this state, which is part of the generated caption of length N, satisfies the following:

$$\begin{aligned} w_{t_{\text{return}}+1} &= \arg\max_{w} p_{t_{\text{return}}+1}(w) \;, w_{t_{\text{return}}+1} \in V_{\text{noun}}, \\ t_{\text{return}} &= \arg\max_{t} \max \text{LSTM}(x_t)(w) \;, 0 \leq t \leq N-1 \end{aligned} \tag{5}$$

4 Dataset

4.1 Construction

We first collected 2.5M images and 28M comments associated with those images from image hosting services, namely Flickr and DeviantArt. Although comments are of different nature from captions, they have been reported to be highly indicative of the sentiment of the images [7], and thus fit our purpose of representing visual sentiments. Sentiment terms that frequently appear on ground truth descriptions of existing standard datasets were manually chosen as queries to retrieve the images. From the collected comments, we count the occurrences of sentiment terms, accompanied by a series of filtering processes as following:

- negation: sentiment terms that are negated are filtered out (*e.g.*, “not very funny”)
- spam: suspicious comments are ignored, and comments with URL are also ignored, regardless of the contents
- color and motion terms: sentiment terms describing specific colors are filtered out. Also, sentiment terms describing motions in the appearance of a gerund (*e.g.*, “jumping”) are filtered out with a few exceptions (*e.g.*,“smiling”).
- first-person subject: sentiment terms used to modify the first-person subject are filtered out (*e.g.*, “I’m serious”)
- inflection: adverbs and comparative forms of adjectives are inflected to their respective original adjective forms (*e.g.*, “happily” or “happier” to “happy”) except they are filtered when followed by an adjective (*e.g.*, “simply” as in “simply beautiful”)
- dual part-of-speech: sentiment terms that have high frequency as a different part-of-speech and require more sophisticated usage of parser are filtered (*e.g.*, “mean”,“pretty”)
- general, non-visual terms: sentiment terms with unclear description criteria that provide no visual clue are manually filtered out from the final counts (*e.g.*, “good,”“bad”)

After filtering and counting of the sentiment terms, we need to determine the appropriate number of classes. We experimented with three different number of classes (20, 50, and 100) determined by the frequency of terms in the comments. According to the number of classes, images without any comment that contains at least one label from the classes are filtered out, and most frequent labels up to maximum of five that appear in the comments for each image

Table 1: Top-1 accuracy of classification by various models. Apart from human evaluation carried out on 1,000 sampled images, all other tests are performed on the entire dataset.

Dataset	Class	Size	SIFT+RGB [22]	VGG [23]	Human
ImageNet [9]	1000	1M	-	71	85
FlickrStyle [16]	20	80k	-	40.7	75.1
Sentiment	100	1.1M	6.8	11.3	16.1
	50	.7M	14.2	20.5	25.8
	20	.5M	19.4	28.7	40.1

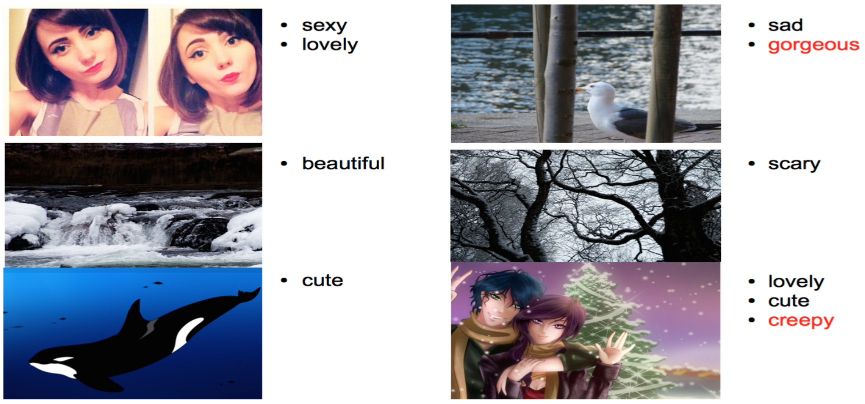


Figure 2: Examples of comment-generated labels. Labels in red color indicate the labels agreed to be inappropriate on Mechanical Turk.

and exist in the classes are selected as the labels for the image. Sizes of the resulting datasets and their performances on various evaluation methods are summarized in Table 1. We refer to this dataset as *Sentiment Dataset* in the rest of the paper.[1]

4.2 Validation

In order to approximate the reliability of the comment-generated labels, we performed a human validation over a subset of our dataset, consisting of 10k images and 19,975 labels. Two workers were assigned per image, and each worker was asked to select the labels that do not seem appropriate given the image. We marked the labels inappropriate if two workers agreed that the label is inappropriate, and 572 labels were agreed to be inappropriate, which amount to 2.9% of the tested labels. While not entirely satisfactory, this yields a fair bound for the reliability of a comment-generated label. Frequent sources of biases were viewers commenting on the overall quality of the images rather than sentiments, or compliments to the uploaders. More refined filtering process considering these biases will further enhance the reliability. Figure 2 shows examples of comment-generated labels and the labels that turned out to be inappropriate.

In order to comprehend the proximity of classification models' performances to human capability, we performed classifications by humans for each number of classes on Mechanical Turk with a subset of dataset, consisting of 1,000 images respectively. Table 1 shows the performances of neural networks and human classification on each dataset with varying number of classes. It is noteworthy that it is difficult even for humans to achieve high accuracy due to the unique nature of sentiments in which subjectivity prevails. However, since we are assuming a single label per image in the test instead of multi-label, the actual accuracy is

[1] http://www.mi.t.u-tokyo.ac.jp/static/projects/sentidata/

Table 2: Sentiment score and number of images for each class.

Class	POS	NEG	OBJ	Images	Class	POS	NEG	OBJ	Images
angry	0	.875	.125	25,824	lovely	.625	0	.375	123,004
beautiful	.750	0	.250	254,905	sad	.125	.750	.125	75,263
crazy	.625	.500	-	37,810	scary	0	.750	.250	30,773
creepy	0	.875	.125	28,830	sexy	.625	0	.375	72,186
cute	.625	0	.375	325,606	simple	.875	.500	-	46,874
dirty	0	.750	.250	16,417	stunning	.750	.625	-	24,049
funny	.500	.500	-	85,590	ugly	0	.750	.250	21,840
gorgeous	.750	0	.250	71,712	unique	.500	0	.500	24,981
handsome	.625	0	.375	28,404	weird	0	.250	.750	51,072
hot	.625	0	.375	48,486	young	.625	.250	.125	39,612

supposedly higher. Table 2 shows the final 20 classes, their sentiment scores as defined on SentiWordNet [1], and the number of images in the dataset belonging to the class.

Some of the traditional hand-crafted features have been known to correlate well in sentiment classification. We applied SIFT and RGB color histogram features followed by a linear SVM in a similar manner as [22] to our dataset, and compared the performances. The results were not as competent as in binary classification, again confirming the complication of multi-class sentiment classification of images and that hand-crafted features may not be adequate for more elaborate classification tasks.

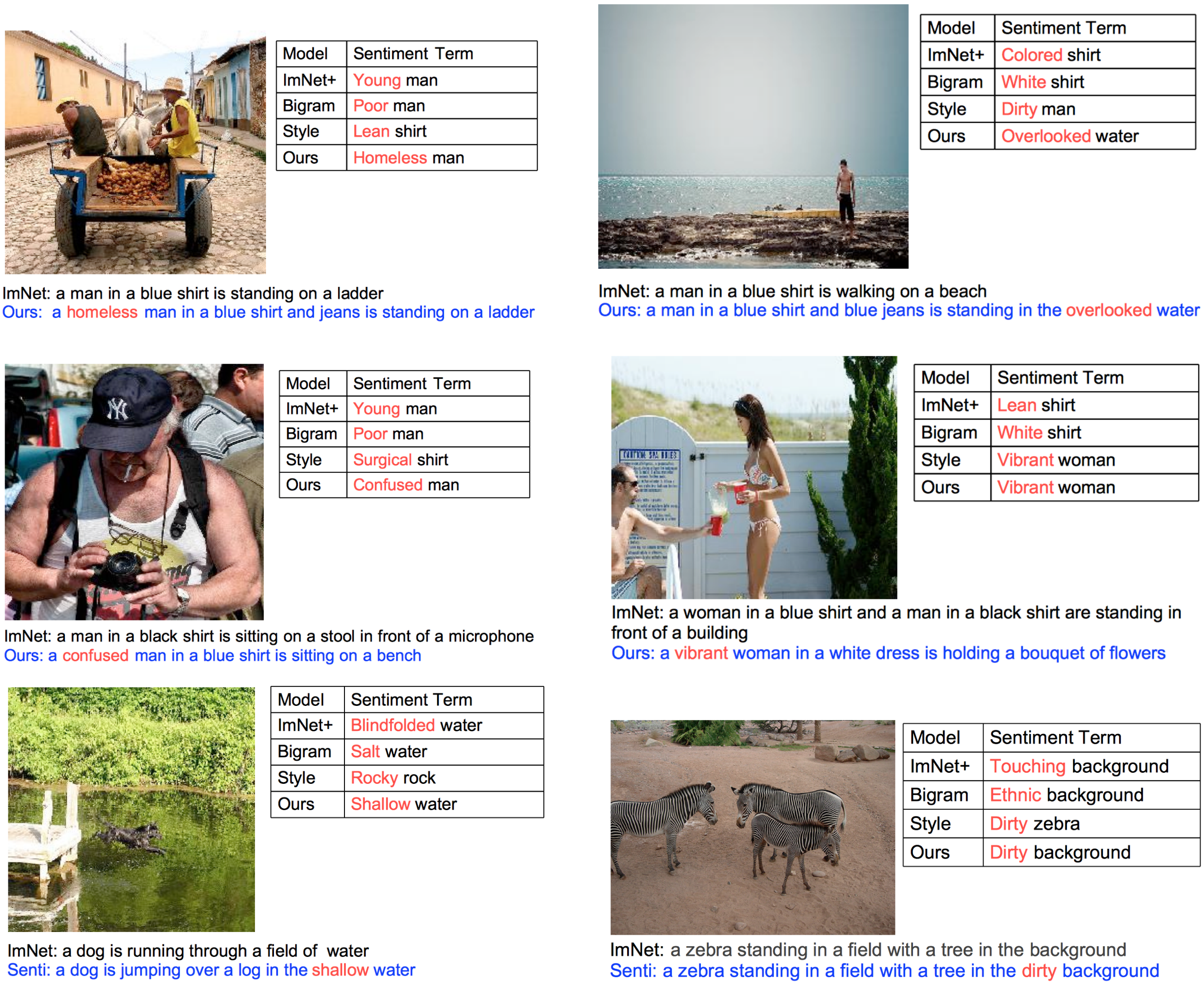
5 Experiments

5.1 Setting

We chose VGG with 19 layers [23] as our network model. Presumably, the characteristics of our sentiment dataset substantially deviate from datasets devoted to object classification task, and we thus aim to adjust the network parameters slightly more aggressively. We fine-tune the layers from the first fully-connected layer (fc6) and on, as opposed to the conventional approach in which only the last fully-connected layer (fc8) is fine-tuned. The initial hyper-parameter setting for fine-tuning is as follows; gaussian weights, initial learning rate of 0.001, step decay of 0.1 at every 20k iterations, etc. Features are extracted from the images via fine-tuned network above using Caffe framework [15].

We compare the performance of our proposed method with those of four baselines. Note that sentiment terms are force-inserted in all models except for the first baseline. :

- ImageNet: ImageNet features with conventional LSTM caption generation without sentiment term force-inserted
- ImageNet+: ImageNet features with conventional LSTM caption generation with sentiment terms force-inserted by LSTM
- Bigram: ImageNet features with sentiment terms chosen by an external bigram corpus, namely Google Web Trillion Word Corpus [6]. No additional features were added, but most likely sentiment term was inserted according to the external bigram corpus, regardless of the probability distribution determined by LSTM, and thus regardless of the features from corresponding images.
- Flickr Style: ImageNet features combined with features from Flickr Style dataset [16] that contains forced sentiment terms from LSTM in the same way as our model. Caffe



Model	Sentiment Term
ImNet+	Young man
Bigram	Poor man
Style	Lean shirt
Ours	Homeless man

ImNet: a man in a blue shirt is standing on a ladder
Ours: a homeless man in a blue shirt and jeans is standing on a ladder

Model	Sentiment Term
ImNet+	Colored shirt
Bigram	White shirt
Style	Dirty man
Ours	Overlooked water

ImNet: a man in a blue shirt is walking on a beach
Ours: a man in a blue shirt and blue jeans is standing in the overlooked water

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

ImNet: a man in a black shirt is sitting on a stool in front of a microphone
Ours: a confused man in a blue shirt is sitting on a bench

Model	Sentiment Term
ImNet+	Lean shirt
Bigram	White shirt
Style	Vibrant woman
Ours	Vibrant woman

ImNet: a woman in a blue shirt and a man in a black shirt are standing in front of a building
Ours: a vibrant woman in a white dress is holding a bouquet of flowers

Model	Sentiment Term
ImNet+	Blindfolded water
Bigram	Salt water
Style	Rocky rock
Ours	Shallow water

ImNet: a dog is running through a field of water
Senti: a dog is jumping over a log in the shallow water

Model	Sentiment Term
ImNet+	Touching background
Bigram	Ethnic background
Style	Dirty zebra
Ours	Dirty background

ImNet: a zebra standing in a field with a tree in the background
Senti: a zebra standing in a field with a tree in the dirty background

Figure 3: Examples of captions with and without sentiment term. Sentiment words generated by each model are shown in the table. See Supplementary Material for more examples.

provides the CaffeNet model fine-tuned on Flickr Style dataset, which achieves about 39.2% accuracy on its own test data. Using VGG 19-layers and fine-tuning from fc6 as we did in our model slightly boosts up the accuracy to 40.7%, and we refer to Flickr Style features as those extracted by this network.

Experiments are carried out on three standard datasets for image captioning task; Flickr 8k [14], Flickr 30k [29], and Microsoft COCO [19]. In each dataset, we build vocabulary V consisting of the words that appear twice or more in the ground truth captions. Vocabulary of sentiment terms V_{senti} are also built in the same way for each dataset.

5.2 Evaluation & Discussion

Figure 3 shows some of the figures and captions with and without sentiment terms, along with all sentiment terms generated by the models. For ImageNet+ and bigram models, sentiment terms are frequently distant from dominant sentiment of the image. Bigram model results in terms that are commonplace, *e.g.*, ‘hot dog’,‘punk rock,‘smart phone,’ yet frequently irrelevant to the image.

Table 3 shows the performances of our model and baselines on a number of automatic evaluation metrics. First, note that BLEU scores [21] are seemingly impaired for all models in which additional terms are inserted. This is inevitable since there are a plethora of ground truth captions that do not contain any sentiment term, and insertion of sentiment

Table 3: Performances of the captions generated by each model on MS COCO determined by automatic evaluation metrics. Note that no additional features were added in first three models. See Supplementary Material for performances on other datasets.

Dataset	**Model**	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**	**METEOR [10]**	**Cider [26]**
MSCOCO	ImNet	62.0	42.4	28.0	18.7	12.1	62.3
	ImNet+	56.6	36.2	21.9	13.1	11.8	44.1
	Bigram	56.7	36.2	22.0	13.3	11.7	44.1
	Style [16]	55.5	34.8	20.7	12.4	11.3	38.2
	Ours	56.5	35.9	21.7	13.0	11.6	43.0

Table 4: Performances of each model on human evaluation

Dataset	**% Appropriateness**	**Avg. Rank**	**Agreement**
ImNet+	.404	2.17	.209
Bigram	.383	2.66	.260
Style [16]	.367	2.98	.301
Sentiment	**.448**	2.25	**.307**

terms will inevitably lower the overall resemblance to those captions, especially as the size of n-gram grows. Since the captions from ImageNet+ and bigram models are exact replica of the original ImageNet model except for the sentiment term, their BLEU scores decrease less than two other models. Since new features are added in the remaining two models including our proposed model, their scores deviate slightly more from ImageNet model, but our model's scores are comparable to those of two fore-mentioned models, especially as the size of dataset grows, and consistently outperform the scores by Flickr Style model.

In order to compensate for limitations of evaluation metrics for dealing with sentiment terms, we also resort to human evaluation, and interpret it as a complementary criterion of evaluation. We performed two types of human evaluation tasks on Mechanical Turk. In the first task, workers were given an image and one of the captions from four models with sentiment terms, and were asked to determine whether the sentiment term is appropriate. In the second task, workers were given an image and all four captions with sentiment terms, and were asked to rank the captions in consideration of both semantic accuracy and appropriateness of the sentiment terms. Two workers were assigned per image in the second task. In both tasks, the same set of 2,000 images from MS COCO was used.

Our model was able to receive the highest appropriateness rating in the first task, which demonstrates that our model was more frequently able to capture the dominant sentiment in the image and generate appropriate terms. In other words, newly added features in our model were more compatible with sentiment terms in the ground truth captions, and the prevalent sentiment of the images. On the other hand, our model was below ImageNet+ model in the ranking task, although only by a close margin. A possible cause is that some of the sentiment terms in ImageNet+ model's captions were considered *compatible* with the image, even when it is not a dominant sentiment in the image (*e.g*. "lean shirt,""young man"). It may also be a reason for its agreement score being lower. Table 4 summarizes the results from human evaluation. Inter-rater agreement was calculated based on [3]. All agreements fall into a range of 'moderate' agreement according to [18].

6 Conclusion

We tackled a novel problem of image captioning with sentiment terms. We introduced a method to inexpensively build a dataset on sentiments, trained in a form of multi-label learning, and exploited the learned features on long short-term memory to generate image captions with sentiment terms. It was comparable on automatic evaluation metrics to conventional models, and human evaluators found the captions from our model to be more appropriate with regards to the sentiment of the image.

Acknowledgement

This work was funded by ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan).

References

[1] 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.

[2] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. *Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior*. In *CVPR*, 2005.

[3] E. Bennett, R. Alpert, and A. Goldstien. *Communications Through Limited-Response Questioning*. *Public Opinion Quarterly*, Vol.18, pp.303-308, 1954.

[4] D. Borth, R. Ji, T. Chen, T. Breuel, and S. Chang. *Large-Scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs*. In *ACM Multimedia*, 2013.

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

[6] T. Brants, and A. Franz. *Web 1T 5-gram Version 1*. *Philadelphia: Linguistic Data Consortium*, 2006.

[7] Y. Chen, T. Chen, W. Hsu, H. Liao, and S. Chang. *Predicting viewer affective comments based on image content in social media*. In *ICMR*, 2014.

[8] X. Chen and C. Zitnick. *Mind's Eye: A Recurrent Visual Representation for Image Caption Generation*. In *CVPR*, 2015.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. *ImageNet: A Large-Scale Hierarchical Image Database*. In *CVPR*, 2009.

[10] M. Denkowski and A. Lavie. *Meteor Universal: Language Specific Translation Evaluation for Any Target Language*. In *EACL Workshop on Statistical Machine Translation*, 2014.

[11] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. *Long-term Recurrent Convolutional Networks for Visual Recognition and Description*. In *CVPR*, 2015.

[12] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. Platt, C. Zitnick, and G. Zweig. *From Captions to Visual Concepts and Back*. In *CVPR*, 2015.

[13] S. Hochreiter and J. Schmidhuber. *Long Short-Term Memory*. *Neural Computation*, Vol. 9, pp.1735-1780, 1997.

[14] M. Hodosh, P. Young, and J. Hockenmaier. *Framing Image Description as a Ranking Task: Data, Models, and Evaluation Metrics*. *Journal of Artificial Intelligence Research*, Vol.47, pp.853-899, 2013.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. *Caffe: Convolutional Architecture for Fast Feature Embedding*. In *ACM Multimedia*, 2014.

[16] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller. *Recognizing Image Style*. In *BMVC*, 2014.

[17] A. Karpathy and L. Fei-Fei. *Deep Visual-Semantic Alignments for Generating Image Descriptions*. In *CVPR*, 2015.

[18] J. Landis and G. Koch. *The Measurement of Observer Agreement for Categorical Data*. In *Biometrics*, Vol.33, pp. 159-174, 1977.

[19] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. Lawrence Zitnick. *Microsoft COCO: Common Objects in Context*. In *ECCV*, 2014.

[20] A. Mathews, L. Xie, and X. He. *SentiCap: Generating Image Descriptions with Sentiments*. In *AAAI*, 2016.

[21] K. Papineni, S. Roukos, T. Ward, and W. Zhu. *BLEU: A Method for Automatic Evaluation of Machine Translation*. In *ACL*, 2002.

[22] S. Siersdorfer, E. Minack, F. Deng, and J. Hare. *Analyzing and Predicting Sentiment of Images on the Social Web*. In *ACM Multimedia*, 2010.

[23] K. Simonyan and A. Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. In *ICLR*, 2015.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. *Going Deeper with Convolutions*. In *ILSVRC*, 2014.

[25] G. Tsoumakas, I. Katakis, and I. Vlashavas. *Random k-labelsets for Multi-label Classification*. *IEEE Transactions on Knowledge and Data Engineering*, Vol.23, pp.1079-1089, 2015.

[26] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. *CIDEr: Consensus-based Image Description Evaluation*. In *CVPR*, 2015.

[27] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan *Show and Tell: A Neural Image Caption Generator*. In *CVPR*, 2015.

[28] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*. In *ICML*, 2015.

[29] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. *From Image Description to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions*. *Transactions of the Association for Computational Linguistics*, Vol.2, pp.67-78, 2014.

[30] M. Zhang and Z. Zhou *A Review on Multi-Label Learning Algorithms*. In *ICLR*, 2015.