Deep Attention Learning Mechanisms for Social Media Sentiment Image Revelation

Maha Alghalibi[1*], Adil Al-Azzawi[2, 3], Kai Lawonn[1]

[1] Computer vision Dept., University of Koblenz-Landau, Koblenz, Germany.
[2] EECS Dept., University of Missouri Columbia, Columbia, USA.
[3] CS Dept., College of Science, University of Diyala, Iraq.

* Corresponding author. Tel.:015779114824; email: mahaalghalibi@yahoo.com
Manuscript submitted January 10, 2019; accepted March 8, 2019.
doi: 10.17706/ijcce.2020.9.1.1-17

Abstract: Sentiment analysis systems can handle social media images by interpreting the embedded emotional responses in those images. This represents an interesting and challenging problem that tries to figure out the high-level content of large-scale visual data based on algorithms devised from computer vision. This paper presents a system to analyze social media images and visualize the implied emotions from each image as (Happy, Sad, and Neutral). The objective of this work is to introduce a system model with features extraction basis utilizing some adequate technique of machine learning. The applied methodology is pivoted on implementing the required system through several steps of processing. This involves social media image displaying and video frames grabbing, image features extraction, then embedded emotions patterns classification and recognition utilizing a proper convolutional neural network (CNN). Flickr and Twitter datasets were utilized while the pertinent algorithm was developed using "Matlab2017b" platform. This can help social media users visualizing their interests besides forming a better scope of visualization. It will further assist companies in envisaging the mood of users/costumers towards their stock prices in order to set competitive prices for both sides. We design a Deep Attention Network Mechanisms (DANM) to achieve a higher level of social media sentiment image analysis and classify them as (Highly positive mood and highly negative mood). The DANM produces features maps basis utilizing the adequate focusing technique of machine learning based on a proper convolutional neural network (CNN). The proposed CNN training system has proven better results with respect to accuracy and efficiency in comparison with some other similar works. When experimentations on both real and synthetic datasets were conducted, the system showed a percentile improvement of about 14.2%. This system is applicable to a broad horizon of applications such as studying the emotional response of humans on visual stimuli, visual sentiment analysis algorithms and modeling, building machine learning-based robust visual sentiment classifier, as well as in most online websites that involve visual data mining for business intelligence, e-commerce, stock market prediction, political vote forecasts, and video gaming.

Keywords: Deep learning, sentiment analysis, attention neural network, convolutional neural network, visualization.

1. Introduction

Sentiment visualization techniques have evolved and spread to deal with complex multidimensional data sets, including geospatial, temporal, and relational aspects. The issue of visual sentiment analysis in social

media involving images is hereby quite new and challenging.

As a matter of fact, images represent the easiest medium through which people can express their emotions on social networking sites. Social media users are thereby increasingly using images and videos to express their opinions and share their experiences. Sentiment analysis of such large-scale visual content can better help the user to extract sentiments towards events or topics such as those in image tweets. So that prediction of sentiments from visual content is complementing textual sentiment analysis.

Several recent works in this regard are there using initially pixel-level features, then mid-level attributes (change between neighboring pixels), and more recently deep-level visual features. This was reached through adopting some adequate computer vision algorithms acclimatized towards visual sentiment analysis utilizing some unsupervised ANN machine learning frameworks. Deep-level learning has made significant advances in tasks related to both vision and language. Consequently, the task of higher-level semantic understanding, such as machine translation, image aesthetic analysis, and visual sentiment analysis have become more amenable. A more interesting yet difficult task is to bridge the semantic gap between computer vision and visual sentiment analysis, and thereby helps in to solve more challenging problems. These orientated activities have thus achieved some reasonable performance in visual sentiment analysis.

Nonetheless, due to the complex nature of visual content, the performance of visual sentiment analysis is still unsatisfactory.

Furthermore, there are some other works on analyzing sentiments using multi-modalities, such as text and image. Late fusion is whereby employed to combine the prediction results of using n-gram textual features and mid-level visual features. More recently a cross-modality consistent regression (CCR) schemes have been proposed for joint textual-visual sentiment analysis. In fact, this approach is employing deep-level visual and textual features to develop a regression model. The successes of deep-level learning make the understanding and jointly modeling vision and visual content a feasible and attractive research topic.

In order to introduce a system model that is capable of analyzing social media images and visualizing the implied emotions from each image as (Happy, Sad, and Neutral) it is inevitable to resort to an adequate technique of 'Machine Learning'. Machine learning is an essential part of artificial intelligence where techniques and algorithms can be investigated with the aim of permitting computers to be trained. It is a procedure that replicates/simulates the manner in which the brain of human being functions, aiming to furnish computers with intelligence. Comprehensively revised approaches for machine learning often incorporate artificial neural network (ANN) with its most famous technique named support vector machine (SVM). For any machine learning models, the 'datasets' consist of two parts, the input part, and the output part. The output is often the features of attention, means the part that is aimed to be predicted or categorized, whereas the input is the set of constituents that might have effects upon the output. Machine learning tries to set correlations between the outputs with the input, through setting functions that approximate between the two parts with formulations yet to be estimated. Thereby, a 'convolutional neural network' (CNN, or ConvNet) is a class of deep-level, feed-forward artificial neural networks, most commonly applied in analyzing visual imagery. CNN's use a variation of multilayer perceptron's designed to require minimal preprocessing [1]. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights

Architecture and translation invariance characteristics [2], [3] CNN's were inspired by biological processes in that the connectivity pattern between neurons resembles the animal visual cortex organization [4]. Individual cortical neurons respond to stimuli only in a restricted zone of the visible field known as the receptive field. The receptive fields of different neurons partially overlap such that they lid the whole visual field. CNN's use relatively little preprocessing compared to other image classification

algorithms. This means that the network can be trained in a way which makes it similar to the filters that are hand-engineered in traditional algorithms. [5]

To presents a system that aims to analyze social media human images and visualize the implied emotions for each image, the presented system model inevitably needs to resort to some adequate techniques for features extraction and classification, followed by facial expression analysis. This practically means to obtain a set of measured data (samples) in the image under consideration. It is then required to derive values of some aspects in that image (called features) intended to be informative and non-redundant (i.e., to be discriminative information). So the feature is some valued aspect that would possibly be different among those samples in order to make a decision on it. Therefore, 'feature extraction' is the process of collecting that feature from a set of samples. Whereas, 'features classification' denotes the grouping of features based on some criteria like similarity (clustering) [6]. When the input data to an algorithm is huge to be processed, and it is suspected to be redundant (e.g., the repetitiveness of images presented as pixels), then it can be turned into a reduced subset of features called feature vector. Determining such a subset of the initially extracted features is called feature selection. The chosen features are expected to include the relevant information from the input data so that the required task can be performed by using this reduced representation instead of the complete initial data. Features classification is thus often related to features selection as this would optimize the machine learning algorithm and possibly assist noise removal of unrelated features. This will further facilitate the subsequent learning and generalization steps, and in some cases will lead to better human interpretations. Feature extraction is therefore related to dimensionality reduction; many machine learning practitioners believe that properly optimized features extraction is the key to effective model construction. Some very important areas of application hereby are computer vision, image processing, and machine vision where various features, such as the desired parts or shapes of a digitized image and video stream, can be isolated and detected using algorithms.

Facial expressions are the facial changes in response to a person's inner sentimental states, intentions, or social communications. 'Facial expression analysis' is an active research topic for behavioral scientists since the relevant work of Darwin in 1872. In 1978 an early attempt was done to automatically analyze facial expressions through tracking the motion of 20 identified spots on an image sequence [7]. Afterward, much progress has been made to build computer systems to help us understand and use this natural form of human communication. Facial expression analysis indicates to computer systems that try to analyze automatically and Computer systems rely on facial expression analysis to automatically analyze and understand both facial motions and facial feature changes from visual information. For sentiment or emotion analysis, higher-level knowledge is required. For example, although facial expressions can transfer emotion, they can also express intention, cognitive processes, physical effort, or other interpersonal expressions. Interpretation is assisted by context, body gesture, voice, individual differences, and cultural factors as well as by facial configuration and timing. Computer facial expression analysis systems need to analyze the facial actions in any case of context, culture, gender, and so on. The accomplishments in related areas such as psychological studies, human movement analysis, face detection, face tracking, and face recognition make the automatic facial expression analysis possible. It can be applied in many areas such as emotion and paralinguistic communication, clinical psychology, psychiatry, neurology, pain assessment, lie detection, intelligent environments, and multimodal human-computer interface (HCI).

This paper aims to articulate the significant sides of implementing an efficient-accurate system for image sentiment analysis and visualization based on the concepts of utilizing Attention convolutional neural network (ACNN) technique.

2. Related Works

The spectrum of related works in the pertinent arena involves pretty good participation. One of them is the work presented by You *et al*. (2017) regarding visual sentiment analysis through attending local image regions. They have studied the impact of local image regions on visual sentiment analysis. The proposed model utilizes the recently studied attention mechanism to jointly discover the relevant local regions and build a sentiment classifier on top of these local regions. Their model is capable of automatically discovering sentimental local regions of given images [1]. Jindal and Singh (2015) introduced their work about image sentiment analysis using deep-level convolutional neural networks (CNN) with domain-specific fine-tuning. In this work presented an image sentiment prediction framework that is maintained with a CNN. Specifically, this framework is pertained to large-scale data for object recognition to further perform transfer learning. Extensive experiments were proceeded on manually labeled Flickr image data. To make use of such labeled data, they employed a progressive strategy of domain-specific fine-tuning of the deep-level CNN [2]. Islam and Zhang (2016) published their work concerning visual sentiment analysis for social images using transfer learning approach. They used hyper-parameters learned from a very deep-level convolutional neural network to initialize the network model to prevent overfitting. They conduct extensive experiments on a Twitter image dataset [3]. You *et al*. (2015) exhibited some robust image sentiment analysis using progressively trained, and domain transferred deep-level networks. As motivated by the needs in leveraging large scale yet noisy training data to solve the extremely challenging problem of image sentiment analysis, they employed the convolutional neural network (CNNs). They have first designed a suitable CNN architecture for image sentiment analysis. They obtained half. By using a baseline sentiment algorithm, a million training samples have been used to label Flickr images; they employed a progressive strategy to fine-tune the deep-level network. Furthermore, they improved the performance on Twitter images by inducing domain transfer with a small number of manually labeled Twitter images [4]. Gupta and Gajarla (2016) published their work about emotion detection and sentiment analysis of images. The possibility of using deep-level learning to predict the emotion depicted by an image has been explored. Their results look promising and indicate that neural nets are capable of learning the emotions essayed by an image and in automatic tag predictions for images uploaded on social media websites [8]. Wang and Li (2015) revealed their work about sentiment analysis for social media images. They showed that neither visual features nor the textual features are by themselves sufficient for accurate sentiment labeling. Thus, they provided a way of using both of them, and formulate sentiment prediction problem in two scenarios: supervised and unsupervised. They developed an optimization algorithm for finding a local-optima solution under the proposed framework [5]. Yuan *et al*. (2015) presented their work regarding sentiment analysis using social multimedia. They introduced a comprehensive review of sentiment analysis based on visual content and textual content [6]. Jin *et al*. (2018) published their paper regarding a novel approach to analyze the facial expressions from images through the learning of a 3D morphable face model and a quantitative information visualization scheme for exploring this type of visual data. A 3D face database with various facial expressions was employed to build a nonnegative matrix factorization (NMF), part-based morphable 3D face model. From an input image, a 3D face with expression could be reconstructed iteratively by using the NMF morphable 3D face model as a priori knowledge. Whereby, basic parameters and a displacement map were extracted as features for facial emotion analysis and visualization. Based on the features, two support vector regressions were trained to determine the fuzzy valence–arousal (VA) values to quantify the emotions. The continuously changing emotion status could be intuitively analyzed by visualizing the VA values in VA space [7]. Kucher *et al*. (2018) presented their review paper about the state of the art in sentiment visualization. They presented a survey of sentiment visualization techniques based on a detailed categorization [8]. They described the background of sentiment analysis, and they further

introduced the categorization for sentiment visualization techniques [9]. Siersdorfer *et al*. (2010) presented analyzing and predicting the sentiment of images on the social web. They studied the connection between the sentiment of images expressed in metadata and their visual content in the social photo-sharing environment Flickr. They thereby considered the bag-of-visual-words representation as well as the color distribution of images and make use of the SentiWordNet thesaurus to extract numerical values for their sentiment from accompanying textual metadata. They, therefore, perform a discriminative feature analysis based on information-theoretic methods and apply machine learning techniques to predict the sentiment of images [10]. Chen *et al*. (2014) presented their work about predicting viewer affective comments based on image content in social media. While current studies are busy in analyzing visual effect, concepts intended by the media content publisher, their work, in contrast, focuses on predicting what viewer affect concepts (VAC) would be triggered when the image is perceived by the viewers [11]. Last but not least, Mandhyani *et al*. (2017) have introduced a novel model for image sentiment analysis. They proposed a model based on the mid-level features of the images that combines the techniques of SentiBank, CNN (Regions with CNN) and Senti Strength. Results of their extensive experiments conducted on Flickr image dataset showed that this approach achieved better sentiment classification accuracy [12].

3. Background Theory

Deep learning approaches have demonstrated incredible execution in computer vision and pattern recognition assignments. Deep learning allows automated learning of feature sets for particular problems alternatively of hand-crafted design. Convolutional Neural Network (CNN) example is one of the most popular types of deep learning methods [13]-[15] utilized in image processing. Convolutional neural network (CNN) can be regarded as a specific type of ANN which is of feed-forward basis. One of the characteristic features of this network is that It can 'learn'; permitting abstraction besides representation in several (multiple) levels. Actually, CNN is nothing but a perceptron of multi-layers; it is precisely composed of four distinct layers. The 1st is the input layer, the 2nd is the convolution layer, the 3rd is the downsampling layer, and the 4th is the output layer [15].

Furthermore, with respect to the architectural point of view, the second (convolution) layer and the third (downsampling) layer could, in turn, be composed of several (multiple) layers. As its structure is so simple, and due to its limited training parameters number, besides its adaptability, the architecture of the convolutional neural networks is widely preferred. One can find two main extensions of the general CNN version; namely the Region-based Convolutional Network (with acronym R-CNN) besides the Fully Convolutional Network (with the acronym FCN). FCN [16], [17] substitutes the fully connected layers in CNNs by full convolution layers and assigns class labels to each pixel in the image instead of one label per image block [17].

Assume that the network input which is presented by the $x_0 = x$, and the network outputs are, $x_1, x_2, \ldots, x_L$. Where each output $x_l = f_l(x_{l-1}; w_l)$ is deeply computed from the previous output $x_{1-1}$by applying the function f_l with the parameters of w_l [17]. The data flowing through the network represents a feature field; $x_l \in R^{H_l \times W_l \times D_l}$. Since the data x has a spatial structure, H_l and W_l are spatial coordinates, and D_l is the depth of channels. The function s the the f_l act as local and translation invariant operators, therefore, the network is called con; volitional [18]. CNNs are applied to distinguish between different classes by producing such as a vector of probabilities that denoted by $\hat{y} = f(x)$ for all tested image. If y is the true label of image x, CNN performance of true label y of image x is measured by a loss function $l_y(\hat{y}) \in R$ which assigns a penalty to classification errors [18].

3.1. Mathematical Approach of the Deep Convolutional Neural Network

Assume that we have some $N \times N$ square neuron layer which is followed by the convolutional layer. If we to use an $m \times m$ filter with kernel size $\omega \times \omega$, in this case, the convolutional layer output will be of size $(N-m+1) \times (N-m+1)(N-m+1) \times (N-m+1)$. In order to compute the pre-nonlinearity input to some unit $x_{ij}^{l}$the next layer that we need to sum up the contributions (weighted by the filter components) from the previous layer cells [18], [19]:

$$x_{ij}^{l} = \sum_{a=0}^{m-1} \sum_{b=0}^{m-1} w_{ab} y_{(i+a)(j+b)}^{l-1} \tag{1}$$

Then, the convolutional layer applies its nonlinearity [20]:

$$y_{ij}^{l} = \sigma(x_{ij}^{l}) \tag{2}$$

Also, pooling layer reduces the size of their input and allows multi-scale analysis. Carpooling and average-pooling are the most popular pooling operators. These operators calculate the highest or the average value within a small spatial block [17]. It has been deemed to consider the case as 'ideal' whenever the pooling is of a 2 × 2 filter size having a step (stride) of 2 [18]. Finally, the fully-connected layer connects to all the neurons of the previous layer. Fully connected layers are typically used as the last layer of the network and perform the classification [19].

3.2. Attention Learning Mechanisms

The Attention term is defined as a type of action that guide directly to the object. In other words, it is defined as "giving need" which is the ability mind to allocate the uneven consideration across a field of sensation [21]. Moreover, it helps to focus and bring certain input to the core of the attention. In the same, diminishing or ignoring the others [20].

Technically, in the neural network, the attention action helps in term of the credit assignment. The main challenge of that action is a long-range dependency. In another word, the prediction is to become more impacting and affected by other facts [22]. The core probability model of the attention network is based on the Markov Assumption [20], [21] which is aimed to introduce a model that consists of different probability numbers [23].

$$P(w_1 w_2 \dots w_n) \approx \prod_i P(w_i | w_{i-k} \dots w_{i-1}) \tag{3}$$

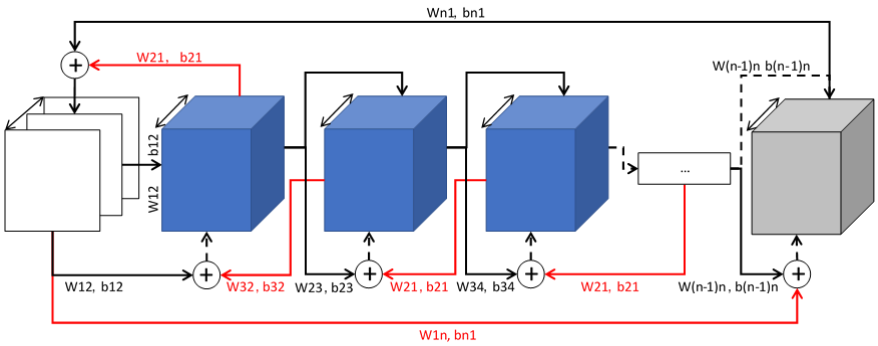


Fig. 1. The attention learning mechanisms. The red line shows the attention layer feed batch for each epoch to combine each weight and bias for each layer, as well the black lines illustrate the regular feed epoch during each iteration for the whole deep learning attention model [23].

At a high-level, the attention network enables the neural network to focus on relevant parts of your input more than the irrelevant parts when doing a prediction task. However, the attention network, in this case, can capture information in a human level [22].

Attention network mechanism is based basically on the sequence-to-sequence models which in this case the design model can capture the essence of the entire input sequence in a single hidden state as is shown in Fig. 1 [23].

4. Proposed System

Based on the complexity of the sentiment image analysis in the social network, we propose a Deep Attention model that bases on design a deep network that has the attention learning-based mechanism which we called Deep Attention Model for sentiment images analysis (DAM). The whole framework of the proposed network is illustrated in Fig. 2. We can notice that the main design of the deep network has been based on the attention mechanism feedback. In our design, the feed attention mechanism has two main parts. The first one is the feedforward attention stage and feed backward stage. In the first stage, the deep model learns and focusing on the high-level image features that will be extracted from the low-level image features. Those features acquired from the image details and the feedforward stage abstracts them to discriminative features. However, the follows stage (feed backward) try to acquire the low-level features in which the high-level features are returned to learned to extract more learned low-level features.

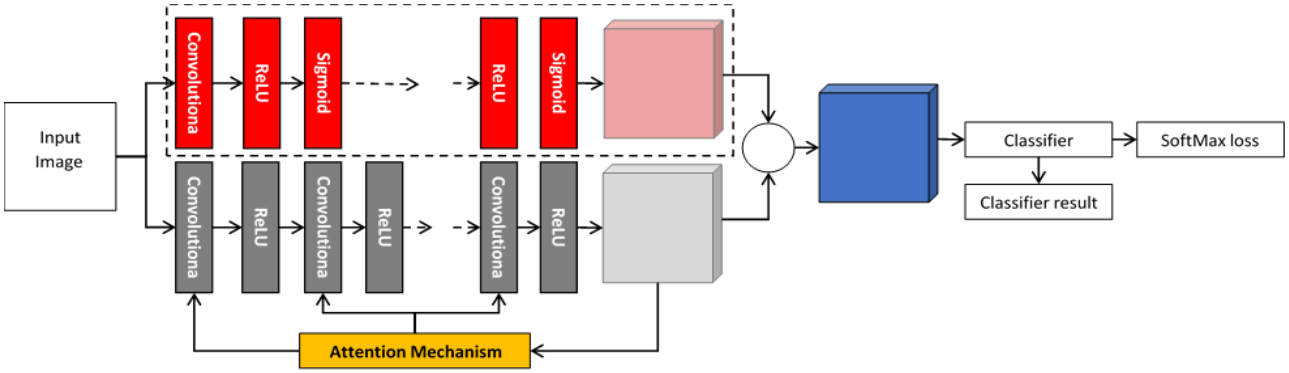


Fig. 2. System flowchart of the deep attention-mechanism for sentiment image classification.

In more details, the image features (high and low-levels) are extracted based on using a stack of CNN blocks (convolutional layers). The CNN blocks (red and gray) are stacked together in which features are comprises based on the convolution, pooling, and non-linear transformation. In another word, the red blocks in our design illustrate the original CNN blocks that are used mainly for high-level feature extraction during the first feature extraction stage (feedforward pass), while the gray blocks illustrate the second stage (feed backward). In more details, the gray blocks (CNN's) use the high-level features that are extracted from the red blocks (CNN's) and stack together based on the attention mechanism, convolution, pooling, and non-linear transformation.

5. Structure of the Deep Attention Network Model

The whole structure of our Deep Attention network has nine double layers in total (18 layers). The first feedforward structure has nine deep layers, in addition to the backward feed layers which also has a total of nine layers. The first five layers from our structure are convolutional layers followed by the next three fully connected layers. The SoftMax function is the main learning-based model that is used last fully connected layer. The main SoftMax function is given in Equation (4) [24].

$$L_S = -\frac{1}{m}\sum_{i=1}^{m} log\frac{e^{W_{y_i}^T f_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T f_i + b_j}} \quad (4)$$

The reason for using the SoftMax is that the SoftMax model can provide a significant distribution that is able to distinguish between the two classes (highly positive/highly negative) in our binary classification problem. The whole structure of our Deep Attention Model fundmantily maximizes the essential logistic objective of the multinomial regression in which equivalent to maximizing the log-probability by maximizing average attained of the logistic-function in the last fully connected layer. In this case, the performance of the Deep Attention network based on the across of the final distribution that is an arrangement for the final prediction which is done based on achieving the final labels of each problem class.

The original image input size that was feeding to the first layer is (224×224×3). Each image is 227 by 227 width and height and three channels since the dataset has colored images. The first convolutional layer of our Deep Attention model is constructed based on using 256 kernels the density features map that is constructed from the first layer is 5×5×48. The ReLU (non-linearity) scheme is applied to the output of the first layer (wholly-connected). The second convolutional is constructed based on using 256 kernels the density features map that is constructed from the first layer is 5×5×48. The output map from the first layer is 11×11×3 based on using stride 4 and padding 0.

Moreover, the second layer is the same first convolutional layer using the same block based on the attention mechanism (feed backward) convolutional layer. The second layer is another convolutional layer followed by the normalization layer (pooled and normalization layers). The output of the second convolutional layer is connected to the third convolutional layer base on using 384 kernels. Each kernel is 3 ×3×256. Then, the fourth convolutional layer has also 384 kernels of size 3×3×192 while the final fifth convolutional layer has 256 kernels. Each kernel size is 3×3×192.

The full structure of our proposed system (Deep Attention Model) for image sentiment analysis and classification are described in Table 1 below:

Table 1. Propose Deep Learning Structure Description

Layer Number	Layer Type	Ker.	Size	Description
I1	Image Input	-	227x227x3	images normalization
C1	Convolution	96	11x11x3	stride [4 4], padding [0 0 0 0]
R1	ReLU	-	-	ReLU
A1	Attention Model	1	-	Attention
N1	Normalization	-	-	normalization
P1	Max Pooling	1	3x3	stride [2 2] and padding [0 0 0 0]
C2	Convolution	256	5x5x48	stride [1 1], padding [2 2 2 2]
R2	ReLU	-	-	ReLU
A2	Attention Model	1	-	Attention
N2	Normalization	-	-	normalization
P2	Max Pooling	1	3x3	stride [2 2] and padding [0 0 0 0]
C3	Convolution	384	3x3x256	stride [1 1], padding [2 2 2 2]
R3	ReLU	-	-	ReLU
A3	Attention Model	1	-	Attention
C4	Convolution	384	3x3x192	stride [1 1], padding [2 2 2 2]

R4	ReLU	-	-	ReLU
A4	Attention Model	1	-	Attention
C5	Convolution	256	3x3x192	stride [1 1], padding [2 2 2 2]
R5	ReLU	-	-	ReLU
A5	Attention Model	1	-	Attention
P6	Max Pooling	1	3x3	stride [2 2], padding [0 0 0 0]
F7	Fully Connected	1	4096	fully connected layer
R7	ReLU	-	-	ReLU
A7	Attention Model	1	-	Attention
D7	Dropout	-	-	dropout
F8	Fully Connected	1	4096	fully connected layer
R8	ReLU	-	-	ReLU
A8	Attention Model	1	-	Attention
D8	Dropout	-	-	dropout
F9	Fully Connected	1	4096	fully connected layer
R9	ReLU	-	-	ReLU
A9	Attention Model	1	-	Attention
D9	Dropout	-	-	dropout

6. Experimental Results

6.1. Dataset

The data set contains over fifteen thousand sentiment-scored images on typical positive/negative sentiment. Data set contains URL of images, sentiment scores of highly positive, positive, neutral, negative, and highly negative, and contributor agreement. Some samples of the training and testing dataset are shown in Fig. 3 [25]:

Highly Negative Sentiment Image Samples

highly positive Sentiment Image Samples

Fig. 3. Samples of the sentiment images dataset, the first row shows some samples from the highly positive sentiment images, while the second row shows the highly negative image samples.

The whole dataset consists of 4000 images have been divided into 2000 images as a highly negative image and 2000 images as a highly positive image. The dataset is divided into 70% of images per category

to train (1399 images for training) and specify 30% as a validation set to test (601 images for testing).

Our network after it has been trained by specifying training options (parameters) as is shown in Table 2.

Table 2. Training Function Parameters

Function	Parameter
Training Function	Sigmoid Function
MiniBatchSize	10
MaxEpochs	6
Shuffle	'every-epoch
InitialLearnRate	1e-4
ValidationData	Used
ValidationFrequency	3
Max Epochs Number	20
MaxIterations Number	3360
Iteration per Epoch	168

The initial learning rate is set to a small value to slow down the learning. Also, specify the validation data and a small validation frequency. For fine-tuning, we want to change the network ever so slightly. The network is changed during training is constrained by the learning rates. Here we do not change the learning rates of the original layers, i.e., the ones preceding the last 3. The rates of these layers are already small, so they are not required to be decreased more. It is further possible to fix the weights of the early layers frozen through setting them all to zero. In this case, instead, we boost the learning rates of the new layers we added so that they change faster than the rest of the network. This way previous layers do does not change that much, and we quickly learn the weights of the newer layer.

6.2. Evaluation Criteria

The performance of the proposed framework can be evaluated using various parameters including classification accuracy, detection rate, and false positive rate, the given parameters True Positive (TP) which refers to correct detection of positive cases. True Negative (TN) which refers to the correct detection of negative cases. False Positive (FP) which refers to incorrect detection of positive cases into negative class. Finally, the False Negative (FN) which refers to the incorrect detection of negative cases into a class positive.

The evaluating performance of emotion detection system is calculated by using three measures called Recognition Rate (RR), Precision (PR), Sensitivity (SE), Specificity (SP) [26]. The formula for calculating these measures are given as in Eq. (5), (6), (7), and (8) respectively. The first performance result is the Recognition Rate which is defined as the ratio between the numbers of correct recognition decision to the total number of attempts as it is given in Eq. (5) [27].

$$Accuracy = \frac{TP}{TP + TN} * 100 \tag{5}$$

Secondly, the sensitivity is defined as the ratio between the numbers of retrieved prediction that are relevant to the number of retrieved detections as itis given in Eq. (6) [25]:

$$Precision = \frac{TP}{TP + TN} \tag{6}$$

Thirdly, the specificity is defined as the ratio between the numbers of true negative prediction and the total number of negative detection as it is given in Eq. (7) [27]:

$$Specificity = \frac{TN}{TN + FN} \tag{7}$$

Finally, the precision is defined as the ratio between the numbers of true negative prediction and the total number of negative detection as it is given in Eq. (8) [27]:

$$Precision = \frac{TP}{TP + FN} \tag{8}$$

6.3. Training Experimental Results

Fig. 4 shows the training accuracy and the lost function score. It is apparent that the loss score started from a higher score and described till reached the lowest loss score by achieving 10% loss score. However, the accuracy starts from the lower score 30% and keep increasing till reached the highest accuracy after consuming 150 iterations to reach the highest score which is almost 90% on the training dataset.

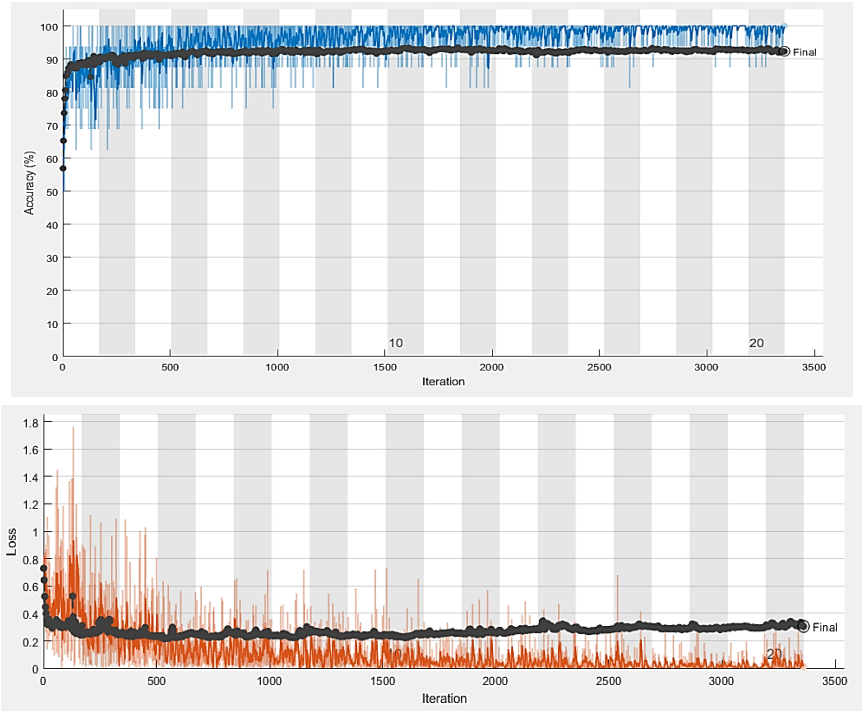


Fig. 4. The overall performance of the training accuracy and the lost function score during the training phase. The blue plot illustrates the training accuracy while the read plot illustrates the loss function, in both plots, the dash lines show the average training score and the average loss function for each epoch.

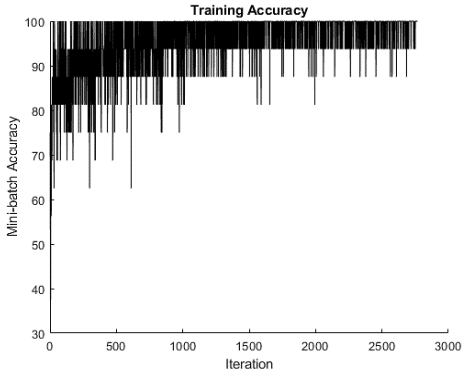


Fig. 5. Illustrates the mini-batch training accuracy for each iteration during the training phase.

Fig. 5 illustrates the mini-batch accuracy based on each iteration during the training step. Table 3 also shows the mini-batch accuracy achieved the highest accuracy of 100% and the lost score is very low by

reaching 0.0093. The different parameter has been tuned during the training phase, the most stable and the appropriate parameters.

Table 2 illustrates the performance of the training results based on reporting the accuracy and the loss function for each epoch. It shows that the total iterations for each epoch and the time consuming for each one. Although, it illustrates the mini-batch accuracy based on each epoch and the mini-batch loss function score. Finally, Table 2 illustrates fine-tuned parameters that have been used based on each epoch.

Table 2. Performance Results of the Training Accuracy and the Loss Function Score for each Iteration

Epoch	Iteration	Time Elapsed	Mini-match accuracy	Mini-batchLoss	Base Learning Rate
1	1-150	00:00:17	75.00%	0.5830	1.0000e-04
2	200-300	00:00:33	75.00%	0.6243	1.0000e-04
3	350-500	00:00:54	100.00%	0.0031	1.0000e-04
4	550-650	00:01:10	100.00%	0.0033	1.0000e-04
5	700-800	00:01:26	93.75%	0.1597	1.0000e-04
6	850-1000	00:01:47	93.75%	0.4944	1.0000e-04
7	1050-1150	00:02:04	100.00%	0.1246	1.0000e-04
8	1200-1300	00:02:20	93.75%	0.0479	1.0000e-04
9	1350-1500	00:02:43	100.00%	0.0800	1.0000e-04
10	4501550-1650	00:02:59	100.00%	0.0154	1.0000e-04
11	5001700-1800	00:03:15	93.75%	0.1471	1.0000e-04
12	1850-2000	00:03:39	100.00%	0.0073	1.0000e-04
13	2050-2150	00:03:56	100.00%	0.0075	1.0000e-04
14	2200-2350	00:04:19	100.00%	0.0220	1.0000e-04
15	2400-2500	00:04:38	100.00%	0.001	1.0000e-04
16	2550-2650	00:04:56	100.00%	0.0448	1.0000e-04
17	2750-2768	00:05:09	100.00%	8.3442e-05	1.0000e-04

6.4. Testing Performance and Experimental Results

Fig. 4 shows the training progress of the proposed model where the blue line illustrates training progress on the training dataset while the orange line illustrates the loss score on the training dataset. Total of 20 epochs has been used to train our model with a total of 3500 iterations. We can notice that the proposed system can reach the 90's during the first epoch and the lost function has been dramatically decreased in the same epoch.

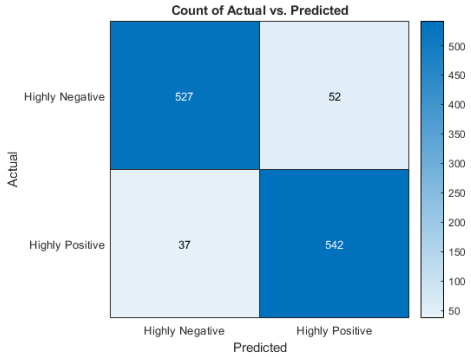


Fig. 6. The confusion matrix of the experimental testing result.

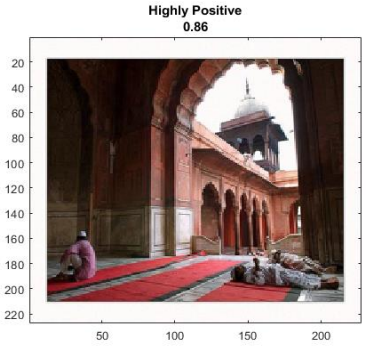
Fig. 6 illustrates the confusion matrix of the testing dataset. It is noticed that the proposed approach (Deep attention network) achieves 92.31% In contrast, the most recent approach for sentiment image classification using regular deep learning approach achieved 78.1% on the same dataset.

The whole performance results of the testing results are shown in Table 3. We can notice that the proposed system achieves 93.44% Sensitivity, 91.25%Specificity, and 91.02% precision. In the other hand, the proposed system achieves 93.61% on the negative predictive, 87.50% on the false positive rate of 89.80% on the false discovery rate, 65.60% on the false-negative rate. Finally, the deep attention network achieves 92.31% accuracy, 92.21% Fi-score measurement, and 84.66% Matthews Correlation Coefficient.

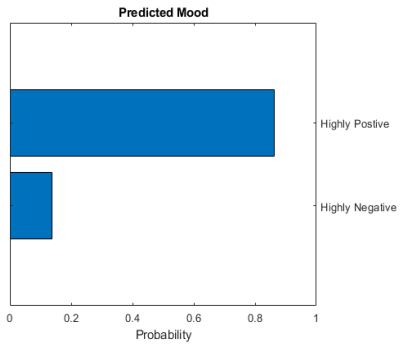
Table 3. The Experimental Results on the Testing Dataset

Measure	Value
Sensitivity	0.9344
Specificity	0.9125
Precision	0.9102
Negative Predictive Value	0.9361
False Positive Rate	0.0875
False Discovery Rate	0.0898
False Negative Rate	0.0656
Accuracy	0.9231
F1 Score	0.9221
Matthews Correlation Coefficient	0.8466

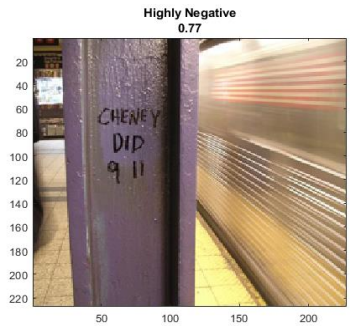
Fig. 7 shows some random examples that have been randomly selected from the testing dataset. In this case, the proposed approach can predict a confident prediction score to assign the image to the correct label. It is also showing that some cases have less confident scores than the other based on the color variation and the complexity of the tested images. Fig. 7 also shows the high protective probability for each testing image that has randomly chosen to form the testing dataset. It also shows the correct label prediction as well as the final score compared with the other label and its prediction score too. Some testing image cases got very high confident score by predict the correct label using 1 or 100% confidante score as is shown in Fig. 7 (e) and (f) also in Fig. 7 (g) and (h) while some other cases get between the 60's and 80's as is shown in Fig. 7 (a) and (b), also in Fig. 7 (c) and (d).



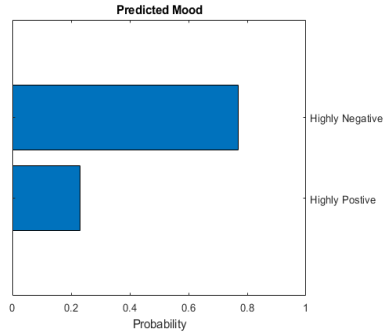
(a)



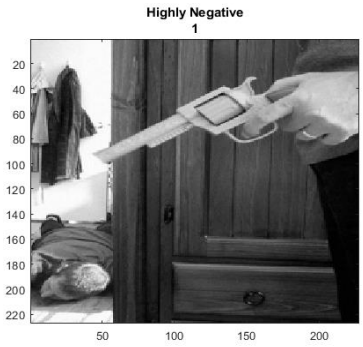
(b)



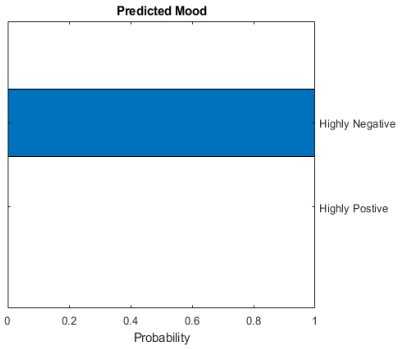
(c)



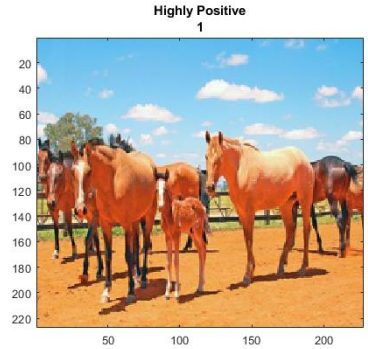
(d)



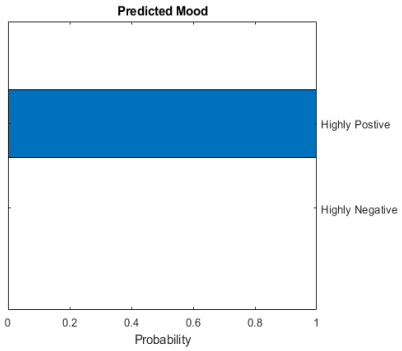
(e)



(f)



(g)



(h)

Fig. 7. Some examples of the sentiment analysis testing experimental results using deep attention network.

7. Conclusion

Sentiment analysis is a challenging problem that tries to figure out the high-level content of large-scale visual data based on algorithms devised from computer vision. In this paper, we design a Deep Attention Network Mechanisms (DANM) to achieve a higher level of social media sentiment image analysis and classify them as (Highly positive mood and highly negative mood). The DANM produces features maps basis utilizing the adequate focusing technique of machine learning based on a proper convolutional neural network (CNN). The proposed network presents a higher accuracy and efficiency in the performance results by achieving 92.31%higher than the most recent work by achieving 78.1% that has been tested in the same dataset.

Conflict of Interest

There is no conflict of interest.

Author Contributions

Maha Alghalibi conceived, planned and carried out the idea and the experiments. Adil Al-Azzawi. Contributed to the interpretation of the results and wrote the manuscript with support from Maha Alghalibi. Kai Lawonn supervise the project.

Acknowledgment

This work was partially supported by Ministry of Higher Education and Scientific Research (MHESR), Iraq, University of Koblenz Landau, Germany, HCED (Higher Committee for Education Development in Iraq).

References

[1] You, Q., & Jose, S. (2012). Visual sentiment analysis by attending on local image regions. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)* (pp. 231–237).

[2] Jindal, S., & Singh, S. (2015). Image sentiment analysis using deep convolutional neural networks with domain-specific fine tuning. *Proceedings of 2015 International Conference on Information Processing (ICIP)* (pp. 447–451).

[3] Islam, J. (2016). Visual sentiment analysis for social images using transfer learning approach. *Proceedings of 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom)* (pp. 124–130).

[4] You, Q., Luo, J., Jin, H., & Yang, J. (2013). Robust image sentiment analysis using progressively trained and domain transferred deep networks. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[5] Wang, Y. (2015). Sentiment analysis for social media images. *Proceedings of 2015 IEEE International Conference on Data Mining Workshop (ICDMW)*.

[6] Yuan, J., You, Q., & Luo, J. (2015). Sentiment analysis using social multimedia. *Multimedia Data Mining and Analytics*. Switzerland: Springer International Publishing Switzerland.

[7] Jin, H., Wang, X., Lian, Y., & Hua, J. (2018, Feb.). Emotion information visualization through the learning of 3D morphable face model. *Vis. Comput.*

[8] Gajarla, V. (2015). Emotion detection and sentiment analysis of images. Georgia Institute of Technology.

[9] Kucher, K., Paradis, C., & Kerren, A. (2017). The state of the art in sentiment visualization. *Computer Graphics Forum, 37(1)*, 71–96.

[10] Siersdorfer, S., Minack, E., Deng, F., & Hare, J. (2010). Analyzing and predicting sentiment of images on the social web. *Proceedings of 18th ACM Int. Conf. Multimedia* (pp. 715–718).

[11] Chen, Y., Chen, T., Hsu, W. H., Liao, H. M., & Chang, S. Predicting viewer affective comments based on image content in social media categories and subject descriptors. *Proceedings of the 2016 ACM Multimedia Conference*.

[12] Mandhyani, J., Khatri, L., Ludhrani, V., Nagdev, R., & Sahu, P. S. (2017). Image sentiment analysis. *Int. J. Eng. Sci. Comput., 7(2)*, 4566–4569.

[13] Ciresan, D., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in Neural Information Processing Systems*, 2843–2851.

[14] Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., & Shen, D. (2015). Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *Neuro Image, 108*, 214–224.

[15] Reddick, W. E., Glass, J. O., Cook, E. N., Elkin, T. D., & Deaton, R. J. (1997). Automated segmentation and classification of multispectral magnetic resonance images of the brain using artificial neural networks. *IEEE Transactions on Medical Imaging, 16(6)*, 911–918.

[16] Tajbakhsh, N., Gotway, M. B., & Liang, J. (2015). Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks. *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 62–69). Springer.

[17] Wang, H., Cruz-Roa, A., Basavanhally, A., Gilmore, H., Shih, N., Feldman, M., Tomaszewski, J., Gonzalez, F., & Madabhushi, A. (2014). The cascaded ensemble of convolutional neural networks and handcrafted features for mitosis detection. *Proceedings of the SPIE Medical Imaging* (pp. 90410B–90410B). International Society for Optics and Photonics.

[18] Cruz-Roa, A. A., Ovalle, J. E. A., Madabhushi, A., & Osorio, F. A. G. (2013). A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 403-410).

[19] Al-Azzawi, A., Al-Sadr, H., Cheng, J., & Han, T. X. (2018). Localized deep norm-CNN structure for face verification. *Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Orlando, FL, USA.

[20] Choi, J., Lee, B.-J., & Zhan, B.-T. Multi-focus attention network for efficient deep reinforcement learning. Association for the Advancement of Artificial Intelligence.

[21] Tang, G., Sennrich, R., & Nivre, J. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. *Proceedings of the Third Conference on Machine Translation (WMT): Vol. 1. Research Papers* (pp. 26–35). Belgium, Brussels, October 31 - November 1, 2018.

[22] Marijn, F., Stollenga, J. M., Faustino, G., & Juergen, S. (2014). Deep networks with internal selective attention through feedback connections. *Proceedings of the 27th International Conference on Neural Information Processing Systems: Vol. 2*. (pp. 3545-3553).

[23] Kim, Y., Denton, C., & Rush, L. H. A. M. (2017). Structured attention networks. *Proceedings of the ICLR 2017*.

[24] Al-Azzawi, A., Hind, J., & Cheng, J. (2018). Localized deep-CNN structure for face recognition. *Proceedings of the 2018 11th International Conference on Developments in eSystems Engineering (DeSE)*.

[25] Image Sentiment Polarity. Retrieved from https://data.world/crowdflower/image-sentiment-polarity

[26] Convolutional neural networks for visual recognition. Retrieved from http://cs231n.github.io/convolutional-networks/

[27] Xu, R., Tao, Y., Lu, Z., & Y. Zhong. (2018). Attention-mechanism-containing neural networks for high-resolution remote sensing image classification. *Remote Sens.*, *10(10)*, 1602.

Copyright © 2020 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).

Maha Al-Ghalibi was born in Basrah, Iraq. She obtained her bachelor degree in 2004 in computer science from Basrah University. In 2012 she did master in computer science from BAMU University, India. Currently she is a PhD student at the Faculty of Computer Science, Institute of Computational Visualistics, University of Koblenz-Landau. Her primary interests are visual analytic, big data visualization.

Adil Al-Azzawi was born in Diyala in 1978. He received the BS.C degree in software engineering from University of Baghdad (2001) and the high diploma from Iraqi Institute for Computer and Informatics (2002), the MS.C degree in computer science from University of Technology (2005) and is currently earning the Ph.D. in Electrical Engineering and Computer Science Department (EECS), University of Missouri-Columbia. His field of specialization is machine learning (ML), deep learning, bioinformatics, biometric, and computational intelligence (CI). He is an assistant professor (2012) at University of Diyala–Iraq. Assist. Prof.

Al-Azzawi is a member of Iraqi Engineers Union, member of Iraqi Teachers Union, and member of Iraqi Association for Information Technology.

Kai Lawonn was born in 1985 in Berlin, Germany. He received his diploma in mathematics from the Freie Universita t Berlin. In 2014 he did his PhD at the University of Magdeburg. Currently he is an assistant professor at the University of Koblenz-Landau.