Image emotion classification using deep learning

A B. Tech Project Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of

Bachelor of Technology

by

Nikhil Agarwal
(130101054)

under the guidance of

Dr. Arijit Sur

to the

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI - 781039, ASSAM

CERTIFICATE

*This is to certify that the work contained in this thesis entitled "**Image emotion classification using deep learning**" is a bonafide work of **Nikhil Agarwal** (**Roll No. 130101054**), carried out in the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati under my supervision and that it has not been submitted elsewhere for a degree.*

Supervisor: **Dr. Arijit Sur**

Associate Professor,

May, 2019 | Department of Computer Science & Engineering,

Guwahati. | Indian Institute of Technology Guwahati, Assam.

Acknowledgements

I would like to thank my BTP guide Dr. Arijit Sur for guiding me through out my entire thesis project. I would also like to give a special thanks to Mr. Sathish Basavaraju for his availability and help. A final thanks to all those people who at any point helped me with my project.

Contents

List of Figures

We propose a new deep network method to classify images into emotions they may seem to generate within the viewer of the image. It seems very intuitive that not every part of the image is equally important in determining the emotion it conveys. Existing models have used the entire image to generate deep level features for this task. Learning the features only from the entire image may confuse the model. In this paper, we have proposed a model which learns to attend differently to different part of the images based on the context of the image. The crucial part here is that the machine learns it self on how to attend to images based on this context. That is, how much importance is to be given to what part is used effectively as a parameter to be learned by the model. Existing models either used art and psychology theories or have used deep learning without attention based mechanisms.

Using deep learning will be even more effective when we are introducing the attention mechanism into the model for this problem. That is because how much importance we are giving to different parts of the images can have an huge impact on how accurate we are in predicting the correct class of the image. Source of much scientific work in the field of deep learning has been achieved by trying to understand how human minds perceives and solves a problem. It is very clear that very certain parts of the image can summarize the entire image in terms of the emotions the image is trying to convey. Certain features like a smile, a cloud, a knife etc. can give very different emotions to the image.

Studies and researches in psychology have shown already that reflections of human emotion varies significantly corresponding to different stimuli of the visual nature. These researches have motivated computer scientists to estimate the emotional responses of humans provided a sequence of visual information. This has led to the development of blossoming research field known as affective image analysis, attention towards which has incremented in contemporary times. But the twofold challenges of subjectivity and complexity of affective level image analysis makes it tougher as compared to image analysis and semantic level.

In spite of this, the time and difficulty involved in development of features designed manually is the main cause that has led to scientists searching for automating techniques. Thus, the use of deep learning has emerged in the field of image emotion classification.

The advantage of using CNN is that it gives a framework to learn features automatically which can provide deep image representations. Brilliant work has been seen in a lot of tasks of visual recognition like segmentation and classification of images as well as detection of objects and recognizing scenes.

Prior Work and Limitations

Prior work on visual emotion classification may be more or less categorized in to approaches which are categorical [MH10, ZGJ$^+$14, PCSG15] and dimensional [NGP11, LSAJ$^+$12, BCL11]. Approaches which are dimensional map the emotions to a three or two D representation, whereas the categorical one classify the picture into one of some predefined classes of emotions which is more direct and hence people can understand this easier as compared to approaches which are dimensional.For this reason classification approach has gained more popularity. Yanulevskaya et al. [YvGR$^+$08] classified emotions on works of art based on features of Wiccest and Gabor and SVM. Lenz and Solli [SL09] made a descriptor of images relying on emotions based on colors which takes support from experiments done in psychology to categorize pictures. In [RXL$^+$16], features of emotions that are varying are taken from view that is both local and global to estimate emotion. Machajdik et al. [MH10] introduced a mixture of features made from hand relying on psychology and art which include variance in color, semantics of image and composition. Zhao et al. [ZGJ$^+$14] proposed an invariant and much exhaustive features which are visual to get insights about emotions of image designed in correspondence to principles of art. The effectiveness of these visual features picked by hand is limited to datasets that are small images of whom have been chosen from restricted domains like photos of art or paintings that are abstract

Organization of The Report

A literature review is done first to understand the problem better and get an insight into the existing methods. Some work outside this immediate domain has also been looked up which definitely played a role in coming up with the final architecture.

The model is proposed and explained along with the intuition behind it. Details regarding what loss function is used and why, variations within the implementation of the model etc. have been described in detail.

Sources of database and the process of acquiring, training and testing will be discussed. Results acquired after applying the model have been mentioned.

Lastly, we discuss the scope of improvement within the model. We also discuss about the problems faced and conclusions that can be drawn from the results.

Chapter 1

Literature Review

1.1 Current State of the Art method

The paper title ”Learning Multi-level Deep Representations for Image Emotion Classification” [RXX16] by Tianrong Rao, Min Xu,Dong Xu is currently the state of the art method for the task of image emotion classification.

Their basic idea is to use three different levels of image representations image semantics (meaning of the image and relations between different parts of the image), image aesthetics(rules regarding appreciation of beauty, solidarity etc. of the nature) and low level visual features through Multiple instance learning (MIL). Their architecture is called Mldr net (Multiple level deep representations) and unifies these different levels to deal with the noisy data that has been collected from over the Internet.

Their proposed method saw a 6% increase in testing accuracy on a dataset which consisted of both real life images and abstract art images.

1.1.1 Architecture of the model

Previous CNN based models used only a single CNN to get deep representations which does not take in to account all the different level factors. In an attempt to combine high level semantics, mid level aesthetics and low level visual features, three different CNNs were combined to propose the MIL. The emotion classes for this problem have been predefined namely Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear, Sadness making 8 classes.

To partition the picture to patches at threefold levels or scales, they have used pyramid segmentation to get varied deep level representations about the emotions related to the images. The CNN model they have used finally were determined experimentally. 1, 4 and 16 were the number of patches they used at the three different scales.

1.1.2 Convolutional Neural Network

Before formalizing the Mldr, we look into the CNN model. Provided with a training sample (x, y) CNN comes up with representations for each layer of the input images using the convolutional layers and the fully connected layers. The output is a single 1 X N vector of real numbers having unbounded range (where N is the number of classes). To map these real numbers in the range (0,1) we use the softmax function. The probability that the image belong to a certain class can be given by:

$$p_i = \frac{\exp h_i}{\sum_{i=1}^{i=N} \exp h_i}$$

where h_i is the value from the i^{th} channel of the output of the last fully connected layer.

The loss function related to this probability is the cross entropy function and can be defined as:

$$L = -\sum_i y_i \log p_i$$

where $y_i = 1$ for true class i and 0 for other classes.

Convolutional neural network models have a pecking order of filters and as we move a layer ahead, the representation becomes deeper. For the extraction of the medium level features of aesthetics of image and lower level visual features of images, the have used different versions of the Alexnet. These models will contain lesser number of Convolutional layers so that they do not extract very high level features instead of mid and low level features.

For extracting the higher level image semantics, the popular Alexnet [KSH12]is used. For the mid level and low level features also the CNNs are used which are a little less deep than the AlexNet and are its modifications only. Finally the average of these three outputs is taken to predict the class.

1.2 Inspiration: Hierarchical Attention Network

[YYD^{+}16] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, Eduard Hovy came up with Hierarchical Attention Network (HAN) for the problem of Document Classification. There model has two levels of attention namely word attention and sentence attention. The idea is to give more importance to words that play a major role in the semantic of the sentence and similarly give more weight to sentences that represent the document better. The key points that distinguishes HAN from previous works is that instead of filtering from the context words, the context of the sentence or the document is used to decide the importance of the words or the sentence.

Although the attention layer is the one which is of importance to us and not the entire model for proposing our model, we nonetheless have a quick look at the entire model with a little extra attention to the Hierarchical Attention.

1.2.1 Hierarchical Attention

A two level based attention is used. Firstly on the words and then secondly on the sentences. Let the i^{th} sentence be denoted as s_i (among a total of L sentences) and let it have T_i words. Then w_{it} is the t^{th} word in the i^{th} sentence where t belongs to some number between 1 to T_i. The model proposed by [BCB14] Bahdanau et. al does a projection of the Document in to the vector space. A classifier is build on these vector representations to categorize the documents. The following module explains how the document vector is obtained from the word vectors progressively by usage of the structure that is hierarchical.

The words are first encoded into their equivalent annotation vectors with the help of a bidirectional RNN. We are however not mentioning its details here.

1.2.2 Word Attention

The model assumes and correctly so that not all words are equally relevant to the context of a particular sentence and hence weights are learned corresponding to this relevance. A new vector representation of the sentence is formed by taking a weighted average of the word vectors where the weights are the one we learned through attention. The following equations are central to the attention mechanism.

$$u_{it} = \tanh(W_w h_{it} + b_w)$$
$$\alpha_{it} = \frac{\exp u_{it}^T u_w}{\sum_{it} \exp u_{it}^T u_w}$$
$$s_i = \sum_t \alpha_{it} h_{it}$$

First we give our annotation vector into a multi layered perceptron and use the first layer output which is a hidden representation of the annotation. Then we calculate a score of each hidden representation from a context vector u_w. This score is then normalized using softmax function to get the final weights of each word. These weights are then used to get the vector representation of the sentence. The context vector however is a variable and needs to be learned.

For the sentences also we apply RNN encoding followed by attention mechanism to get the document vectors just as we were getting sentence vectors using word attetion.

Chapter 2

Proposed method

Here we discuss about the model we have used on the problem of classifying images into emotions.

To recall, the problem statement is to classify images into one of the 8 classes which have the labels Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear, Sadness.

2.0.1 Input details

The input to the model is a batch of 50 images (or for that matter any n number of images). Each image is resized to 227*227 pixels before using it. Each pixel has information of 3 colors namely Red, Green and Blue and hence the input batch becomes a vector of dimension 227*227*3. We will cover the entire training dataset by running iterations till each image has contributed 50 times (or for that matter some bigger number n if possible). When one such cycle is completed and entire dataset has been used completely an epoch has said to be completed. We thus train our model for 50 epochs.

2.0.2 Output details

The output of our model is a 1*8 vector for each image of the batch (batch size for testing is also 50 but can be varied according to different requirements and limitations) where we have a score for each class which can be any real number. Hence the output is a vector of size 50*8. We apply softmax across each row of this matrix to scale the scores in the range 0 to 1 and so that their sum becomes 1. This scaled output is now ready to be compared with the expected output to give accuracy results and batch loss for training in case of training and batch loss for test in case of testing.

2.0.3 Loss function

Defining loss function appropriately is of utmost importance as the parameters of any model gets trained on it. We are using the cross-entropy loss which is the total negative log likelihood of scores assigned to known correct classes along all the samples.

The insight of the loss function is to punish bad results. The lesser the score of the correct class, the more will be it negative log and hence the summation captures the loss incurred on the entire batch.

2.1 Parent Model: The AlexNet

[KSH12] The AlexNet model was the winner of the 2012 ImageNet Image Classification task. We first revisist the AlexNet model in some detail. We then describe our actual model which combines AlexNet with Hierarchical Attention layer.

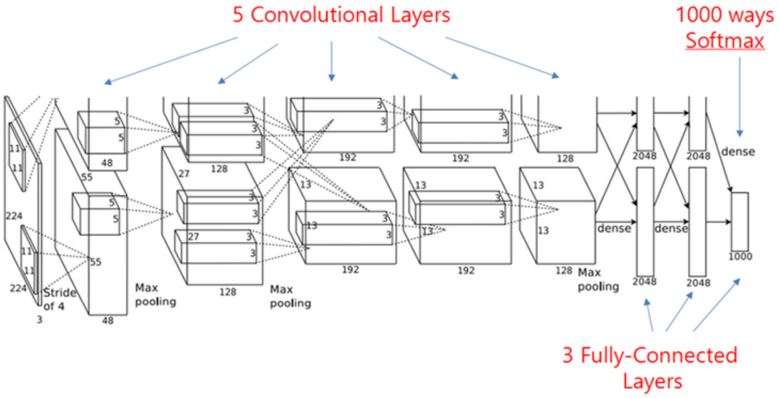

Fig. 2.1 Architecture of the AlexNet

AlexNet uses ReLu instead of classical tanh because of the faster training time involved with ReLu (maximum(0, x))

AlexNet also uses Overlap pooling with s = 2 which is the distance between two adjacent pooling regions and z = 3 which is the size of the area the max of which will be used. Overlap pooling has empirically proven to be less proven to the problem of over fitting.

2.1.1 The Architecture of Alexnet

The AlexNet consists of eighth layers which have parameters/weights, five of which are convolutional layers and three of which are fully connected layers. The last layer output is put into a 1000 way softmax to produce the class scores.

The first, second and fifth convolutional layer is followed by a max pooling layer about which we mentioned. However the non linearity ReLu is applied after every layer output be it convolutional layer or fully connected layer.

The first convolutional layer takes the input of size 224*224*3 and using 96 kernels of size 11*11*3 with a stride pf 4 pixels reducing its size to 55*55*96.

The second convolutional layer has the filter size as 5*5*48 and the number of kernels is 256.

The filter size of third convolutional layer is 3*3*256 and the number of kernels is 384. The fourth layer has filters of size 3*3*192 and the number of such filters also called the number of kernels is 384. The fifth layer has similar dimensions filter as of fourth layer of size 3*3*192 but the number of such filters reduces down to 256. There are 4096 neurons in each fully connected layer.

2.1.2 The problem of over-fitting: How to counter it?

This neural net model has 60 million parameters! This is just too many parameters to train even with the large dataset for ImageNet. To overcome the problem of overiftting, two methods are common:

Augmenting Data

One can use 5 or 10 patches of the same image from slightly different positions and train data on this increased data set. The testing is done by taking average of the 5 or 10 patches taken from the same original image. The problem with data augmentation is that it increases the training time by multiple times.

Dropout

By dropout rate of 0.5, with a probability of 0.5 each neurons value is mapped to zero and that neuron does not participate in backward propagation. This reduces the twisted correlations among different neurons preventing over-fitting. During testing time however, each neuron participates although there output is halved which corresponds to the dropout rate of 0.5. In case of a dropout rate of 0.7, this number would be 0.3.

2.2 Proposed Model

The base model we have used remains as AlexNet. Upto the first 5 convolutional layers the model is exactly same. The output of the 5th convolutional network (after max pooling) is given to an attention layer which also takes as input parameters the dimensions of the pooling layer output (apart from the batch size which is not a property of the pooling layer but instead specified by us). A series of reshape, transpose, multiplication, addition and softmax functions are applied to get the final values as the output. Basically, first the values are multiplied with a weight matrix, then a bias is added to it (equivalent to hidden layer of a MLP). This vector is then put into a softmax to get the attention weights. The weighted average using these attention weights is calculated using element wise multiplication of the attention weight matrix with the output of the pooling layer.These are then reshaped into 6*6*256 vector to be fed into the remaining layers of the AlexNet which are the fully connected layers. We only use this attetion AlexNet upto its pen ultimate layer. We apply the original AlexNet model concurrently on the same image batch upto its pen ultimate layer. We then feed the concatenation of these two sub networks to the final output layer. The final output layer is modified to give scores of 8 classes instead of 1000. The same softmax is then applied on this output vector along each sample of the batch and cross entropy(negative log likelihood) loss function is used to train the parameters.

2.2.1 The Attention Layer

The heart of our model lies in the attention layer. The attention layer gets as input a 6*6*256 sized vector. We can visualize this as 256 layers of 6*6 two dimensional matrix of values which are the feature maps. Let the point (i, j, k) denote the value in the i^{th} row and j^{th} column in the k^{th} of such 256 layers. If we keep i and j fixed and vary over k, we get a vector of length 256. This 256 length vector can be viewed as specifying 256 different qualities of the image or conceptual dimensions of the image at the immediate locality of the point (i, j). We have 6 x 6 or 36 such 256 feature vectors. We apply the attention model to learn 36 attention weights which represent the importance of each of our feature. We apply a softmax based single layer perceptron model to learn these 36 weights. We then do element wise multiplication of each of the 36 features along the 3_{rd} dimension to get a new 6*6*256 vector to be fed into the remaining network which has three fully connected layers. The maths associated with the learning model has been explained below:

$$H = W^T X + B$$

$$H = softmax(H)$$

$$X_i = H_i \cdot X_i$$

where i ranges from 1 to 36 along the 3rd dimension of our 6*6*256 matrix.

Our model learns the weight matrix W and the bias vector B which helps in calculating the attention weight matrix H. Thus we are not learning the weight matrix H directly. Instead we are learning matrices W and B which interact with the input X to produce H. Hence we can say that attention layer's output is based not only on the parameters directly learned by the model but also on the context of the image.

2.3 Conclusion

The idea of attention based classification takes into account the context of the image and how different parts of the image have different relevance to the problem. This specificity has been missing in existing methods and we are expecting our model to outperform the existing best method. Currently however the model is in its training phase and we don't have concrete results to support our work. The partial results have however been mentioned in the next section.

Chapter 3

Experimental Results

3.1 Dataset Used

The dataset was available through access to a meta data set. [YLJY16] have provided files which contain links to images and the name of the file which contains this link gives the emotion to which the linked images belong to. A special thanks to all those people who worked for making this dataset available.

The dataset is available in two forms: noisy dataset and well dataset.
The images which have been labelled with high confidence are referred as well dataset and the ones which have not been marked with at least 50% confidence constitute the noisy data set. We have distributed data into 2 parts. 80% for training and 20% testing for both noisy and well data set. Validation data set is not used though it can be used.

3.2 Results

The model achievhed a highest accuracy of **54%** on well dataset on partial training which is 5% less than the current state of the art of around 59% ans is expected to perform better with more training. The training on the noisy dataset is still running.

3.3 Conclusion

The idea of attention is intrinsic to humans. When we are exposed to any image, it is only the very crucial parts of the image which play a role in determining the emotion conveyed by that image. We have attempted to replicate this human nature in the form the attention layer introduced in our model.

Although the model is currently in the process of training, we are expecting it to give better results than the current state of the art accuracies of 45% on noisy dataset and 58% on well data set because on a very basic level it makes sense to attend differently to different part of the images.

Chapter 4

Conclusion and Future Work

The idea of using the attention layer to attend to images is interesting since not every part of the image is equally important in generating emotions within the viewer. The idea is intuitive and is an attempt to replicate the way in which a human perceives an emotion when they view an image.

Since the model is not trained and the partial results from partial training are not enough. However the results are good enough for us to expect excellent results after the model is extensively trained. We are hopeful to be able to get better results than the existing methods.

The approach of image attention based mechanisms might be extended to the domain of videos. Video emotions classification and emotional labeling or captioning of the videos can use the idea of Attention based neural network too.

References

[BCB14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[BCL11] Sergio Benini, Luca Canini, and Riccardo Leonardi. A connotative space for supporting movie affective recommendation. *IEEE Transactions on Multimedia*, 13(6):1356–1370, 2011.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[LSAJ$^+$12] Xin Lu, Poonam Suryanarayan, Reginald B Adams Jr, Jia Li, Michelle G Newman, and James Z Wang. On shape and the computability of emotions. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 229–238. ACM, 2012.

[MH10] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 83–92. ACM, 2010.

[NGP11] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. A multi-layer hybrid framework for dimensional emotion classification. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 933–936. ACM, 2011.

[PCSG15] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 860–868, 2015.

[RXL+16] Tianrong Rao, Min Xu, Huiying Liu, Jinqiao Wang, and Ian Burnett. Multiscale blocks based image emotion classification using multiple instance learning. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 634–638. IEEE, 2016.

[RXX16] Tianrong Rao, Min Xu, and Dong Xu. Learning multi-level deep representations for image emotion classification. *arXiv preprint arXiv:1611.07145*, 2016.

[SL09] Martin Solli and Reiner Lenz. Color based bags-of-emotions. In *Computer Analysis of Images and Patterns*, pages 573–580. Springer, 2009.

[YLJY16] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. *arXiv preprint arXiv:1605.02677*, 2016.

[YvGR+08] Victoria Yanulevskaya, Jan C van Gemert, Katharina Roth, Ann-Katrin Herbold, Nicu Sebe, and Jan-Mark Geusebroek. Emotional valence categorization using holistic image features. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 101–104. IEEE, 2008.

[YYD+16] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT*, pages 1480–1489, 2016.

[ZGJ+14] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 47–56. ACM, 2014.