CNN for Image Sentiment Analysis with Triplet Loss

Zheng Lu, Yunhe Feng
Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville
Email: {zlu12, yfeng14}@vols.utk.edu

***Abstract*—Understanding the underlying attitude of people towards a given image automatically is important for many applications. The rapid growth of social media provides new opportunities and challenges to design image sentiment inference systems. Due to the weak relation between low-level visual features and sentiment, recent work focus on constructing mid-level semantic representation that can both be easily detected from the input images and be mapped to the sentiment class. However, although current convolutional neural networks are good at identifying objects, it is not as effective when applied for mid-level semantic representations with adjective parts. To overcome this challenge, we adopt a variant of convolutional neural networks with triplet loss to perform end-to-end learning that can automatically generate the most suitable mid-level features. We show the effectiveness of our approaches by detecting both mid-level representations and predicting sentiment of various image datasets and show significant improvement against the baseline approaches.**

I. Introduction

Recent years, social media platforms have seen a rapid growth of user-generated multimedia contents. Billion of images are shared on multiple social media platforms such as Instagram or Twitter which contributes to a large portion of the shared links. Through image sharing, users are usually also express their emotions and sentiments to strengthen the opinion carried in the content. Understanding users' sentiments provide us reliable signals of people's real-world activities which are very helpful in many applications such as predicting movie box-office revenues [2], political voting forecasts [12]. It can also be used as the building block for other tasks such as the image captioning [16].

Automatic sentiment analysis recognizes a person's position, attitude or opinion on an entity with computer technologies [14]. Text-based sentiment analysis has been the main concentration in the past. Only recently, sentiment analysis from online social media images has begun to draw more attentions. To simplify the task, previously, the sentiment analysis mainly focuses on the opinion's polarity, i.e., one's sentiment is classified into categories of positive, neutral and negative. However, as pointed out in many recent studies [1], [3], [5], [19], it faces the unique challenge of large "affective gap" between the low-level features and the high-level sentiments.

Recent work resort to extract manually designed mid-level representations from low-level features for image sentiment analysis tasks, e.g., visual sentiment ontology (VSO) in [3], mid-level attributes in [19]. Such mid-level representations usually include both adjective and noun parts such as "happy dogs", "creepy house" and have strong sentiment. So approaches based on mid-level representations usually outperforms methods that inferring sentiment directly from low-level features. Rapid developments in Convolutional Neural Network (CNN) [6], [11], [13], [15] push the transformations of computer vision tasks. There are also several efforts to apply CNN to image sentiment analysis [4], [5], [18]. However, recent works on applying convolutional networks still borrow the network architectures from the image classification tasks [1], [4], [5], [18]. Although convolutional neural networks are very effective at image classification tasks, they are not as effective when separating adjective parts that widely found in most mid-level representations for image sentiment analysis purpose.

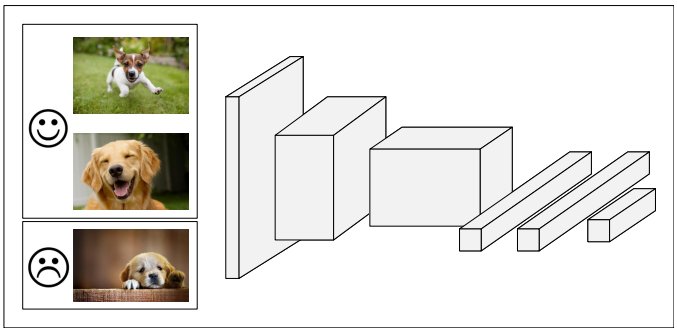

Fig. 1. Image sentiment analysis with triplet loss

To overcome this challenge, we first let the convolutional neural network trained on the noun part, then we adopt the triplet loss to replace surrogate loss used in the convolutional neural networks to force the network to focus more on the adjective part in the mid-level representations. When applying triplet loss, the network loot at three data points at the same time, the anchor point, the positive point and the negative point, where the anchor point and the positive point belongs to the same class and the negative point belongs to a different class. Then during training, the network tries to minimize the distance between anchor point and positive point and maximize the distance between the anchor point and the negative point.

A major challenge for triplet loss is mining triplets, as

trivial triplets are not uninformative to the network and too hard triplets make the network hard to learn "normal" features [7]. When applying triplet loss to mid-level representation classification for image sentiment analysis, we select triplets from the same noun class, with the anchor point and the positive point belongs to the same adjective class while negative point belongs to a different adjective class. In this way, the triplet loss forces the network to learn the adjective part of the mid-level representations and the triplets are informative and moderately hard for the network to learn.

A. Contributions

Our contributions can be summarized as follows:

- We apply a two-stage learning scheme for the network to learn the adjective and noun part of mid-level representations in the image sentiment analysis separately;
- We replace surrogate loss in convolutional neural networks with triplet loss and design a triplet selection approach to force the network learn the adjective part of mid-level representations;
- We perform extensive experiments on several real word datasets to show the effectiveness of our approach in both mid-level representation classification and sentiment prediction.

The rest of this paper is organized as follows. In Section II we show previous works on image sentiment analysis. We explain our system structure and triplet selection scheme in Section III. The evaluations of our proposed method are in Section IV. We conclude our work in Section V.

II. Related Works

The majority of work in image sentiment analysis focus on manually design meaningful mid-level representations for the task for image sentiment prediction. [3] trying to fill in the "affective gap" between the low-level features and the high-level sentiment by a set of mid-level representation called visual sentiment ontology that consists of more than 3,000 Adjective Noun Pairs (ANP) such as "beautiful flower" or "disgusting food". The authors also published a large-scale dataset called SentiBank that is widely used in later works. They also extend their work into a multilingual settings in one of their later work [8]. [19] adopts a similar methodology as [3]. The authors also construct a mid-level representations for better classification except that they choose different scene-based mid-level attributes than [3]. In addition to that, they also include face detection to enhance the performance of images containing human faces. [1] studies the sentiment analysis of images of social events. It designs specific mid-level representations of each event class and classifies sentiment of each image without the help of texts associated with the image. Due to the challenge to manually collecting labeled data for image sentiment analysis, [17] proposes an unsupervised method to facilitate social media images sentiment analysis with textual information associated with each image.

As convolutional neural networks are found to be very effective at image classification tasks, there are many efforts try to apply CNN to the image sentiment analysis task. [5] try to apply CNN based on AlexNet [11] to automatically extract features based on ANPs proposed in [3]. [18] also applies CNN to image sentiment analysis. They propose a method to progressively training CNN by keep training instances with distinct sentiment scores towards sentiment polars and discard training instances otherwise. [4] studies how to fine-tune AlexNet-styled CNN to achieve better performance on image sentiment analysis tasks.

Our work falls into the category of using convolutional networks to solve the image sentiment analysis task. Unlike previous works that mostly focus on modifying the neural network structure to increase the prediction accuracy, we adopt a different loss function, i.e., triplet loss [7], and modifying the training procedure to better detect mid-level representations in the input image.

III. The design

We propose to solve the image sentiment analysis with the convolutional neural network to detecting mid-level representations. We use an AlexNet-styled network and perform a two-stage training scheme where the network first learns the noun part of mid-level representations. Then we rearrange the dataset and train the network with triplet loss to force the network to learn the adjective part.

A. Preliminary

We adopt the Adjective Noun Pairs (ANP) proposed in [3] as the mid-level representations for image sentiment analysis. ANPs are frequently found in practice and reflect strong sentiment. Each ANP contains exactly one adjective and noun word, such as "beautiful flower" or "creepy house".

We use polarized sentiment model that each image can either be classified as positive or negative.

B. Network structure

Here we show the architecture of our convolutional neural networks. The network contains 5 convolutional layers and 2 fully connected layers with a softmax layer as the output layer. There are max-pooling layers at the first, second and last convolutional layers. Batch normalization is performed before activation for each convolutional layers and fully connected layers. ReLU is used as the activation function for all layers. Adam [9] is used as the optimizer of the network. We use cross entropy and triplet loss as the loss function for each training stage respectively. The detailed size of each layer is shown in Figure [**?**]

C. Two stage training

It is widely shown in the literature [6], [11], [13], [15] that convolutional neural networks are good at image classification tasks. In [4], the authors have shown that convolutional neural networks pre-trained on image classification datasets can be quite effective at image sentiment prediction tasks. So in our work, we adopted a two-stage training scheme that let the network learn the noun and adjective part of ANPs gradually.

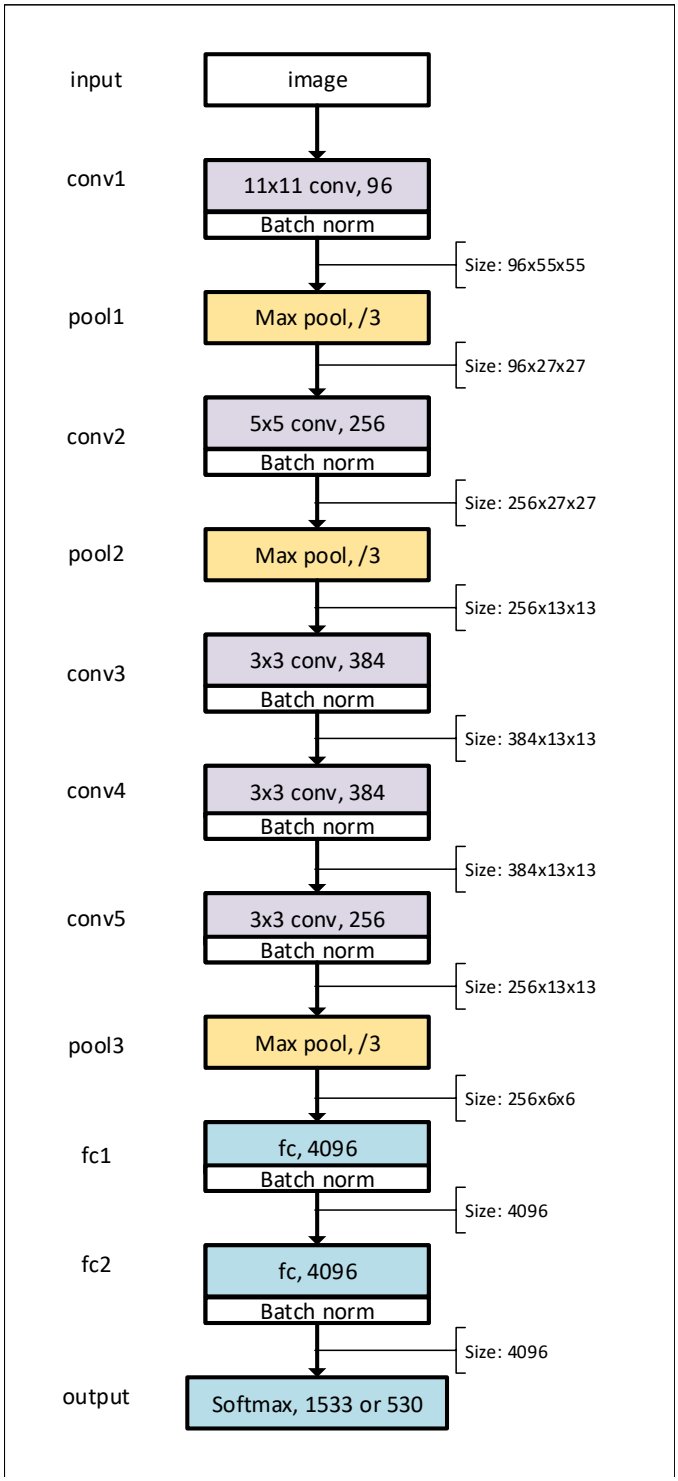

Fig. 2. Network structure

The first stage of training is let the network learn the noun part of ANPs. This stage is the same as the image classification task. We first extract all noun words from all ANPs. Then we set the last softmax layer with the size of the number of noun words and use cross-entropy loss to train the network.

The second stage requires us to modify the network structure and the loss function. We first replace the softmax layer with a new one that has the size of the number of ANPs and initialize it randomly. Then we replace the loss function with triplet loss and train the network again using images from the same dataset. To train triplet, the data need to be arranged in triplets, each triplet consist of an anchor point, a positive point from the same class of the anchor point, and a negative class from a different class than anchor point. The triplet loss is defined below, where D is the distance between different data points and m is the margin to ensure the network does not form trivial models where distances of all pairs of data points are 0.

$$\mathcal{L}_{tri}(\theta) = \sum_{a,p,n} [m + D_{a,p} - D_{a,n}]_+.$$

To training with triplet loss, a crucial step is to mining good triplets as there are n^3 number of possible triplets when the dataset contains n data points. If the selected triplets are too simple, the network will stop learning quickly. However, if the triplets are too hard, the network will become unstable [7]. To ensure the network learn meaningful information from the training, we arrange the dataset in a way that in each triplet, all three image belongs to the same noun class. However, the negative image has the different ground truth ANP with the anchor image and the positive image. For example, we may select both the anchor image and positive image from "beautiful flower", and select a negative image from "dead flower". Ideally, the negative image should have the different sentiment as the anchor point. There are cases where a noun class only has one ANP or has all ANPs with the same sentiment, but even though we can hardly select meaningful triplets to train the network in such a case, the network will not suffer from it when predicting sentiment of the image.

When the network finished training, it can classify images to ANPs. Then with the ANP, we can predict the sentiment of the image. A simple solution would be predicting the sentiment of the image with the ANP with the highest probability of the softmax output layer. However, in many cases, it is more robust to use top-k ANPs with the highest probability of softmax layer. For example, it is very common that there are multiple ANPs in a single image, sometimes even with a different sentiment, or the top ANP prediction is not correct.

IV. Evaluations

In this section, we show the results of experimental evaluation of our convolutional neural networks for both Adjective Noun Pair classification and image sentiment prediction.

A. Datasets

We collect several public available data and list them in table I. The SentiBanck-Flickr dataset is crawled from the flickr website with each ANP as keyword. All the rest datasets are manually annotated by multiple workers. Collecting data for image sentiment analysis is challenging because not only the image needs to be manually labeled, but also each image normally require multiple people to label it as it is very

TABLE I
DATASETS

Dataset	size	class	p/n	workers
SentiBank-Flickr [3]	316,000	1553 ANPs	N/A	auto
SentiBank-twitter [3]	603	2	3.53	3/3
twitter [18]	1269	2	1.54	3/5
Bing [1]	8,812	3	2.76	2/3

Datasets with the no. of images and no. of mid-level representations if applicable. We also list how the label is generated, the number following manual method shows how many workers label each image. There are much more positive images than negative images as people tend to engage more with positive posts [10]. We discard all the neutral images in all datasets.

common for people to have different opinions on the same image. In the end, only a portion of the labeled data can be used, for example, images with all 5 annotators agree on the sentiment label. We show in table I the number of annotators for each image of each dataset and the minimum number of workers are required to agree on the label.

Due to the size of the datasets, we use the SentiBank-Flickr dataset as our training/validation dataset, and the rest three dataset as our testing dataset. All the images are center cropped to 227 x 227 before feeding to the network. There are 1553 different ANPs in the SentiBank-Flickr dataset belongs to 530 different noun classes.

B. Experiment settings

We evaluate our neural network on a single server with a nVidia P5000 GPU. The learning rate of the network is set at 0.0001 for both training stages. We use a batch size of 64 and train the network for 300 epochs (first stage 200 epochs and second stage 100 epochs). The training time of 300 epochs over 316,000 images is around 3 days and the testing over all three testing sets takes less than an hour. We evaluate our CNN with triplet loss and compared the results with the CNN proposed in [4].

C. ANP classification

TABLE II
FIVE-FOLD CROSS-VALIDATION ACCURACY FOR ANP CLASSIFITION

Model	top-1	top-5	top-10
Baseline CNN [5]	14.76%	18.63%	21.47%
CNN w/ triplet loss	18.33%	25.75%	29.13%

We show the five-fold cross-validation results for our network after two-stage training on SentiBank-Flickr dataset in table II. We can see that by using two-stage training with triplet loss, the ANP classification accuracy has increased for all cases. The highest increase in accuracy is the top-10 accuracy which is higher than 1.4 times the accuracy of the baseline algorithm. This shows that the two-stage learning indeed helps with the classification of the mid-level representation.

To further show that the triplet loss can force the network to learn more on the adjective part in the ANPs, we show several examples that the baseline network failed to recognize the adjective part but our network succeeded in figure 3. We

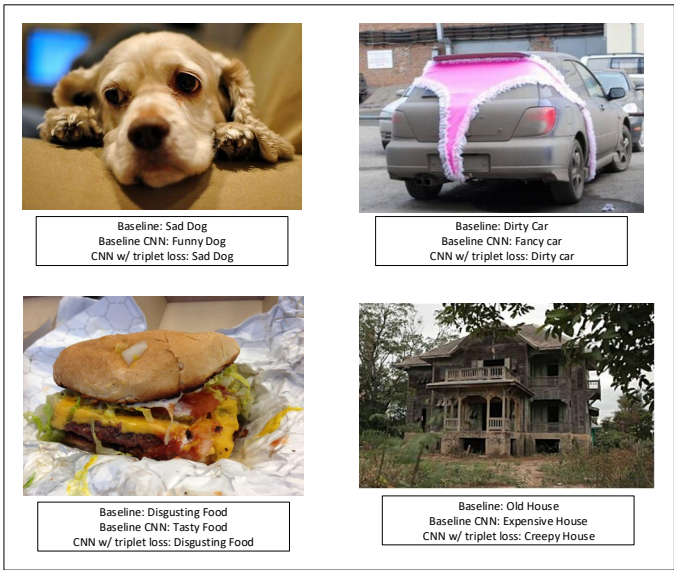

Fig. 3. Example of CNN with triplet loss learning adjective part of ANPs

can see that in these selected examples, CNN with triplet loss can correctly classify the adjective part which the baseline CNN failed to do so. Even CNN with triplet loss failed to predict the correct adjective part for the "old house" image in figure 3, the predicted adjective "creepy" is much closer to the ground truth than the adjective "expensive" predicted by the baseline CNN.

D. Sentiment prediction

TABLE III
IMAGE SENTIMENT PREDICTION ACCURACY

Model	SentiBank-twitter	Twitter	Bing
Baseline CNN [5]	82.4%	74.1%	66.8%
CNN w/ triplet loss	88.3%	76.3%	73.9%

We show in this section the image sentiment prediction based on the predicted ANP by our network on three different testing datasets in III. The final sentiment prediction is made by taking all top-10 ANPs predicted by the neural network. We can see that by increasing the ANP classification accuracy, the image sentiment prediction can reach a higher accuracy.

V. CONCLUSION

In this paper, we adopt a variant of convolutional neural networks with triplet loss to perform end-to-end learning that can automatically generate the most suitable mid-level features for image sentiment analysis tasks. We also designed a specific triplet selection schemes for mid-level representations commonly used in image sentiment analysis. We show the effectiveness of our approaches by detecting both mid-level representations and predicting sentiment of various image datasets and show significant improvement against baseline approaches.

REFERENCES

[1] U. Ahsan, M. De Choudhury, and I. Essa. Towards using visual attributes to infer image sentiment of social events. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 1372–1379. IEEE, 2017.

[2] S. Asur and B. A. Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010.

[3] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232. ACM, 2013.

[4] V. Campos, B. Jou, and X. Giro-i Nieto. From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction. *Image and Vision Computing*, 2017.

[5] T. Chen, D. Borth, T. Darrell, and S.-F. Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[8] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S.-F. Chang. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 159–168. ACM, 2015.

[9] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[12] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2, 2010.

[13] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[14] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[16] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[17] Y. Wang, S. Wang, J. Tang, H. Liu, and B. Li. Unsupervised sentiment analysis for social media images. In *IJCAI*, pages 2378–2379, 2015.

[18] Q. You, J. Luo, H. Jin, and J. Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, pages 381–388, 2015.

[19] J. Yuan, S. Mcdonough, Q. You, and J. Luo. Sentribute: image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 10. ACM, 2013.