Journal of KIIT. Vol. 19, No. 11, pp. 37-43, Nov. 30, 2021. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2021.19.11.37

Efficient Visual Sentiment Detector using Knowledge Distillation

Hyun Wook Kang*, Kwangil Kim**

"This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW supervised by the IITP(Institute of Information & communications Technology Planning & evaluation)"(2016-0-00017)

Abstract

In this paper, distillation method is used to build neural networks for efficient visual sentiment detector. An ensemble of many layers or a single large model are commonly used at training stage to increase the performance. Although the delay is not as much as it is for training, a lot of resources are still needed at deployment stage. As a result, mobile devices with limited computing power and resources cannot fully deploy those large networks. On the other hand, the performance is decreased using small networks for low-powered devices. To tackle this problem, we used knowledge distillation method to improve the performance of small networks. Using distillation method, the proposed small network can perform as much as VGGNet, which is trained with large number of parameters. The distilled model that is trained fast can be used to build an ensemble model and it is expected to contribute to build a real-time visual sentiment detector.

요 약

논문에서는 효율적 시각 감정 인식을 위해 지식 추출 방법을 사용하는 신경망 모델을 제안한다. 높은 성능을 위해 많은 수의 신경망 모델로 앙상블을 구성하거나 대규모의 단일 신경망 모델을 사용하여 대규모 훈련데이터로 훈련시키는 것은 흔한 일이다. 그러나 이러한 무거운 모델은 학습 단계 만큼은 아니지만 테스트 단계에서도 많은 자원을 필요로 하기 때문에, 이동식 기기 같은 저성능 장치에서는 사용에 제약을 받으며, 저성능 장치를 위한 작은 모델은 무거운 모델만큼 성능을 발휘할 수 없다. 이 문제를 해결하기 위하여 본 논문에서는 지식 추출을 통해 작은 모델의 성능을 개선하였다. 지식 추출 방법을 통해서 제안하는 작은 모델은 무거운 모델인 VGGNet과 비슷한 성능을 보일 수 있다. 빠르게 학습된 제안 모델은 앙상블 모델에 활용될 수 있으며 실시간 감정 인식기에도 기여할 것으로 기대된다.

Keywords

visual sentiment analysis, neural network, pre-trained model, deep learning, distillation

* Department of Computer Science and Engineering, Dongguk University, Seoul
- ORCID: https://orcid.org/0000-0002-5350-7273
** Dongguk Institute of Convergence Education, Dongguk University, Seoul
- ORCID: https://orcid.org/0000-0002-9371-7873

· Received: Oct. 31, 2021, Revised: Nov. 18, 2021, Accepted: Nov. 21, 2021
· Corresponding Author: Kwangil Kim
Dongguk Institute of Convergence Education, Dongguk University
Tel.: +82-2-2290-1404, Email: kikim6114@dongguk.edu

Ⅰ. Introduction

Automatic assessment of sentiment analysis can be applied in many areas[1][2]. For example, an automatic tagging system can be constructed if the evoking emotion can be detected from an image someone has posted to social media. Most of the earlier studies on sentiment analysis has made use of textual contents [3]-[5]. However, Borth et al.[2] gives an example in which the valuable sentiment is conveyed primarily by images but texts. Thus, analyzing visual contents such as images and videos is an important aspect of sentiment analysis research.

Analyzing affective images is more challenging due to the different human responses evoked by the same stimuli. Previous studies on visual sentiment analysis mostly uses deep CNNs[1][6]-[8] due to their success in classification tasks since introduced by Krizhevsky et al.[9]. Nowadays, there are various types of CNNs, such as AlexNet[7], VGGNet[10], ResNet[11], and etc.

However, these large networks have one critical problem that they require large storage capacity and computing power. Although training and deployment of the model have very different objectives, typical neural network models use similar models for training stage and deployment stage[12]. For example, small networks should be utilized for automatic sentiment tagging systems because they need to operate in real time. However, the performance is decreased when using small networks. Therefore, knowledge distillation is a promising method to improve the performance of small networks that can be deployed with limited computing power.

To achieve this goal, we train a small network by transferring the knowledge in a large teacher network to the target small network. At training stage, the teacher network is trained with many layers. The teacher network inherits all the layers from ResNet [11] model to its second last fully connected layer and additional layers are added. Then the teacher network is fine-tuned with affective images dataset, ArtPhoto[14], that has 8 emotion classes. Once the teacher network has been trained, the knowledge is transferred from the teacher to the student network. The logits from teacher network preserve probabilities for each class and their values are greater than 0 and less than 1. Logits correspond to soft targets because they do not give one true answer like hard targets. Therefore, soft targets have more information than hard targets. Using this information from soft targets, student network can learn faster with less epochs and higher learning rate because they reduce the variance in the gradient during training. This method is originally introduced by Hinton et al.[12] and it is called distillation. Inspired by this work, we use distillation method, to train a scalable and real time visual sentiment detector with affective images.

This paper is organized as follows:
In section 2, previous methods for visual sentiment analysis and distillation methods will be reviewed.
In section 3, the details about model architecture will be explained.
In section 4, the used dataset is described.
In section 5, the comparative results are given.
In section 6, we analyze the results.
In section 7, the final conclusion is made.

Ⅱ. Related Works

2.1 Visual Sentiment Analysis

There are two main approaches for predicting sentiments from images. The first method uses hand-crafted features[13] and another one uses deep learning frameworks[8]. Regarding hand-crafted features, a combination of low-level features was defined by [14] according to psychology theories. These low-level features are composed of color, texture, and composition. Later, efforts have been made to find more robust features by [15]. Further studies have been made to predict personalized emotion perceptions

by Zhao et al.[16][17], in which they proposed the multi-task hypergraph learning and released IESN dataset. They have motivated the following studies of subjectivity in visual sentiment analysis by jointly considering different emotional aspects such as visual content, social context, temporal evolution and location influence.

Recent methods use CNNs to extract emotions from images[1][6]-[8]. Adjective Noun Pairs(ANP) are exploited by [18] to detect sentiments depicted in images. A novel progressive CNN architecture called PCNN is used by You et al.[19]. Yang et al.[6] optimized retrieval and classification simultaneously by employing the multi-task framework. Later, global and local information is considered by [1][20] to recognize emotions from images.

2.2 Distillation Method

Distillation method is to transfer knowledge from a cumbersome model to a small model[12]. Motivated by [21], Hinton et al.[12] introduced a new training method called distillation. They used logits, which are inputs to the softmax, to train the deployment network. They do not use class probabilities output from softmax because the values are so close to zero, which have little influence on the cross entropy cost function during the transfer stage. Thus, logits from teacher network are used to learn a small network instead of class probabilities output from softmax. By matching the logits, the training time is faster because the variance of the gradient is smaller between training cases[12].

III. Methodology

Our efficient visual sentiment detector is trained with logits from a large teacher network. The model architecture of teacher network and student network is shown in Table 1 and Table 2, respectively.

Table 1. Teacher network architecture

Name	Output size	Layers
conv1	112 x 112	7x7, 64, stride 2
conv2_x	56 x 56	3 x 3 max pool stride 2 $\begin{bmatrix} 1\times1,\ 64 \\ 3\times3,\ 64 \\ 1\times1,256 \end{bmatrix}\times3$
conv3_x	28 x 28	$\begin{bmatrix} 1\times1,128 \\ 3\times3,128 \\ 1\times1,512 \end{bmatrix}\times4$
conv4_x	14 x 14	$\begin{bmatrix} 1\times1,\ 256 \\ 3\times3,\ 256 \\ 1\times1,1024 \end{bmatrix}\times6$
conv5_x	7 x7	$\begin{bmatrix} 1\times1,\ 512 \\ 3\times3,\ 512 \\ 1\times1,2048 \end{bmatrix}\times3$
	1x1	average pool, 8-d fc

Table 2. Student network architecture

Layer(Type)	Output size
Conv2D	222 x 222
MaxPooling2D	55 x 55
Conv2D	53 x 53
GlobalAveragePooling2D	128-d
Dense	512-d
Dense	8-d

To fine-tune the teacher network, we use ResNet [11] as the backbone network. From the input layer to the second last fully connected layer with 2048 units are used to build the teacher network. The layer is regularized by dropout with 50% and reduced to 8 dimensions, which is the final teacher network, to output the logits. Once the teacher network is trained, the student network is trained with the logits from the large teacher network to calculate the error rates between them.

The student network has two hidden layers and two fully connected layers, which is a lot smaller compared to the teacher network.

The overall architecture of matching logits between student network and teacher network is shown in below Fig. 1.

The first and second hidden layer of the student network have 64 and 128 units, respectively.

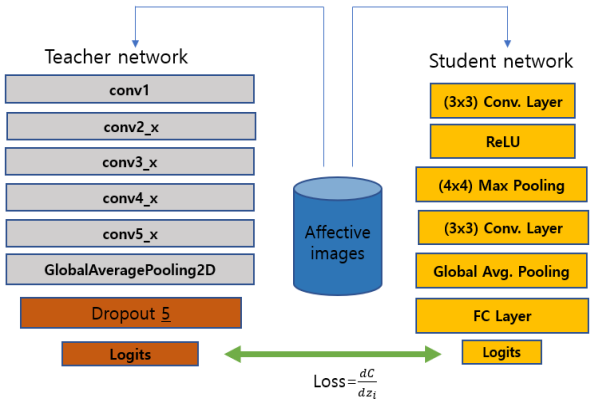

Fig. 1. Overall architecture of proposed system. The student network learns from a large teacher network by matching its logits to the logtis of its teacher

The first hidden layer is followed by ReLU activation and max pooling. The second hidden layer is followed by global average pooling and two fully connected layers. The last fully connected layer has 8 nodes. The respective logits are the outputs from the last fully connected layer of teacher and student network. The temperature of logits is set to 4 because it showed highest prediction accuracy. The error rates between the logits of the student network and the teacher network are calculated as in equation 1

$$\frac{dC}{dz_i} = \frac{1}{T}(q_i - p_i) = \frac{1}{T}\left(\frac{e^{z_i/T}}{\sum_j e^{z_j/T}} - \frac{e^{v_i/T}}{\sum_j e^{v_j/T}}\right) \quad (1)$$

The respective cross-entropy gradient is denoted as $\frac{dC}{dz_i}$ for the corresponding logit of the distilled model, which is denoted as z_i. The logits from the teacher is denoted as v_i and the produced soft targets are represented with p_i. The temperature is denoted as T. By learning the parameters from the teacher network, the small student network can converge faster.

Ⅳ. Dataset

ArtPhoto [14] is a public benchmark dataset for affective images. The emotion of each image is defined by the owner. The images are collected from photo sharing sites and the total number of images is 806. We used this dataset because they are easily accessible and can be downloaded from the internet. In ArtPhoto, there are eight emotion classes as follows: amusement, anger, awe, contentment, disgust, excitement, fear, and sad.

Ⅴ. Experiments

Among the dataset, 90% samples are used for training and 10% samples are used for testing. The teacher network is trained with 30 epochs and the batch size is set to 20. We used adam optimizer with learning rate 1e-5 for the teacher network. On the other hand, the student network is trained with less epochs and smaller learning rate of 10 and 1e-3, respectively. To compare the accuracy and training time with other models, we also show the results of VGGNet and AlexNet.

The VGGNet model is enhanced with additional layers. The second fully connected layer of VGG16 is extracted followed by two fully connected layers. Dropout is applied to the first additional layer with 50%. Then the number of nodes is reduced to 8 nodes and softmax function is applied.

Regarding AlexNet, two fully connected layers are removed from AlexNet[7] model due to memory limitation. Instead, one fully connected layer with 2000 nodes is added before softmax function. Both VGG16 and AlexNet use the same hyperparameter settings as the teacher network.

Ⅵ. Results

Below Table 3 shows the respective elapsed time of training time for distilled model, teacher model, VGGNet and AlexNet model.

As shown in above Table 3, the training times for each temperature value of the proposed model is a lot

less than the rest of the models by a significant margin.

Table 4 shows the comparative results of the accuracy for each model. The accuracy is the rate of predicting the correct emotions evoked from test images out of 8 emotion classes. The teacher model shows the highest accuracy compared to the rest of the models. VGGNet shows the next highest accuracy.

However, large networks such as the teacher model and VGGNet can not be deployed in mobile devices with less storage capacity and limited computing power due to their large number of parameters. Although there is about 6% decrease in the accuracy of the distilled model, it can reach performance as much as the VGGNet showing the next highest performance.

Due to the reduced number of parameters, the proposed model can predict much faster.

Table 3. Training time of all models. Different training times of the Distilled model to distill the knowledge in a Teacher network is given according to different temperature values.

(unit: seconds)

Temperature	Distilled model	Teacher model	VGGNet	AlexNet
3	28.21	189.08	76.02	376.38
4	**28.42**			
5	28.61			
6	28.68			
7	28.42			

Table 4. Accuracy (%) of predicting emotions by all models. Respective accuracy of the Distilled model is given for different temperature values, where T denotes the corresponding temperature value.

	Distilled model	Teacher model	VGGNet	AlexNet
# of parameters in millions	0.146M	23.604M	142.471M	540.305M
Accuracy	34.5% (T=3)	43.2%	**38.3%**	23.5%
	37.4% (T=4)			
	35.3% (T=6)			
	35.3% (T=7)			
	35.3% (T=8)			

The number of parameters of the proposed model decreased roughly by 23M (99%) and 142M (99%) compared to teacher model and VGGNet, respectively. Thus, it is expected to contribute to real-time visual sentiment detector.

Ⅶ. Analysis

The distilled model can be deployed in devices with limited hardware. In addition, there is less latency during prediction due to their less number of parameters. Although there is a loss in accuracy compared to its teacher model, it shows similar performance as much as a large VGGNet. In addition, the accuracy is higher roughly by 13% compared to the variation of AlexNet without distillation. The proposed model is more scalable because it can be retrained with different affective images according to different user objectives due to the fast retraining time.

However, accuracies of existing models trained by directly mapping sentiments to the visual contents is not as good as image classification. It is due to the gap between low-level visual features and high-level emotion concepts. In the literature, this gap is defined as affective gap[14]. In the future, we plan to bridge this gap to improve the performance of large teacher networks.

Ⅷ. Conclusion

In this paper, we propose an efficient visual sentiment detector by training a small network with distillation methods. Although a large network can be well trained to show a good performance in classification task, the large computation overhead makes it hard to be generalized to different user objectives. By utilizing distillation method, the deployment model of visual sentiment detector can be efficiently trained.

The performance of efficiently trained visual sentiment detector shows similar performance as much as the large VGGNet. In addition, the proposed model can predict much faster due to its less number of parameters. Thus, it is expected to contribute to a real-time visual sentiment detector.

The performance of student network is dependent on teacher network. Thus, improving the performance of large teacher network can act on the performance of small student network. However, when neural network models are applied to visual sentiment analysis, accuracies are not as good as image classification. To improve the performance of neural network models for emotion recognition, Yang et al. [1] utilized the local information, from which the sentiment is evoked. Inspired by this work, we plan to improve the performance of large teacher networks by giving attention to Region of Interests(ROI) that elicits the sentiment as a future work.

References

[1] J. Yang, D. She, Y. Lai, P. L. Rosin, and M. Yang, "Weakly supervised coupled networks for visual sentiment analysis", In Proceedings of the IEEE conference on computer vision and pattern recognition, Lake Salt City, UT, USA, pp. 7584-7592, Jun. 2018.

[2] D. Borth, R. Ji, T. Chen, T. Breuel, and S. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs", Proceedings of the 21st ACM international conference on Multimedia, Barcelona Spain, pp. 223-232, Oct. 2013. https://doi.org/10.1145/2502081.2502282.

[3] B. Pang and L. Lee, "Opinion mining and Sentiment Analysis", Information Retrieval, Vol. 2, No. 1–2, pp. 1-135, Jun. 2008. https://doi.org/10.1561/1500000011.

[4] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis", Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 347-354, Vancouver British Columbia Canada, Oct. 2005. https://doi.org/10.3115/1220575.1220619.

[5] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment Strength Detection in Short Informal Text", Journal of the American Society for Information Science and Technology, Vol. 61, No. 12, pp. 2381-2594, Dec. 2010. https://doi.org/10.1002/asi.21416.

[6] K. Peng, T. Chen, A. Sadovnik, and A. Gallagher, "A mixed bag of emotions: Model, predict, and transfer emotion distributions", In Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, pp. 860-868, Jun. 2015. https://doi.org/10.1109/CVPR.2015.7298687.

[7] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and benchmark", In Proceedings of the AAAI conference on artificial intelligence, Phoenix Arizona, Vol. 30, No. 1, pp. 308–314, Feb. 2016.

[8] J. Yang, D. She, Y. Lai, and M. Yang, "Retrieving and classifying affective images via deep metric learning", Thirty-Second AAAI Conference on Artificial Intelligence, Vol. 32, No. 1, Apr. 2018.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", Proceedings of the 25th International Conference on Neural Information Processing Systems, Vol. 60, No. 6, pp. 84-90, Dec. 2012. https://doi.org/10.1145/3065386.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", 3rd International Conference on Learning Representations, May 2015.

[11] 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, Las Vegas, NV,

USA, pp. 770-778, Jun. 2016.

[12] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network", NIPS, Mar. 2014.

[13] V. Yanulevskaya, J. C. van Gemert, K. Roth, A. K. Herbold, N. Sebe, and J. M. Geusebroek, "Emotional valence categorization using holistic image features", 15th IEEE international conference on Image Processing, San Diego, CA, USA, pp. 101-104, Oct. 2008. https://doi.org/10.1109/ICIP.2008.4711701.

[14] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by Psychology and Art Theory", Proceedings of the 18th ACM international conference on Multimedia, New York, NY, USA, pp. 83-92, Oct. 2010. https://doi.org/10.1145/1873951.1873965.

[15] S. Zhao, Y. Gao, X. Jiang, H. Yao, T. Chua, and X. Sun, "Exploring Principles-of-Art Features For Image Emotion Recognition", Proceedings of the 22nd ACM international conference on Multimedia, New York, NY, USA, pp. 47-56, Nov. 2014. https://doi.org/10.1145/2647868.2654930.

[16] J. Yang, D. She, and M. Sun, "Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network", Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Vienna, Austria, pp. 3266-3272, Aug. 2017. https://doi.org/10.24963/ijcai.2017/456.

[17] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, and T. Chua, "Predicting Personalized Emotion Perceptions of Social Images", Proceedings of the 24th ACM international conference on Multimedia, New York, NY, USA, pp. 1385-1394, Oct. 2016. https://doi.org/10.1145/2964284.2964289.

[18] T. Chen, D. Borth, T. Darrell, and S. Chang, "DeepSentiBank: Visual Sentiment Concept Classification with deep Convolutional Neural Networks", Oct. 2014.

[19] Q. You, J. Luo, H. Jin, and J. Yang, "Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks", Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin Texas, pp. 381-388, Jan. 2015.

[20] Q. You, H. Jin, and J. Luo, "Visual Sentiment Analysis by Attending on Local Image Regions", Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco California USA, pp. 231-237, Feb. 2017.

[21] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model Compression", Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, pp. 535-541, Aug. 2006. https://doi.org/10.1145/1150402.1150464.

Authors

Hyun Wook Kang

2015 : BS of Information Technology from the University of Newcastle, Australia

2021 ~ present : Research intern. in Department of Computer Science, National University of Singapore.

Research interests : sentiment analysis, video understanding, zero-shot learning, cross-modal retrieval, AI and NLP

Kwangil Kim

1981 : B.S degree in Department of Industrial Engineering, Hanyang Univerisity

1983 : M.S degree in Department of Management Science, KAIST

2017 ~ present : Professor in Department of Convergence Education, Dongguk Univerisity

Research interests: Artificial Intelligence, Data Science