2015 International Conference on Information Processing (ICIP)
Vishwakarma Institute of Technology. Dec 16-19, 2015

Image Sentiment Analysis using Deep Convolutional Neural Networks with Domain Specific Fine Tuning

Stuti Jindal and Sanjay Singh
Department of Information & Communication Technology
Manipal Institute of Technology, Manipal University, Manipal-576104, INDIA
stuti.jindal11@gmail.com, sanjay.singh@manipal.edu

Abstract—**Images are the easiest medium through which people can express their emotions on social networking sites. Social media users are increasingly using images and videos to express their opinions and share their experiences. Sentiment analysis of such large scale visual content can help better extract user sentiments toward events or topics, such as those in image tweets, so that prediction of sentiment from visual content is complementary to textual sentiment analysis. Significant progress has been made with this technology, however, there is little research focus on the picture sentiments. In this work, an image sentiment prediction framework is built with Convolutional Neural Networks (CNN). Specifically, this framework is pretrained on a large scale data for object recognition to further perform transfer learning. Extensive experiments were conducted on manually labeled Flickr image dataset. To make use of such labeled data, we employ a progressive strategy of domain specific fine tuning of the deep network. The results show that the proposed CNN training can achieve better performance in image sentiment analysis than competing networks.**

I. Introduction

These days Internet has become a major platform for communication and information exchange, providing us a rich repository of peoples perspectives and sentiments regarding a copious spectrum of topics. Such knowledge is embedded in sundry facets, such as blogs, comments and tags on microblogging sites. The scrutiny of such information in the domain of sentiment analysis plays an indispensable role in behavior study, which aims to understand and predict human decision making and allows a wide range of applications in business intelligence, stock market prediction and political vote forecasts.

Most of photo posts on social networks hardily contain any description or caption. Hence, lots of opinions and emotions are conveyed through visual content only. Till now the computational analysis of sentiment mostly converges on the textual content. Restrained efforts have been conducted to analyze sentiments from visual content such as images and videos, which is becoming a pervasive media type on the web. In this era of big data, it has been discerned that the exploration of visual content can provide us with reliable or complementary online social signals [1].

The Deep Learning [2] framework enables robust and accurate feature learning, which in turn produces the state of the art classification performance for images. For images related tasks, Convolutional Neural Network (CNN) are widely used due to the usage of convolutional layers. It takes into consideration the locations and neighbors of image pixels, which are important to capture useful features for visual tasks as shown in [3].

In addition to the existing benchmarks that only includes positive or negative labels for images, a 7-scale sentiment rating has been introduced in [4], which accounts for neutral images and different sentiment strength of the same polarity.

Instead of focusing on defining and training mid-level attributes related to emotional perception of [1] and [4], a framework has been used that efficiently transfers CNNs learned on a large-scale dataset to the task of visual sentiment prediction as shown in [5]. The transfer learning has a major advantage over those standard approaches because there is no requirement of domain knowledge from psychology or linguistics.

We intend to find out whether applying CNN to visual sentiment analysis provides advantages. To that end, in this work we address a major challenge as how to fine tune with a small scale labeled training data that is different from the dataset used for image classification. Since the dataset is unlike, it might be best to perform learning only in top layers during backpropagation as they have less dataset specific features. The experimental results suggest that this domain specific fine tuning is effective for improving the performance of neural network.

Rest of the paper is organized as follows. Section II discusses the most recent works in the area of image sentiment prediction. Section III explains about the dataset collection and construction process. Section IV describe the proposed computational system for sentiment prediction. Section V briefs about the experimental results. Section VI discusses about the results obtained and finally Section VII concludes this paper.

II. Related Work

Previous work on visual sentiment analysis has mostly been conducted to develop mid-level attributes for selecting features from low-level image features. Motivated by the fact that sentiment involves high-level abstraction, which may be easier to explain by objects or attributes in images. Borth et al. [4]

978-1-4673-7758-4/15/$31.00 ©2015 IEEE

and Yuan et al. [1] has proposed to employ visual entities or attributes as features for visual sentiment analysis. The major drawback of these approaches is that the training process requires lots of domain knowledge of psychology or linguistics to define the mid level attributes and human intervention to fine tune the sentiment prediction results.

Deep compositional architectures introduced by Krizhevsky et al. [3] has outperformed all known image classification pipelines on ImageNet large scale visual recognition challenge ILSVRC 2012 [6]. Deep convolutional neural networks (CNN) used by Krizhevsky et al. [3] are layered classifiers with millions of parameters. As large scaling datasets have been introduced by Russakovsky et al. [6], fully supervised learning of CNNs have become possible without over fitting huge amount of parameters.

Recent studies by Donahue et al. [7] and Oquab et al. [8] shows that the parameters of CNN trained on large-scale dataset such as ILSVRC can be transferred to object recognition and scene classification tasks when the data is limited, resulting in better performance than traditional hand-engineered representations. Xu et al. [5] has proposed a novel sentiment analysis framework based upon convolutional neural network for visual sentiment prediction. It shows that the image representations from the CNN trained on a large-scale dataset could be efficiently transferred for sentiment analysis. Our work is motivated by the work of Xu et al. [5], wherein we apply the concept of transfer learning Deep CNN from large-scale image classification to the problem of sentiment prediction using small scale datasets which are disparate in nature from the pretraining image dataset, for domain specific fine tuning.

III. The Data Set

To evaluate the proposed method we have taken real world dataset from a major microblogging sites, namely Flickr [9]. The dataset used [10] is publicly available and the details of the data collection and ground truth labeling are discussed in this section. Also we have constructed a 7 scale sentiment rating data set contrasting to the benchmark methods. Process of dataset collection and construction is discussed in this section.

A. Data Collection

We utilize the Flickr photo posts that have detectable sentiment content. Typically, the tags indicate the users sentiment for the uploaded images. To annotate the dataset, a survey is conducted among the UG and PG students of Manipal Institute of Technology. Two hundred and sixty seven students participated in this survey and annotated 1000 images for the respective sentiment. A 7 scale sentiment rating is used which accounts for neutral images and different sentiment strength of the same polarity. This lexicon lists the neutrality or polarity of frequently used words with subjectivity clues out of seven sentiment levels namely depressed, very sad, sad, neutral, happy, very happy and excited. For instance, according to the lexicon, happy is positive in the strongly subjective sense while sad is negative in the strongly subjective sense.

B. Ground Truth Labeling

To obtain ground truth of the collected photo posts, we asked several annotators to label the data. Each image were assigned to exactly 3 annotators. Annotators were asked to provide a sentiment score out of a 7-scale labeling scheme ranging from 1 to 7.

Capturing sentiment strength is also important along with capturing sentiment polarity. Indeed, fine-grained categorization in sentiment strength is widely accepted in text analysis as shown by Thelwall et al. [11]. It has been explicitly shown in [5] that utilizing sentiment strength is a better choice for describing sentiment strength in visual content rather than fine-grained categorization.

After collecting all the annotations, we took the majority vote out of the 3 scores for each image; that is an image annotation is considered valid only when at least 2 out of 3 annotators agree on the exact label (out of 7 possible labels). A total of about 3000 annotations were completed for the images of the dataset on the website. From the collected annotations, the majority vote is taken for each image. Overall, a set of 806 images is collected with image-text combined ground truth out of the total 1000 images. This dataset is now available publicly [12].

IV. The Computational System

This section explains about the proposed computation system for image sentiment prediction. Figure 1 shows the image sentiment prediction framework for our system. We deploy a suitable convolutional neural network architecture for visual sentiment analysis. Moreover, we employ a progressive training strategy that leverages the training results of convolutional neural network. The details of the proposed framework will be described in the following sub-sections.

A. Deep Convolutional Neural Networks

For the pretrained CNN, we have used the deep architecture of Krizhevsky et al. [3]. It is composed of eight learned layers that are five convolutional and three fully connected with weights. The CNN is composed of seven internal layers and ultimately a soft-max layer. The hidden layers are five successive convolutional layers followed by two fully connected layers. The nonlinearity of each neuron in this CNN is modeled by Rectified Linear Units (ReLUs) [13]

$$f(x) = max(0,\ x)$$

which accelerates learning as compared with saturating nonlinearity. The CNN takes a 224×224 pixel RGB image as input. Each convolutional layer convolves the output of its previous layer with a set of learned kernels, followed by ReLU nonlinearity, and two optional layers, local response normalization and max pooling. The local response normalization layer is applied across feature channels, and the max pooling layer is applied over neighboring neurons. The output of the last

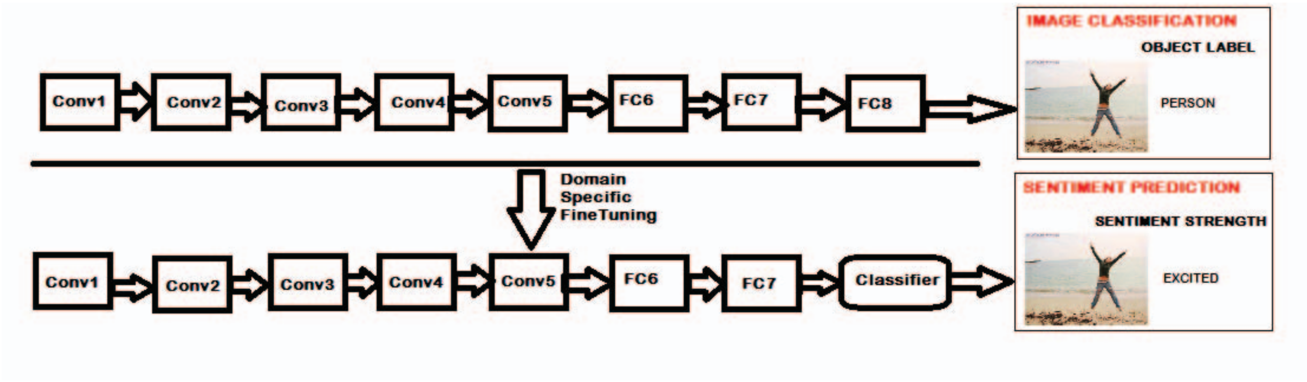

Fig. 1. The overall architecture of the proposed visual sentiment prediction framework. The CNN is first trained on a large scale dataset (ImageNet) for image classification. The parameters of pretrained layers are transferred to the problem of sentiment prediction for generating image representations using domain specific fine tuning.

fully connected layer is fed to the classifier which produces a distribution over the thousand of class labels.

The advantage of using this architecture of Convolutional Neural Network is that it maximizes the average log probability across training cases by predicting the correct label on distribution.

B. Image Classification

The open source implementation of Jia et al. works [14] named Caffe has been used [15]. In our research work, Caffe is used to execute the network of Krizhevsky [3] to train the CNN on ILSVRC 2012 dataset [6]. It is a subset of ImageNet, consisting of around 1.2 million labeled data with 1000 different classes. All the images in ILSVRC 2012 are quality-controlled and human annotated for the presence or absence of 1000 object categories.

C. Transfer Learning

Training an entire Convolutional Network from scratch with random initialization is difficult because it is relatively rare to have a dataset of sufficient size. Instead, it is common to pretrain a ConvNet on a very large dataset (in this case ImageNet, which contains 1.2 million images with 1000 categories) and then use the ConvNet either as an initialization or a fixed feature extractor for the task of interest.

The learned parameters are transferred to the task of sentiment prediction, where the images are from a different domain and the labeled data is limited. The strategy is to not only replace and retrain the classifier on top of the ConvNet on the new dataset, but to also fine tune the weights of the pretrained network by continuing the backpropagation. It is possible to fine tune all the layers of the ConvNet, and it is also possible to keep some of the earlier layers fixed due to over fitting concerns and only fine tune some higher level portion of the network.

The soft-max layer is removed while all the other internal layers of the pretrained CNN are kept fixed. We consider the activations from the 7th layer neurons as the image-level representation, which is a 4096 dimensional feature. The 7th layer output of pretrained CNN generalizes well for object recognition and detection.

Our work is motivated by the observation that the earlier features of a ConvNet contain more generic features such as edge detectors or color blob detectors, but later layers of the ConvNet becomes progressively more specific to the details of the classes contained in the original dataset. In case of ImageNet for example, which contains many dog breeds, a significant portion of the representational power of the ConvNet may be devoted to features that are specific to differentiating between dog breeds. Hence, the learning rate filter parameter of the convolutional layers are set while fine tuning so that the network trains with respect to the generic features. Also, the learning rate is initially assigned as 10^{-4} and it is increased by 10^{-1} after every 1000 iterations to make sure that it does not overshoot the minima.

D. Classifier

The work of Borth et al. [4] has proved that the logistic regression model leads to better performance than Support Vector Machine (SVM) classifiers for sentiment prediction. In this work, we employ logistic regression as the classifier on top of the generated features. For the 4096 dimensional activations, a logistic regression model is trained with each type of features.

TABLE I
RESULTS OF THE PROPOSED METHOD ON THE BENCHMARK FLICKR DATASET IN COMPARISON TO THE FC7 METHOD

Method	Prediction Accuracy
FC7 Method(Xu et al.[5])	49
Proposed Method	53.5

V. EXPERIMENTAL RESULTS

This section presents the experimental results of the proposed approaches presented in Section IV for the visual sentiment detection task.

The results showed that all the seven sentiments were being recognized by the network. On thoroughly implementing the network for different set of images, it was learnt that the network interprets the three extreme sentiments 1(depressed), 4(neutral) and 7(excited) with high probability. On the other hand, it fails to distinguish properly between sentiments 6(very happy) and 5(happy) and sentiments 2(very sad) and 3(sad). Also, the networks bemuses between sentiments 3(sad) and 5 (happy) and classifies some of the images with 0(neutral) rating.

We compare the performance of the baseline approach of FC7 method [5] with our proposed training strategy on the same Flickr dataset. From Table.I, it can be seen that the sentiment prediction accuracy of the most recent work [5] is 49% while the prediction accuracy of the proposed method is 53.5%. This shows the power of the suggested approach as a result of domain specific fine tuning. This observation can be explained by the fact that the prior approach suffers to be fine tuned on discrete dataset features and is not able to reach its full potential.

This set of results clearly demonstrates that the proposed visual sentiment prediction procedure is able to successfully utilize the power of the pre-trained Convolutional Neural Network by transferring domain knowledge from the image classification domain to the sentiment prediction domain, and by effectively utilizing the 4096 dimensional representations of the images in its sentiment prediction classifiers.

Therefore, the results provided in this section for the first time suggest that Convolutional Neural Networks are highly promising for visual sentiment analysis for domain specific fine tuning on small scale dataset which are non-identical to the large datasets, used for pretraining the network.

VI. Discussion

Domain specific fine tuning improves the performance of the CNN model; it outperforms the recent work by Xu et al. [5]. This improvement is significant given that we only use 806 images for domain adaptation. This suggests that the domain specific fine-tuning stage helps the model to find a better local minimum.

It is important to reiterate the significance of this work over the state-of-the-art works [1], [4], [5]. We are able to directly leverage a much smaller labeled data set for training. The smaller data sets, along with the proposed training strategy, gives rise to better generalizability of the trained model and higher confidence of such generalizability.

One of the limitations faced during the experimental work was working in the CPU mode for training the CNNs. When the training is done in CPU mode, it takes around 100 seconds for each iteration, while fine tuning in a GPU mode is faster. However, other than speed, it does not have any affects on the network prediction accuracy.

VII. Conclusion

Visual sentiment analysis is a challenging and interesting problem. While vast majority of previous works of sentiment analysis on social web were conducted on text, we propose to focus on the analysis of images, one of the dominant media types of online microblogging services. In this paper, we have used convolutional neural networks to solve this problem.

We have designed a new training strategy to overcome small scale training samples. To evaluate the proposed method on real-world data, we constructed a sentiment benchmark from the photo posts on Flickr, which has a rich repository of images and associated tags reflecting user's emotions. We have also introduced a 7-scale granularity of sentiment rating, which is more comprehensive compared with the bi-polar labeling scheme in the existing datasets that was used by Xu et al. [5].

Both progressive training and transfer learning inducted by a small number of confidently labeled images have yielded notable improvements. The experimental results suggest that convolutional neural networks that are properly trained on the suggested method can outperform for the highly challenging problem of visual sentiment analysis.

There are several interesting future directions for us to explore. First, we intend to adapt the CNN to the sentiment images with the user-tagged data of Flickr via semi-supervised learning. Furthermore, we would like to apply our research results to many applications in different domains, such as video gaming and election polls.

References

[1] 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*, ser. WISDOM '13. New York, NY, USA: ACM, 2013, pp. 10:1–10:8. [Online]. Available: http://doi.acm.org/10.1145/2502069.2502079

[2] Wikipedia, "Deep learning — wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Deep_learning&oldid=673293335, 2015, [Online; accessed 29-Jan-2015].

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[4] 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*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 223–232. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502282

[5] C. Xu, S. Cetintas, K. Lee, and L. Li, "Visual sentiment prediction with deep convolutional neural networks," *CoRR*, vol. abs/1411.5731, 2014. [Online]. Available: http://arxiv.org/abs/1411.5731

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, pp. 1–42, April 2015.

[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *CoRR*, vol. abs/1310.1531, 2013. [Online]. Available: http://arxiv.org/abs/1310.1531

[8] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1717–1724. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2014.222

[9] Wikipedia, "Flickr — wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Flickr&oldid=667389164, 2015, [Online; accessed 15-Feb-2015].

[10] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the International Conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 83–92. [Online]. Available: http://doi.acm.org/10.1145/1873951.1873965

[11] 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. 2544–2558, 2010. [Online]. Available: http://dx.doi.org/10.1002/asi.21416

[12] S. Jindal and S. Singh, "Manipal Image Sentiment Analysis Dataset," http://dx.doi.org/10.6084/m9.figshare.1496534, 07 2015.

[13] Wikipedia, "Rectifier (neural networks) — wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Rectifier_(neural_networks)&oldid=672864910, 2015, [Online; accessed 20-March-2015].

[14] 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 *Proceedings of the ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 675–678. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654889

[15] Y. Jia, "Caffe: An open source convolutional architecture for fast feature embedding," [Online]. Available: http://caffe.berkeleyvision.org/, 2013.