2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)

XRA-Net framework for Visual Sentiments Analysis

Ashima Yadav[1], Ayush Agarwal[2], Dinesh Kumar Vishwakarma[3]
Biometric Research Laboratory, Department of Information Technology
Delhi Technological University, Bawana Road
New Delhi, India
ashimayadavdtu@gmail.com[1], ayush286@gmail.com[2], dvishwakarma@gmail.com[3]

***Abstract*—The exponential growth of social media has motivated people to express themselves in various forms. Visual media is one of the most effective and popular ways of conveying sentiments or opinions on the web as people keeps on uploading millions of photos on famous social networking sites. Hence, Visual Sentiment Analysis is instrumental in monitoring an overview of the broader public consensus behind a specific topic or issue. This work proposes a deep learning-based architecture XRA-Net (Xception Residual Attention based Network) for visual sentiment analysis. Moreover, the performance of the XRA-Net architecture is evaluated on the publicly available real-world Twitter I dataset, which is further composed of three subsets of the dataset: 3-agree, 4-agree, and 5-agree. The accuracy achieved on these datasets are: 79.2%, 81.2%, and 86.4% respectively, which shows that the proposed architecture has outperformed the state-of-the-art results on all the three subsets of Twitter I dataset as it can focus on the most informative features in an input image, which boosts the visual sentiment analysis process.**

Keywords- Attention, CNN, Deep Learning, Twitter, Visual Sentiment Analysis.

I. INTRODUCTION

The programmed process of expressing an opinion regarding a proffered text is called Sentiment Analysis. Sentiment Analysis also referred to as Opinion Mining, is a division in Natural Language Processing (NLP) that builds methods which try to recognize and extract opinions in the text. Moreover, we see that on a daily basis, 300 million photos are uploaded per day. Hence, extracting the sentiments from the photos uploaded by users becomes a crucial task. This process is known as Visual Sentiment Analysis (VSA). The visual sentiments can be in the form of images, videos, or emoticons, which can strengthen the opinion expressed by the textual modality. Thus, VSA can serve as a crucial process for making sense of the information.

Deep learning is one of the most promising fields which is gaining a vast amount of public attention [1] [2] [3]. When there is an absence of domain knowledge for feature creation, deep learning techniques surpass others as you have to bother less about feature engineering. Deep Learning algorithms try to determine high-level features from data in an incremental way. This eliminates the need for domain expertise and hardcore feature extraction process, which is the most crucial part for the machine learning based methods. Convolutional Neural Networks (CNN), which are one of the most popular deep learning based architecture has shown outstanding results in the field of computer vision for tasks like person re-identification [4], object detection and localization [5], action recognition [6], etc. Hence, motivated by this idea, we have applied deep learning based architecture, CNN for the process of VSA. The significant contribution of our work can be summarized as follows:

- We have proposed a deep learning based architecture XRA-Net (Xception Residual Attention based Network) for VSA by extracting the visual sentiments from the images.
- Further, we show the importance of Attention mechanism for VSA by applying the Residual Attention to focus on the most prominent areas in the images to get the finer sentiments.
- To evaluate the performance of the XRA-Net architecture, we have conducted experiments on the publicly available, real-world, and challenging dataset of Twitter I [7].

The remaining script is organized as follows: Section II summarizes the related work, Section III discusses the proposed XRA-Net architecture, Section IV describes the experimental settings, Section V concludes the script and discusses the future work.

II. RELATED WORK

This section discusses the previous work in the field of VSA. The entire work can be categorized into three major categories: low-level feature-based approaches, mid-level feature-based approaches, and deep learning based approaches.

A. Low-level feature-based approaches

The initial attempt for VSA focuses on low-level feature-based approaches. Siersdorfer *et al.* [8] applied Global and Local RGB Histogram along with SIFT-based bag of visual terms to extract the visual features from half a million Flickr images. The images contained textual metadata which describes the sentiment expressed in them. Initially, numeric sentiment values are assigned to the images based on the textual metadata. Later, this value is combined with the low-level features (Global and Local RGB Histogram along with SIFT-based Bag of Visual Terms) and passed into SVM for the final classification of the images. Vonikakis *et al.* [9] presented a system to create slideshow from family photo collections by focusing on their emotions and other attributes like colors in the picture, the time when it was clicked, and Gist of the scene. The experimental results obtained by combining these attributes show that the system can categorize the photos based on different emotions and could also filter out unwanted emotions. Li *et al.* [10] proposed a method using Bilayer Sparse Representation (BSR) for emotion classification in images. Further, the BSR contains two layers: Global Sparse Representation (GSR) which uses

978-1-7281-5527-2/19/$31.00 ©2019 IEEE
DOI 10.1109/BigMM.2019.00040

global features like Gabor texture feature, HSV color histogram, Bag of words, and Local Sparse Representation (LSR) which segments the image into uniform regions, and for every region color emotion, Gabor text, and HSV color feature is applied.

B. Mid-level feature-based approaches

Low-level features are not able to capture the sentiment expressed in the images properly. Hence, to fill this gap, mid-level features are extracted from the image. Borth *et al.* [11] proposed SentiBank, which provides mid-level visual representation by providing Adjective Noun Pairs (ANP) corresponding to each image. This response is fed as an input feature to the classifier. The authors have used Linear SVM and Logistic Regression as the classifiers. Experimental results prove that the mid-level features have shown better performances in comparison to the low-level features. Jou *et al.* [12] developed the Multilingual Visual Sentiment Ontology (MVSO) containing ANP from multiple languages. The experimental results demonstrate that emotions are not necessarily culturally universal. Yuan *et al.* [13] proposed a novel algorithm Sentribute, which extracts low-level features from visual contents for generating 102 mid-level attributes. They also used the Eigenface emotion detection approach as an added mid-level attribute. Finally, decision fusion is performed to predict the overall sentiment of an image. Zhao *et al.* [14] extracted features from the images based on Principles-of-art Emotion Features (PAEF), which includes six major features: Movement, Symmetry, Harmony, Variety, Emphasis, and Gradation. Finally, the PAEF features are applied to the classifier to predict the final sentiment score. The experiment was performed on the International Affective Picture System, which signifies that the approach gives effective results. Chen *et al.* [15] used Sentibank on Flickr for extracting the 1200 publisher affect concepts from the images and 446 viewers affect concepts from the Flickr comments.

C. Deep learning-based approaches

The increasing demand for deep learning architectures like CNN in computer vision has motivated many researchers to employ such techniques in the area of VSA. You *et al.* [7] trained CNN on a subset of Flickr images and calculate a prediction score on the training data. The trained model is fine-tuned with new training instances, and this new model is used for final classification. The experimental results showed that fine-tuned CNN could outperform low-level features and mid-level attributes. Recently, multimodal sentiment prediction has also gained a lot of interest in which researchers have focused on more than one modality. For example, [16] proposed a Weakly Supervised multimodal deep learning (WS-MDL) model for predicting sentiment from microblogs consisting of image, text, and emoticon modalities. The model includes the following steps: Initially, the model considered the emoticons as the weak labels and used them for training to infer the label noise. Next, CNN (AlexNet) is employed for VSA, and Deep CNN is used for textual sentiment analysis to calculate the prediction scores. The Expectation-Maximization algorithm optimizes the parameters of DCNN and CNN. The results justify that the model has achieved excellent performance for multimodal sentiment prediction. Yang *et al.* [17] developed a framework for learning the affective regions in the images. First, CNN is employed to compute the sentiment score of the image by considering local and global details of the input image. Next, bounding box candidates are generated around every image, and objectness scores are computed. Finally, the sentiment and objectness scores are combined to create the affective regions.

D. Visual Attention

In VSA, the image might contain a lot of irrelevant information which could interfere with the network performance. Hence, attention-based mechanisms are popularly used in computer vision on image-based datasets for avoiding such problems. Li *et al.* [18] introduced a new architecture that uses 3D attention model, which can detect vital elements in a video by strengthening them with large weights for video highlight detection. Hence, unique features from the significant video segments can be fetched to get the highlight score. The 3D attention mechanism increased the accuracy for video highlight detection by retrieving the highlights in spatial and temporal dimensions.

Attention networks are also applied for fine-grained classification. Zhao *et al.* [19] used attention networks to explore the range of attention for fine-grained object classification. They integrated the attention mechanism with LSTM networks. Letarte *et al.* [20] discussed the importance of self-attention systems for sentiment analysis. The network modeled the interactions between all the input word pairs. The authors have used Glove embedding for encoding each word in the sequence. Similarly, Wang *et al.* [21] applied LSTM with attention for aspect-based sentiment analysis so that the model could concentrate on various parts of the sentence. Recently, the attention-based network has also gained popularity for VSA. You *et al.* [22] proposed a framework to fill the gap between visual and textual language using tree-based LSTM model for visual-textual sentiment analysis. They applied attention model so that the model is not controlled by a single modality and can grasp the alignments between the image region and descriptive words. Inspired by the positive results achieved by the attention mechanism, we proposed an architecture which applies Residual attention for the VSA.

III. Proposed Methodology

This section describes the proposed architecture for VSA, as shown in Fig. 1. Our XRA-Net architecture can be divided into two major units: Xception based CNN unit and Residual Attention unit.

A. Xception based CNN Unit

VSA can be categorized as an image classification problem. Moreover, the dataset for VSA is not large enough for training the CNN models. Hence, we have applied the fine-tuning process by initializing the network using the weights of the pre-trained model. The CNN architecture that we have

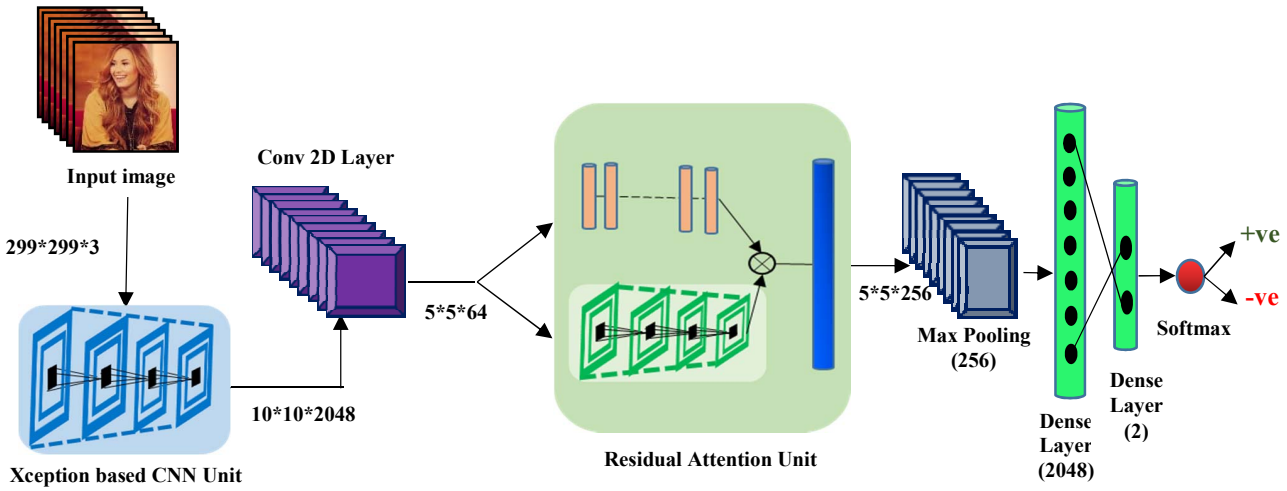

Fig. 1. Underlying architecture of the proposed XRA-Net framework

used for the fine-tuning process is the Xception [23]. The prime motivation for applying this architecture is that it has achieved state-of-the-art results on the ImageNet classification task and has outperformed the results achieved by the famous InceptionV3 architecture. The Xception architecture is composed of 36 deep convolutional layers which are stacked together, having residual connections around them. Hence, the CNN layer helps in getting the input image representation by extracting the features corresponding to every region in the input image. We feed the input image with 299*299 resolution. The Xception units help in getting the input image features, with the output of the last convolutional layer as 10*10*2048. Finally, the output of the Xception unit is passed into a Conv2D layer to get the output shape of 5*5*64.

B. Residual Attention Unit

The word "Attention" refers to the ability of the mind to focus on certain distinguish regions while ignoring the unimportant areas. In computer vision, attention has served as one of the crucial parts for image classification problem. Hence, there exists a need to model the role of attention for VSA. Inspired by the applications of attention for various image classification tasks, we have applied Residual Attention for the task of VSA. The residual attention proposed by [24] is built by stacking multiple attention modules. Since the size of our dataset is small; we observed that by applying all the attention modules, the overall architecture stopped learning, as the attention modules were capturing no new features. Hence, we have implemented only one attention module, followed by end-to-end training the network. The attention module of the residual attention unit is composed of two branches: the trunk branch and the mask branch. The trunk branch is used for feature engineering process, and the mask branch prevents the trunk parameters from getting updated by the wrong gradients. They are used for picking useful features in the image. Thus, the mask branch learns the attention-aware features in the image, which helps in getting the global information about the whole image.

The key component in our attention mechanism is the mask branch *M(x),* as they work as feature selectors which helps in extracting the relevant features and reduce the noise from the trunk features. The output of the attention $H_{p,s}(x)$ is shown in Eq. (1) below:

$$H_{p,s}(x) = \left(1 + M_{p,s}(x)\right) * T_{p,s}(x) \quad (1)$$

Where *p* varies over all the spatial positions, *s* ε *{1,2,…,s}* denotes the channel index, *M(x)* varies from [0.1]. If *M(x)* goes to 0, then *H(x)* will approximate to *T(x)* which are the original features. This idea is called residual attention learning.

C. End-to-end Training

In the proposed XRA-Net architecture, the Xception based CNN unit helps in getting the input image representation where the output shape (5*5*64) of the Conv2D layer is passed into the residual attention unit for getting the attention-aware features by discarding the irrelevant information about the image. The output of the last layer of the residual attention unit is passed into the two dense layers. The first dense layers contain 2048 neurons, and the final dense layer has two neurons for the two classes, namely positive and negative. Finally, softmax classifier is used for binary (positive or negative) sentiment classification.

IV. EXPERIMENTAL RESULTS

This section discusses the dataset details, experimental settings, and compares the proposed architecture with state-of-the-art results.

A. Dataset details and Experimental Settings

To evaluate the performance of the proposed XRA-Net architecture, we have used publicly available, real-world, challenging dataset of Twitter I which is further composed of three subsets of dataset namely: 3-agree, 4-agree and 5-agree.

TABLE III
BASELINE COMPARISON OF THE XRA-NET ARCHITECTURE ON ALL THE THREE DATASETS

Previous work	Methods	3-agree				4-agree				5-agree			
		P	R	F1	ACC	P	R	F1	ACC	P	R	F1	ACC
[8] (2010)	GCH	0.678	0.836	0.749	0.66	0.687	0.84	0.756	0.665	0.708	0.888	0.787	0.684
[8] (2010)	LCH	0.716	0.737	0.726	0.664	0.725	0.753	0.739	0.671	0.764	0.809	0.786	0.71
[8] (2010)	GCH + BoW	0.683	0.835	0.751	0.665	0.703	0.849	0.769	0.685	0.724	0.904	0.804	0.71
[8] (2010)	LCH + BoW	0.722	0.726	0.723	0.664	0.751	0.762	0.756	0.697	0.771	0.811	0.79	0.717
[25] (2013)	SentiBank	0.720	0.723	0.721	0.662	0.742	0.727	0.734	0.675	0.785	0.768	0.776	0.709
[13] (2013)	Sentribute	0.733	0.783	0.757	0.696	0.75	0.792	0.771	0.709	0.789	0.823	0.805	0.738
[7] [2](2015)	CNN	0.734	0.832	0.779	0.715	0.773	0.855	0.811	0.755	0.795	0.905	0.846	0.783
[7] (2015)	PCNN	0.755	0.805	0.778	0.723	0.786	0.842	0.811	0.759	0.797	0.881	0.836	0.773
[26] (2015)	Fine-tuned CaffeNet with oversampling	-	-	-	-	-	-	-	-	-	-	-	0.830
[27] (2016)	DCAN (Alex)+ReKL	-	-	-	-	-	-	-	-	-	-	-	0.838
[28] (2017)	Fine-tuned CaffeNet with oversampling	-	-	-	0.749	-	-	-	0.787	-	-	-	0.830
[29] (2017)	Hybrid SentiBank + Late fusion	0.739	0.809	0.772	0.711	0.765	0.823	0.792	0.734	0.804	0.864	0.833	0.772
[30] (2017)	VGG19 + saliency map	-	-	-	-	-	-	-	-	0.91	0.89	0.90	0.87
[31] (2018)	SentiNet-A	0.82	0.80	0.81	0.777	0.85	0.83	0.84	0.807	0.89	0.87	0.88	0.851
Proposed approach	**XRA-Net**	**0.842**	**0.804**	**0.821**	**0.792**	**0.862**	**0.842**	**0.851**	**0.812**	**0.924**	**0.886**	**0.903**	**0.864**

TABLE I
DATASET DETAILS

Sentiment	5- Agree	At least 4-Agree	At least 3-Agree
No. of Positive samples	581	689	769
No. of Negative samples	301	427	500
Total	882	1116	1269

The images in the dataset were manually annotated by five Amazon Mechanical Turks (AMT) workers. Hence, 5-agree signifies that all the five AMT works gave a unanimous label for the image. The dataset details are shown in Table I. The dataset contained many challenges. Since the users upload the images in the dataset, hence the images are not clear (blur). Also, many images included text in them. Hence, this becomes difficult for the layers to extract the relevant features from such images. Fig. 2. displays some of the challenging images from the dataset.

The proposed XRA-Net architecture is implemented using Python 3, on Windows 10, a 64-bit machine with two NVIDIA Titan RTX GPUs and 128 GB RAM. The network is built using the popular neural network framework Keras. To avoid overfitting, data augmentation techniques are used, which includes horizontal flipping, zooming, rescaling, and shearing. Adam optimizer is applied with 0.001 learning rate. The network is trained for three epochs with 1000 steps per epoch. The binary cross-entropy loss serves as the objective function of the proposed network. The weights of the new layers are initialized with Xavier uniform initializer. The

Fig. 2. Some of the challenging images from the dataset

remaining layers are trained with the weights of the pre-trained model. The entire model took 3 hours to train.

B. Classification Results

For evaluating the performance of the XRA-Net architecture, we have performed five-fold cross-validation accuracy so that each class is represented equally across each of the test fold with 32 batch size. The final sentiment value is computed by averaging the values. Hence, the classification accuracy (ACC) denotes the five-fold cross-validation results on the target datasets. The other evaluation metrics used are Precision (P), Recall (R), and F1 score (F1).

TABLE II
CLASSIFICATION RESULTS OF XRA-NET ON VSA DATASETS

Dataset	P	R	F1	ACC
3-agree	0.842	0.804	0.821	0.792
4-agree	0.862	0.842	0.851	0.812
5-agree	**0.924**	**0.886**	**0.903**	**0.864**

From Table II, we observe that the proposed XRA-Net architecture performs best for the 5-agree dataset and worst for the 3-agree dataset. This is because, as compared to the 5-agree dataset, the 3-agree dataset contains more noisier images as discussed in the previous section.

C. Baseline Comparison

Table III shows the comparative results of the proposed XRA-Net model with similar state-of-the-art works on all the subsets of Twitter I datasets. Our results are highlighted in Bold. The following baselines are used for comparison.

- Low-Level Feature-based approaches: The low-level feature-based approaches include Local Color Histogram (LCH), Global Color Histogram (GCH), and SIFT-based bag of visual terms [8].

- Mid-Level Feature-based approaches: These includes SentiBank [25] which contains 1200 ANPs and Sentribute [13] which uses scene-based attributes as semantic level features. Moreover, Li *et al.* [29] fused the image sentiment value with corresponding ANP responses to predict the final sentiment of the image.

- Deep Learning-based approaches: The deep learning-based approaches used for comparison are: You *et al.* [7] applied progressive CNN (PCNN) for fine-tuning the network, Campos *et al.* [28] and [26] fine-tuned CaffeNet and used oversampling for removing the dataset bias. Wang *et al.* [27] extracted descriptiveness and objectiveness features using DCAN (Deep Coupled Adjective and Noun) network and Rectified Kullback-Leibler loss (ReKL) by using pre-trained Alexnet network. Song *et al.* [31] used VGGNet architecture for producing the image representation and applied visual attention and saliency detection for attending the prominent regions of the images for the VSA process.

Hence, from Table III, we observe that the XRA-Net architecture has outperformed all the baseline results on the three subsets of Twitter I dataset based on Accuracy, Precision, Recall, and F1 score. Further, from the results, we observe that the deep learning based approaches have shown tremendous improvement over the traditional low-level feature based and mid-level feature-based approaches. The Xception unit of the XRA-Net architecture can fetch the local image features, and the residual attention network extracts the relevant, informative features in the image for obtaining the global information about the whole image.

D. Comparison with the Inception V3 model

We have compared our Xception model with Inception V3 model by replacing the Xception based CNN unit with Inception V3 unit and end-to-end training the model again. The results are shown in Fig. 3.

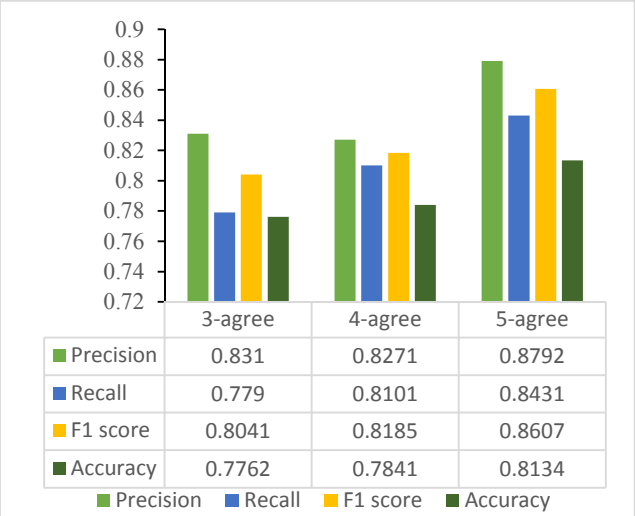

Fig. 3. Comparison of Xception with Inception V3

From Fig. 3, it is evident that the Xception unit outperforms the Inception V3 module on all the four evaluation metrics. Hence, the Xception unit improves the accuracy over Inception V3 unit on all the datasets by 3%-5%.

V. CONCLUSION

This work proposes a deep learning based architecture, XRA-Net (Xception Residual Attention based network) for visual sentiment analysis. The Xception based CNN unit can extract the local image features, whereas the residual attention unit captures the attention-aware features in the image, which helps in getting the global information about the whole image. The efficacy of the XRA-Net architecture is evaluated on publicly available, real-world Twitter I dataset, which contains three sub-datasets: 3-agree, 4-agree, and 5-agree. The accuracy achieved on these datasets are 79.2%, 81.2%, and 86.4% respectively, which shows that the proposed architecture has outperformed all the state-of-the-art results.

In the future, the proposed architecture can be used for movie or TV series analysis by predicting whether the movie/series conveys a positive sentiment or a negative sentiment. Apart from movie analysis, the model can be

extended to capture the feelings or emotions expressed in the videos uploaded by the users on various social networks. It can further be enhanced to obtain the information from multimodal data in the form of text, speech, or video.

VI. References

[1] R. Sawhney, P. Manchanda, P. Mathur, R. Singh and R. R. Shah, "Exploring and Learning Suicidal Ideation Connotations on Social Media," in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Belgium, 2018.

[2] Y. Yin, R. R. Shah and R. Zimmermann, "Learning and fusing multimodal deep features for acoustic scene categorization," in *ACM Multimedia Conference on Multimedia Conference.*, Seoul, 2018.

[3] H. Jangid, S. Singhal, R. R. Shah and R. Zimmermann, "Aspect-Based Financial Sentiment Analysis using Deep Learning," in *Companion Proceedings of the The Web Conference 2018* , France, 2018.

[4] T. Xiao, H. Li, W. Ouyang and X. Wang, "Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, United States, 2016.

[5] E. Ohn-bar and M. M. Trivedi, "Multi-Scale Volumes for Deep Object Detection and Localization," *Pattern Recognition,* vol. 61, pp. 557-572, 2016.

[6] H. Yang, C. Yuan, B. Li, Y. Du and J. Xing, "Asymmetric 3D Convolutional Neural Networks for Action Recognition," *Pattern Recognition,* vol. 85, pp. 1-12, 2019.

[7] Q. You, J. Luo, H. Jin and J. Yang, "Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, USA, 2015.

[8] S. Siersdorfer, E. Minack, F. Deng and J. Hare, "Analyzing and Predicting Sentiment of Images on the Social Web," in *18th ACM international conference on Multimedia*, USA, 2010.

[9] V. Vonikakis and S. Winkler, "Emotion-Based Sequence of Family Photos," in *20th ACM international conference on Multimedia*, Japan, 2012.

[10] B. Li, S. Feng, W. Xiong and W. Hu, "Scaring or Pleasing: Exploit Emotional Impact of An Image," in *20th ACM international conference on Multimedia*, Japan, 2012.

[11] D. Borth, R. Ji, T. Chen, T. Breuel and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *21st ACM international conference on Multimedia*, Spain, 2013.

[12] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara and S.-F. Chang, "Visual Affect Around the World: A Large-scale Multilingual Visual Sentiment Ontology," in *Proceedings of the 23rd ACM international conference on Multimedia.*, Australia , 2015.

[13] J. Yuan, Q. You, S. Mcdonough and J. Luo, "Sentribute : Image Sentiment Analysis from a Mid-level Perspective," in *Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, Chicago, 2013.

[14] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-s. Chua and X. Sun, "Exploring Principles-of-Art Features For Image Emotion Recognition," in *22nd ACM International conference on Multimedia*, Florida, 2014.

[15] Y.-Y. Chen, T. Chen, T. Liu, H.-Y. M. Liao and S.-F. Chang, "Assistive Image Comment Robot - A Novel Mid-Level Concept-Based Representation," *IEEE Transactions on Affective Computing,* vol. 6, no. 3, pp. 298-311, 2015.

[16] F. Chen, R. Ji, J. Su, D. Cao and Y. Gao, "Predicting Microblog Sentiments via Weakly Supervised Multimodal Deep Learning," *IEEE Transactions on Multimedia,* vol. 20, no. 4, pp. 997-1007, 2018.

[17] J. Yang, D. She, M. Sun, M.-m. Cheng, P. L. Rosin and L. Wang, "Visual Sentiment Prediction based on Automatic Discovery of Affective Regions," *IEEE Transactions on Multimedia,* vol. 20, pp. 2513-2525, 2018.

[18] Z. Li, Y. Jiao, X. Yang, T. Zhang and S. Huang, "3D Attention-Based Deep Ranking Model for Video Highlight Detection," *IEEE Transactions on Multimedia,* vol. 20, no. 10, pp. 2693-2705, 2018.

[19] B. Zhao, X. Wu, J. Feng, Q. Peng and S. Yan, "Diversified Visual Attention Networks for Fine-Grained Object Classification," *IEEE Transactions on Multimedia,* vol. 19, no. 6, pp. 1245-1256, 2017.

[20] G. Letarte, F. Paradis, P. Giguere and F. Laviolette, "Importance of Self-Attention for Sentiment Analysis," in *EMNLP Workshop Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP*, Belgium, 2018.

[21] Y. Wang, M. Huang, L. Zhao and X. Zhu, "Attention-based LSTM for Aspect-level Sentiment Classification," in *Conference on empirical methods in Natural Language Processing*, Texas, 2016.

[22] Q. You, L. Cao, H. Jin and J. Luo, "Robust Visual-Textual Sentiment Analysis: When Attention meets Tree-structured Recursive Neural Networks," in *24th ACM international conference on Multimedia*, Amsterdam, 2016.

[23] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *IEEE conference on computer vision and pattern recognition*, United States, 2017.

[24] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang and X. Tang, "Residual Attention Network for Image Classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, United States, 2017.

[25] D. Borth, T. Chen, R.-r. Ji and S.-f. Chang, "SentiBank : Large-Scale Ontology and Classifiers for Detecting Sentiment and Emotions in Visual Content," in *21st ACM international conference on Multimedia*, Spain, 2013.

[26] V. Campos, A. Salvador, B. Jou and X. Giró-i-nieto, "Diving Deep into Sentiment : Understanding Fine-tuned CNNs for Visual Sentiment Prediction," in *1st International Workshop on Affect & Sentiment in Multimedia*, Australia , 2015.

[27] J. Wang, J. Fu, Y. Xu and T. Mei, "Beyond Object Recognition : Visual Sentiment Analysis with Deep Coupled Adjective and Noun Neural Networks," in *Twenty-Fifth International Joint Conference on Artificial Intelligence*, New York, 2016.

[28] V. Campos, B. Jou and X. Giro-i-Nieto, "From Pixels to Sentiment : Fine-tuning CNNs for Visual Sentiment Prediction," *Image and Vision Computing,* vol. 65, pp. 15-22, 2017.

[29] Z. Li, Y. Fan, W. Liu and F. Wang, "Image sentiment prediction based on textual descriptions with adjective noun pairs," *Multimedia Tools and Applications,* vol. 77, no. 1, pp. 1115-1132, 2017.

[30] S. Fan, M. Jiang, Z. Shen, B. L. Koenig, M. S. Kankanhalli and Q. Zhao, "The Role of Visual Attention in Sentiment Prediction," in *25th ACM international conference on Multimedia*, California, 2017.

[31] K. Song, T. Yao, Q. Ling and T. Mei, "Boosting image sentiment analysis with visual attention," *Neurocomputing,* vol. 312, pp. 218-228, 2018.