REDUCING NOISY LABELS IN WEAKLY LABELED DATA FOR VISUAL SENTIMENT ANALYSIS

Lifang Wu[1] *Shuang Liu*[1] *Meng Jian*[1] *Jiebo Luo*[2] *Xiuzhen Zhang*[3] *and Mingchao Qi*[1]

1) School of Information and Communication Engineering, Beijing University of Technology, Beijing, China, 100124
2) Department of Computer Science, University of Rochester Rochester, NY, USA 14623
3) Department of Computer Science and IT, RMIT University, Melbourne, 3000, Australia

ABSTRACT

Deep learning-based visual sentiment analysis requires a large dataset for training. Dataset from social networks is popular but noisy because some images collected in this manner are mislabeled. Therefore, it is necessary to refine the dataset. Based on observations to such datasets, we propose a refinement algorithm based on the sentiments of adjective-noun pairs (ANPs) and tags. We first determine the unreliably labeled images through the sentiment contradiction between the ANPs and tags. These images are removed if the numbers of tags with positive and negative sentiments are equal. The remaining images are labeled again based on the majority vote of the tags' sentiments. Furthermore, we improve the traditional deep learning model by combining the softmax and Euclidean loss functions. Additionally, the improved model is trained using the refined dataset. Experiments demonstrate that both the dataset refinement algorithm and improved deep learning model are beneficial. The proposed algorithms outperform the benchmark results.

Index Terms—Visual sentiment analysis, mislabeled images, deep learning, sentiment conflict

1. INTRODUCTION

With the development of the Internet and smartphones, social networks have become indispensable for people in their daily lives. It is reasonable to estimate the trend of some social events by analyzing the response of people in social networks. People usually express themselves using items that they upload, including text, images, and videos. Therefore, it is important to analyze the sentiment of these items. In recent years, visual presentation has become increasingly popular in social networks. In this paper, we focus on the sentiment analysis of images.

Research has been conducted to predict sentiment using visual features. Early studies used low-level and middle-level features [1-3]; however, the performance of sentiment prediction was not satisfactory. In recent years, deep learning has been used for image sentiment prediction. Xu et al. [4] used a pretrained Convolutional Neural Network (CNN) model to generate features and train classifiers with these features. Chen et al. [5] and Campos et al. [6] generated their model by fine-tuning the pretrained CNN model. You et al. [7] designed a new progressive CNN (PCNN) architecture for sentiment prediction. Experiments on their database, which contained over 1,000 images, demonstrated that their algorithm achieved an accuracy of 74.8%, with state-of-the-art performance.

The sentiment labels of images include both sentiment polarity and strength. Additionally, from the viewpoint of machine learning, the stronger the supervising signal, the more accurate the model. However, existing studies have considered sentiment prediction as a binary classification task, in which only sentiment polarity is used. In this study,

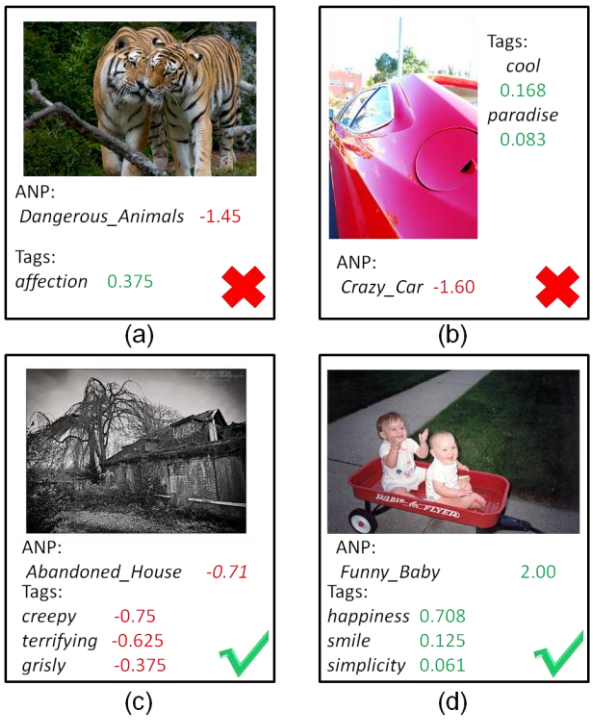

Fig. 1. Example images with ANPs and tags. The labels of the images are obtained by the ANPs. Therefore, (d) is labeled as positive, whereas the other images are labeled negative. Sentiments in (a) and (b) are contradicted with labels. And the sentiments polarity of tag is contradictory with the sentiment label by the ANPs. (c) and (d) have consistent labels. The sentiments from ANP and tags are consistent.

we consider how to use both the sentiment polarity and strength.

Furthermore, deep learning requires a large training set. It is challenging to collect and label a large number of images. One solution is to collect images from social networks. A typical example is SentiBank [2]. The authors obtained the sentiments of ANPs from sentiment lexicons. Then they searched images on Flickr, a visual social network, using ANPs from tags of the images. Additionally, the sentiments of the ANPs were assigned as the labels of the corresponding images.

However, data from social networks is noisy. SentiBank, for example, contains some mislabeled images, whose true sentiments contradict the automatically assigned labels. These images are misleading and possibly result in an ineffective model. Therefore, it is necessary to manage these images. You et al. [7] proposed the PCNN to overcome this problem based on a self-training strategy. They used the entire dataset of SentiBank to train a basic CNN model and selected the images with higher scores. Then the model was fine-tuned using the selected images. The accuracy increased by approximately 2%. Jou et al. [8] proposed a language-specific method to discover ANPs in different languages. More reliable retrievals could be obtained using crowdsourcing and statistics.

By observation, we find that images on Flickr generally include many tags in addition to these in ANPs; that is, ANPs are only a small part of the image tags. Therefore, the analysis is incomplete if we assign sentiment labels to images only considering the sentiment of the ANPs. Furthermore, ANPs and other tags present contradictory sentiment polarities in some images, as shown in Fig. 1 (a) and (b).

Fig. 1(a) is retrieved by the ANP "dangerous animals". It is automatically labeled as a negative sentiment because "dangerous animals" is thought to be negative. In fact, this picture presents the warm affection between two tigers; It presents a definitely positive sentiment. Additionally, from this picture, we can find the tag "affection", which presents a positive sentiment. The same phenomenon can be observed in Fig.1(b). However, in Fig1 (c) and (d), the sentiments of the ANPs and tags are consistent, and they are the same as those presented in the pictures. From these examples, we can observe that some automatically assigned labels using ANPs in Sentibank are misleading and noisy. How could we determine such mislabeled images ?

Consider the image, ANP, and other tags as three parties. The sentiment of the image is unknown. If the sentiment polarities from the ANP and tags are consistent, then this means that two out of three parties agree with the same idea. The tags' sentiments could be considered as enhancements to those of the ANP. The automatically labeled sentiment from the ANP could be more convincing. On the other hand, if the sentiment of the ANP contradicts those of some tags, it means that these two parties present different sentiments. Therefore, the sentiment label of the image from the ANP is not convincing. We further analyze the images that have a sentiment conflict statistically. Additionally, we observed that the voting results of tags' sentiments are more consistent with the sentiments of the corresponding images.

Based on the above observations, we propose a dataset refinement algorithm. We first determine the unreliably labeled images in SentiBank according to the sentiment contradiction of the ANPs and tags. These images are removed from the dataset if the numbers of tags with positive and negative sentiment polarities are identical. The remaining images are labeled again based on voting results on the tags' sentiments. Finally, we obtain a new dataset called refined Sentibank (rSentibank), which includes more images with reliable sentiment labels and fewer images with unreliable labels. Furthermore, we improve the traditional deep learning model by combining softmax and Euclidean loss functions so that more supervisory signals can be introduced. Additionally, the improved model is trained using rSentibank. The experimental results on a public dataset demonstrate that the proposed algorithms outperform the benchmark algorithms.

The contributions of this paper are as follows:

1) We propose an algorithm to denoise the dataset SentiBank using the sentiments of the ANPs and tags of the images. We obtain a refined training set: rSentiBank.

2) We improve the deep learning model by combining the softmax and Euclidean loss functions. The model is effective because more supervisory signals are introduced.

2. REFINE THE SENTIBANK DATASET USING THE SENTIMENT OF THE ANPS AND TAGS

2.1. Observations

SentiBank was proposed by Borth et al. [2]. They used each of the 24 emotions defined in Plutchik's theory to obtain 1,553 ANPs. Then they assigned the sentiment polarity and strength of the ANPs based on SentiWordNet [9] and SentiStrength [10]. Finally, they searched images on Flickr using the ANPs as keywords. The sentiments of the ANPs were automatically used as labels of the corresponding images.

We observe SentiBank and find that each image includes 8.06 tags, on average. Thus, there are more than six tags in addition to the ANP. Furthermore, the sentiment polarities of the tags and ANP are contradictory for some images. Approximately 86,248 images from a total of 452,272 images have such problems according to SentiWordNet. We consider the image, ANP, and other tags as three parties. The sentiment of the image are undetermined. If the sentiment polarities from the ANP and tags are consistent, this means that two out of three parties agree with the same idea. The automatically labeled

sentiments from the ANP can be particularly convincing. These images are retained in the dataset.

We consider what we can do for the images with sentiment conflicts, for example, removing these images from the dataset directly. From our viewpoint, it is possible that these images represent more complicated scenarios. If we can obtain their true sentiment labels, they can be valuable for model training. By further observation, we find that there are approximately 12,089 images for which the numbers of tags with positive and negative sentiments are unequal. Could we obtain the true sentiments of these images by voting of tags' sentiments ?

We randomly select 1,200 samples from the 12,089 images. The user study is implemented by 18 subjects (10 males and eight females). Each subject manually marks the sentiment polarity (positive, negative) of 200 images, which are randomly selected from the 1,200 images. Therefore, Each image are marked by three subjects. From the voting results of the three subjects' marks, we obtain the sentiment polarity of each image, which is treated as the ground truth. By comparing the automatically obtained labels from the voting of tags' sentiments with the ground truth, there are 1,008 images whose sentiment labels are consistent with the ground truth. The labeled sentiments of the remaining 192 images are different from the ground truth.

2.2. Refine the dataset using the sentiments of tags and ANPs

Based on the observations in Section 2.1, we propose an algorithm to refine the dataset using the sentiments of tags and ANPs. The sentiments of tags are obtained from the sentiment lexicon of SentiWordNet [9].

Algorithm 1 : Dataset refinement algorithm

Input :

1) *Sen_ANP*, sentiment of the ANP.
2) Tag_i, $i=1,2..N$, N tags of the image.
3) *Sen_labeled*, automatically labeled sentiment polarity using the ANP. It is assigned 1 for the positive sentiment and 0 for the negative sentiment.

Output:

1) Decision_value

1: The image should be retained in the dataset, and the automatically labeled sentiment could be kept.

0: The image could be retained in the dataset, but the automatically labeled sentiment should be changed.

-1: The image should be removed from the dataset.

2) *Sen_polarity*: final sentiment polarity

3) *Sen_strength*: final sentiment strength

Start

1) IF ($\mathrm{sgn}(\sum Sen\_labeled \oplus Sen(Tag_i)) = 0$)
2) *Decision_value*=1;
3) *Sen_final*=*Sen_labeled*;
4) ELSE
5) Count the tags with positive and negative sentiments, and get *Num_pos* and *Num_neg* , respectively.
6) IF (*Num_pos*=*Num_neg*)
7) *Decision_value*=*-1*;
8) ENDIF
9) IF (*Num_pos*>*Num_neg*) *AND* Sen_ANP=pos) *OR* (*Num_pos*<*Num_neg AND Sen_ANP*=*neg*)
10) *Decision_value=1*;
11) *Sen_polarity*=*Sen_labeled*;
12) *Sen_strength is the sentiment strength of the ANP*;
13) ENDIF
14) IF (*Num_pos*>*Num_neg AND Sen_ANP*=*neg*) OR (*Num_pos*<*Num_neg AND Sen_ANP*=*pos*)
15) *Decision_value=0*;
16) *Sen_polarity*=*NOT(Sen_labeled)*;
17) *Sen_strength is the average of the sentiment strengths of the tags whose sentiment polarities are equal to Sen_polarity*
18) ENDIF
19) ENDIF

End

In the Algorithm 1, $\oplus$, *AND*, *OR*, and *NOT* represent the Boolean XOR, AND, OR, and NOT operations respectively; *Sen(w)* is the sentiment polarity extraction function; and *Sgn(x)* is the sign function. Example images considered by the proposed algorithm are shown in Fig. 2. The images in the top row were removed from the dataset. And the images in the bottom row were retained, but the sentiment labels were changed by voting of tags' sentiments.

Thus, we obtain a refined dataset that includes more images with reliable labels. The dataset is named rSentiBank, which contains 378,113 images.

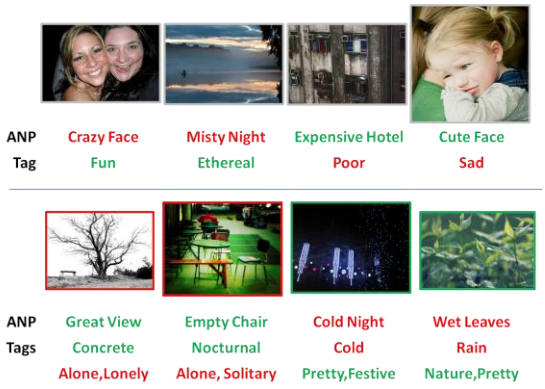

Fig. 2. Example images considered by the dataset refinement algorithm. The images in the first row were removed. The images in the second row were retained, but their sentiment labels were changed by voting of tags' sentiments. The final labels are marked as the bounding box of the images: positive is marked in green, whereas negative is in red.

3. DEEP LEARNING ARCHITECTURE

In this section, we propose the framework ACaffeNet, which is an improved CaffeNet [11]; CaffeNet is the improved edition of AlexNet[12].

Like CaffeNet, ACaffeNet includes five convolutional layers and three fully connected layers. However, ACaffeNet includes two types of loss functions: the softmax loss function represents the variation of sentiment polarity and the Euclidean loss function represents the variation of sentiment strength. The architecture of ACaffeNet is shown in Fig. 3.The loss function is as follows:

$$E = \lambda\left(-\frac{1}{N}\sum_{n=1}^{N}\log(\hat{p}_{nl_n})\right) + (1-\lambda)\left(\frac{1}{2N}\sum_{n=1}^{N}\left\|s_n - s_n\right\|_2^2\right), l_n \in [0,1], \quad (1)$$

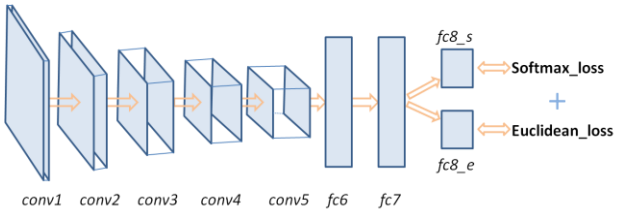

Fig. 3. ACaffeNet; an improved CNN architecture.

where N is the number of images in the training set; λ is the weight to control the trade-off between two types of losses (we set $\lambda = 0.5$ in our experiments); l_n is the binary label of the n^{th} image; $\hat{s}_n$ is the strength prediction of the n^{th} image, whereas s_n is its labeled strength; and $\hat{p}_{nl_n}$ is the prediction probability of the n^{th} image for the l_n class given by

$$\hat{p}_{nk} = e^{x_{nk}} \Big/ \sum_{l=0}^{K-1} e^{x_{nl}} \ . \qquad (2)$$

4. EXPERIMENTAL RESULTS

4.1. Dataset

We test the proposed scheme using the public dataset released by You et al. [7]. The dataset included 1,268 images collected from Twitter. The sentiment of each image was annotated by five users. The images for which the sentiments of the five users agreed formed Subset 5, which included 882 images. Similarly, Subset 4 included 1,116 images for which at least four users agreed on the sentiment. Additionally, Subset 3 included 1,268 images for which at least three users agreed on the sentiment.

4.2. Effect of dataset refinement

To evaluate the effect of the dataset refinement algorithm, we use CaffeNet, which is pretrained on ImageNet and fine-tuned on rSentiBank and SentiBank. The experimental results are shown in Table 1.

Table 1. Prediction accuracy of CaffeNet on SentiBank and rSentiBank

Model	Test Dataset	Training Dataset	
		SentiBank	rSentiBank
CaffeNet	*Subset 3*	*71.08%*	**72.32%**
	Subset 4	*73.75%*	**74.53%**
	Subset 5	*77.89%*	**78.89%**

From Table 1, we observe that the model trained on rSentiBank obtained better performance than that trained on SentiBank. The proposed dataset refinement algorithm reduced the size of the training set, but improved the performance of sentiment prediction.

4.3. Effect of the improved deep learning scheme

We compare the performance of ACaffeNet and CaffeNet. Both models are pretrained on ImageNet and fine-tuned on rSentiBank. The structure of ACaffeNet is shown in Fig. 3. The compared results are shown in Table 2.

From Table 2, we observe that for all three subsets, ACaffeNet with both loss functions outperformed CaffeNet with only the softmax loss function.

Table 2. Prediction accuracy of CaffeNet and ACaffeNet

Training Dataset	Test Dataset	Model	
		CaffeNet	ACaffeNet
rSentiBank	*Subset 3*	72.32%	**72.95%**
	Subset 4	74.53%	**75.96%**
	Subset 5	78.89%	**79.57%**

4.4. Comparison with the state-of-the-art algorithm

We compare the proposed scheme with the state-of-the-art algorithm PCNN [7]. Using the framework of PCNN and replacing CNN with ACaffeNet, we obtained PACaffeNet. The model was first trained on Sentibank and then fine-tuned on rSentibank.

The compared results are shown in Table 3. The proposed AcaffeNet&rSentiBank outperforms PCNN. PACaffeNet has the best results. This demonstrates that the progressive strategy further improved performance. Additionally, using the same PCNN framework, AcaffeNet& rSentiBank is more efficient than the original PCNN.

Table 3. Prediction accuracy of different models

Test Dataset	ACaffeNet & rSentiBank	PCNN[7]	PACaffeNet
Subset 3	72.95%	68.7%	74.21%
Subset 4	75.96%	71.4%	76.95%
Subset 5	79.57%	74.7%	81.45%

5. CONCLUSION

To address the problem of mislabeled training data that is collected from social networks, for deep learning-based visual sentiment analysis, we proposed a pre-processing algorithm to refine the dataset according to the sentiments of ANPs and tags. We further improved the architecture of deep learning by combining the softmax and Euclidean loss functions. The experimental results confirmed the effectiveness of the proposed algorithms.

The dataset refinement algorithms were validated on SentiBank. They could be used to collect a new, larger dataset from social networks. Currently, certain images are removed because their true sentiment labels cannot be obtained automatically. In future work, we will introduce more features to address the problem so that more images can be included.

6. REFERENCES

[1] Siersdorfer S, Minack E, Deng F, et al. Analyzing and predicting sentiment of images on the social web[C] //Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010: 715-718.

[2] Borth D, Ji R, Chen T, et al. Large-scale visual sentiment ontology and detectors using adjective noun pairs[C] //Proceedings of the 21st ACM international conference on Multimedia. ACM, 2013: 223-232.

[3] Yuan J, Mcdonough S, You Q, et al. Sentribute: image sentiment analysis from a mid-level perspective[C] //Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining. ACM, 2013: 10.

[4] Xu C, Cetintas S, Lee K C, et al. Visual sentiment prediction with deep convolutional neural networks[J]. arXiv preprint arXiv:1411.5731, 2014.

[5] Chen T, Borth D, Darrell T, et al. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks[J]. arXiv:1410.8586, 2014.

[6] Campos V, Jou B, Giro-i-Nieto X. From Pixels to Sentiment: Fine-tuning CNNs for Visual Sentiment Prediction[J]. arXiv preprint arXiv:1604.03489, 2016.

[7] You Q, Luo J, Jin H, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks[J]. arXiv preprint arXiv:1509.06041, 2015.

[8] Jou B, Chen T, Pappas N, et al. Visual affect around the world: A large-scale multilingual visual sentiment ontology[C]. Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015: 159-168.

[9] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining[C]//LREC. 2010, 10: 2200-2204.

[10] Thelwall M, Buckley K, Paltoglou G, et al. Sentiment strength detection in short informal text[J]. Journal of the American Society for Information Science and Technology, 2010, 61(12): 2544-2558.

[11] https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet

[12] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C] //Advances in neural information processing systems. 2012: 1097-1105.