An Effective Visual Attention Mechanism

Zhuanghui Wu[a], Guoheng Huang[b] and Lianglun Cheng[c]
School of Computers Guangdong University of Technology (GDUT), Guangzhou, China.
[a]13798115469@163.com, [b]kevinwong@gdut.edu.com, [c]llcheng@gdut.edu.cn

Abstract. Visual sentiment analysis is mainly to study the emotional response of the observer after reading the image. Now, as images become exploding in social networks, the need for visual sentiment analysis is increasing. A major problem with the existing image sentiment analysis is that the emotional label is often related to the local area of the image rather than the whole, but the existing algorithms cannot detect them, so the prediction effect is very poor. In this paper, we propose an attention mechanism to detect areas that are closely related to visual emotions. The experimental results show that the performance of the sentiment classifier based on the detection area has been significantly improved.

Keywords: Visual Sentiment Analysis, Local Image Re- gions, Attention Mechanism.

1. Introduction

VISUAL sentiment analysis has important research significance and commercial value. On the one hand, emotional classification of images can satisfy our search needs. The current image retrieval focuses on content and ignores emotions. On the other hand, as an interdisciplinary subject, studying image sentiment analysis can bring very important enlightenment to the research fields of computer, psychology and aesthetics[1-6,15].

According to research, when humans observe pictures, they tend to pay attention to the selection of local areas rather than the whole area. At the same time, the emotional region is more stimulating to the human perception system than the entire picture[12,15,13,10]. Therefore, the emotional region of the image often determines the emotional label of the image. However, the sentiment analysis model now deals with the entire picture rather than the local area, which is why they cannot detect these most important emotional areas[5-9,11,14].

In our research, we propose a reasonable attention mechanism to discover the emotional regions of the image. In order to compensate for the semantic gap between objective visual features and subjective emotional features, we have added subjective emotional attributes to the attention mechanism. By paying attention to the feature map through the emotional attribute, the attention mechanism can detect the emotional areas closely related to the image. Finally, there is a significant improvement in the performance of classifiers built on these emotional areas. The emotional areas detected is shown as Fig. 1

Fig. 1 The emotional areas detected are shown in the right

As an important innovation in our approach, our proposed attention model has two advantages: (1) it can depict various aspects of various sample images, for example, different concerns highlight different parts of an object. Or highlight the object and its visual environment; (2) reduce the attention map is incorrect, because more attention means that having at least one note will be more correct and relevant. Our contributions are summarized as follows.

We design an attention mechanism to detect emotional regions. It can overcome the semantic gap between objective visual features and subjective emotional attributes.

2. Model

The proposed framework mainly consists of two parts, the attention mechanism and the sentiment classifier, which is shown as Figure 1. The attention model based on the subjective emotional attributes detects areas related to image emotions, and the sentiment classifier is built on these areas and predicts emotions. For a picture, the emotional tags and emotional attributes of the picture are defined as y and a, respectively.

2.1 Feature Extraction

In this work, we extracted the convolutional features of the image CNN and treated them as an array of all local regions. For subjective emotional attribute markers, we use the groom vector as the semantic embedding. Glove vector has been widely used in sentiment analysis.

2.2 Feature Encoding

First, we designed an encoder to encode all the feature maps. The coding formula is as shown in (1).

$$\mathrm{V_{E,i}} = f(W_v \times V_i + b_v) \tag{1}$$

The $f(a) = \max(0, a)$ is a ReLU linear function, and $\mathrm{V_i}$ is the i-th feature map generated by the VGG-16. The W_v and the $\mathrm{b_v}$ are the parameters for the encoder.

Then, we designed another encoder to encode word representations. The coding formula is as shown in (2).

$$\mathrm{a}_E = f(W_a \times a + b_a) \tag{2}$$

The $f(a) = \max(0, a)$ is a ReLU linear function, and a is the truth attribute for the image. The W_a and the b_a are the parameters for the encoder.

2.3 Attention Map Prediction.

Not directly aggregating these local regions to create image representation, we create its representation through a multiple attention mechanism. Basic form of shown as (3)

$$g_i = o_i \cdot v_{E,i} \tag{3}$$

The g_i is the i-th attention map. In the follows, we will calculateo_i, which is attention value. We apply the generated encoding visual vectors vE to the attribute encoding vectors aE to get its initial attention confidence score. Scores in areas related to emotional attributes are expected to be higher. This is driven by the objective function, which means that it weakens the characteristics of the emotionally unrelated regions and strengthens the characteristics of the emotional regions, so that the

region with high confidence score is the emotional region we are looking for. Finally, we normalize the confidence score as shown in Equation 4. The formula is as shown in (5).

$$\hat{o}_i = a_E \cdot v_{E,i} \tag{4}$$

$$o_i = \frac{b(\hat{o}_i)}{\sum_{i=1}^{z} b(\hat{o}_i)} \tag{5}$$

The z is the number of feature maps. We extract features in VGG16, therefore, z is 196. The $b(\cdot)$ is a function guaranteeing that the attention scores, which is a ReLU linear function, $f(a) = \max(0, a)$.

2.4 Sentiment Classifier

Finally, we build a sentiment classifier on top of these local regions as shown in (6). We minimize a softmax loss function.

$$L_{\text{sentiment}} - \text{loss} = -\log(p(s), y) \tag{6}$$

The $p(s) = \sum_{i=1}^{z} o_i$ and the y is the emotional label for the image. Meanwhile, p is a softmax function.

3. Experiment

3.1 VSO Dataset

This is the largest visual sentiment analysis dataset in the publicly available visual sentiment ontology data set [11]. All images in the VSO dataset are from flick, each image corresponds to an ANP (adjective-noun pair), and has an emotional label, which we design as attribute emotions and image emotions, as showed in Fig.3. The VSO data set has a total of 3,244 ANPs and approximately 1.4 million images. Of the 3,244 ANPS, there are 269 adjectives. Among them, 127 attributes correspond to positive emotions. However, the data set is unbalanced. Since there are too many positive images, we randomly construct a balanced data set using negative images. Due to the small number of images of some attributes, we have data enhancements for them, so each attribute has at least 700 images. Finally, we have about 1.1 million images, and we randomly divide the images into 80% for training attribute detectors and attention models, 10% for testing and 10% for verification. The data enhancement method is as follows.

- Flip the image horizontally and vertically.
- Add Gaussian noise to the image.
- Change the intensity of the image.
- Randomly crop the image.
- Randomly scale the image.

3.2 Training Details

In training, the parameters are learned end-to-end on the training images using stochastic gradient de- scent (SGD) with learning rate of 0.001. We use 32 images for each iteration. Each epoch has 1000 iterations and we use 32 images per iteration. In each epoch, we will verify the network whether it is converging or over- fitting in the validation set containing approximately 2,000 images. When the loss function does not improve on the verification set, we stop learning.

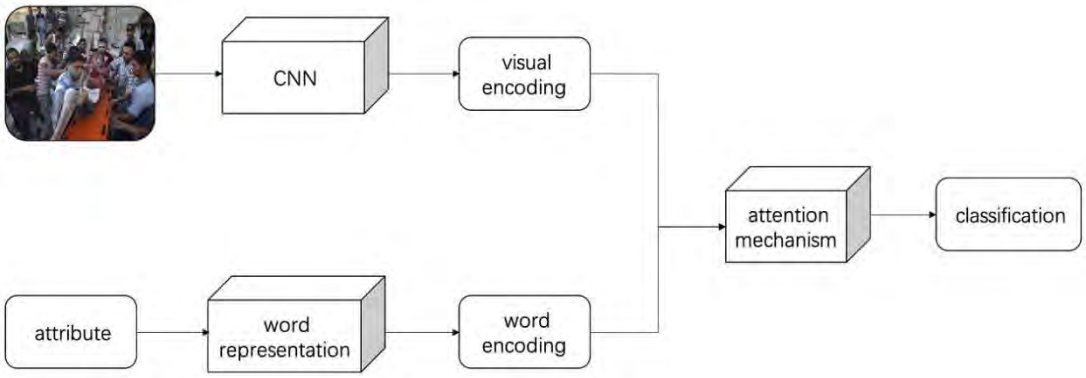

Fig. 2 he proposed framework mainly consists of two parts, the attention mechanism and the sentiment classifier. The attention model based on the subjective emotional attributes detects areas related to image emotions, and the sentiment classifier is built on these areas and predicts emotions.

Fig. 3 The VSO dataset.

3.3 Result Analysis

The performance of attention mechanism is shown as Table 1. The most advanced models available can only achieve an accuracy of 70%, because the emotional regions most closely related to the image emotional tags are not detected. The entire image contains a lot of useless information that will greatly categorize the image model. Therefore, we use the proposed attention mechanism to extract local regions closely related to image emotions. These local regions can reflect the emotional features of images well, thus achieving the accuracy of 100%. The experimental results show that the proposed attention mechanism can be well applied in visual sentiment analysis.

Table 1 The performance on VSO dataset

Model	Category	Accuracy
Sentiment detector[10]	2	70%
Attention mechanism	2	99%

4. Conclusion

In the existing visual sentiment analysis models, they deal with the entire picture rather than the local emotional area. However, in the human visual attention system, local regions are often given priority, which leads to the irreplaceable role of local regions in visual sentiment analysis. In our work, we deal with local areas rather than entire image. Through the attention mechanism, the emotional regions of the image are detected exactly. Finally, the experimental results prove that proposed attention mechanism is superior to the most advanced models available. In the future, we will continue to study how local areas act on visual emotions.

Acknowledgments

This work was sponsored by High-resolution Earth Observation Major Project of China (grant number 83—Y40G33-9001-18/20), Guangzhou Major Science and Technology Projects, China (201604010096), Guangdong Province Applied Science and Technology R&D Special Fund Project (2015B090922013), Guangdong Provincial Science and Technology Plan Project (2016B090918017).

References

[1]. Siersdorfer, S.; Minack, E.; Deng, F.; and Hare, J. S. 2010b. Analyzing and predicting sentiment of images on the social web. In ACM MM, 715–718.

[2]. Deep cross residual learning for multitask visual recognition. arXiv preprint arXiv:1604.01335.

[3]. Karpathy, A., and Li, F. 2015. Deep visual-semantic alignments for generating image descriptions. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3128–3137.

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

[5]. Mnih, V.; Heess, N.; Graves, A.; and kavukcuoglu, k. 2014. Recurrent models of visual attention. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., Advances in Neural Information Processing Systems 27. Curran Associates, Inc. 2204–2212.

[6]. Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

[7]. Escorcia, V.; Niebles, J. C.; and Ghanem, B. 2015. On the relationship between visual attributes and convolutional networks. In CVPR 2015, 1256–1264. IEEE.

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

[9]. Vinyals, O.; Kaiser, Ł.; Koo, T.; Petrov, S.; Sutskever, I.; and Hinton, G. 2015. Grammar as a foreign language. In Advances in Neural Information Processing Systems, 2755– 2763.

[10]. Ma, L.; Lu, Z.; Shang, L.; and Li, H. 2015. Multimodal convolutional neural networks for matching image and sentence. In The IEEE International Conference on Computer Vision (ICCV).

[11]. You, Q.; Luo, J.; Jin, H.; and Yang, J. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In Proceedings of the Twenty- Ninth AAAI Conference on Artificial Intelligence, 381–388.

[12]. Pennington, J.;Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

[13]. Tao Chen, Felix X Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. Object-based visual sentiment concept analysis and application. In ACM MM, 2014.

[14]. Anastasia Giachanou and Fabio Crestani. Like it or not: A survey of twitter sentiment analysis methods. In ACM Computing Surveys (CSUR), 2016.

[15]. Ulrike Rimmele, Lila Davachi, Radoslav Petrov, Sonya Dougal, and Elizabeth A Phelps. Emotion enhances the subjective feeling of remembering, despite lower accuracy for contextual details. In Emotion, 2011.