International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

MCNN: Visual Sentiment Analysis using Various Deep Learning Framework with Deep CNN

[1]Rahul Subhash Gaikwad, [2]Dr. Sharanabasappa C. Gandage

Submitted: 27/10/2022 **Revised:** 24/12/2022 **Accepted**: 24/01/2023

Abstract: Sentiment analysis is a technique for assessing people's opinions and viewpoints based on the text or photos they publish on social networking sites such as Instagram and Twitter. Because it can be difficult to pinpoint the precise thoughts, ideas, and sentiments expressed in a text document or image, sentiment classification is a difficult undertaking. People share their emotions in diverse ways depending on the situation and subject. In this paper, we proposed a hybrid feature extraction and selection technique using numerous deep learning techniques. The features extracted from the image, such as luminance, chrominance, histogram, autoencoder, etc., are validated with a modified convolutional neural network called mCNN. The number of deep CNN layers, size of extracted features, various activation functions, and different optimizers have been used for CNN feeding. In an extensive experimental analysis, our module was tested and compared with two different deep learning modules, such as RESNET and VGGNET. Our proposed module obtains higher accuracy than two conventional deep learning frameworks.

Keywords*: Visual Sentiment analysis, CNN, ResNet50, VGG16, Deep Learning*

1. Introduction

Social networking sites such as Facebook and Instagram have created a public forum where anybody can share their thoughts, feelings, and social commentary with the world. Extraction of information from such information has become a subject of enormous importance as people all over the globe have the ability to share their feelings about many topics linked to politics, academia, tourism, art, commercial applications, or themes of common interest. Understanding users' emotions as they are represented by their comments on different platforms has proven to be a crucial component for the evaluation of people's perspectives about a particular topic, in addition to providing information regarding users' visited websites, shopping habits, etc. Classifying a text's valence according to the user's satisfaction, discontent, or neutrality is a widely popular technique. Although the categorization or number of levels from affirmative to negatives can vary, generally, polarity refers to the mood of a text, which can range from a joyful to a sad state. Series of techniques for collecting useful features and categorizing text into suitable polarity tags are dependent on various approaches to Natural Language Processing (NLP) and Machine Learning (ML). These methods are implemented for sentiment classification. Numerous Deep Neural Networks (DNNs) [32,33] have been successfully used in the area for some time due to the development of deep learning techniques. Convolutional neural networks in general showed to be effective for sentiment analysis applications.

It is beneficial to comprehend the many feelings and sentiments that an image might convey and to immediately forecast affective tags on them, such as joy, affection, sorrow, rage, calm, etc., as pictures play such an important role in modern culture. In this study, we try to forecast the feelings of a picture that belongs to the Pleasure, Shock, Grief, Anxiety, Aggression, and Neutral categories. In order to predict feelings and analyze emotions of several static photos using a highly customized GUI, we implemented our unique convolutional neural network and a pretrained CNN framework. In recent times, sentiment classification has become a more widely used tool on CNN [31,34]. The development of visual information forecasting enabled us to explore convolutional capabilities for visual sentiment analysis and sentiment forecasting in this research. The most crucial need for using convolutional neural network is a large dataset with a variety of images to offer better accuracy. The dataset, a fundamental component needed for continuous expansion, plays the biggest role in the model training and for the forecast.

Research Scholar: Dept of computer science and Engineering Dr. A.P.J. Abdul kalam University, Indore
rahul.gaikwad2k13@gmail.com
Asst Prof : Dept of computer science and Engineering, Dr. A.P.J. Abdul kalam University, Indore
sharangandage@gmail.com

2. Literature Survey

A unique approach called a deep metric network via heterogeneous semantics was proposed in 2021 by Yun Liang et al. [1]. It has recently added captioning characteristics to the image sentiment classification, which was motivated by the fact that image captioning has the capacity to characterize the contents of a picture. Then, it generated the trustworthy latent space of image sentiments using the joint loss as well as the HS characteristics made up of captioning characteristics and visual characteristics. Additionally, in the two distinct types of well-known datasets, the empirical findings demonstrated that the suggested technique surpassed various comparison techniques, as well as the state-of-the-art technique. To execute multi-modal sentimental analysis encompassing both the relevant data and the social relations in 2020, Jie Xu et al. [2] develop an Attention-based Heterogeneous Relational Model (AHRM). It provides a revolutionary continuous dual focus (channel attention and region attention) to emphasize the affective semantic-significant regions and develop a combined image-text representation for taking benefit of the feelings and emotions among image and text. The process then expands the Graph Convolutional Network (GCN) to collect the content data from social settings as an alternative method of learning high-quality representations. It does this by building a diverse connection network from retrieved social structures. On two benchmark datasets, tests were carried, and the results showed that we outperformed the state-of-the-art baselines. Although some pairs are not realistic, the model primarily relies on information with fine-grained connections between words and pictures. Additionally, some photographs might not flow naturally with others, which is also not adequately taken into account in our approach. In order to enhance effectiveness even more, the program plans to develop a more logical data model ahead.

In 2020, Yingying Pan et al. [3]This work designs an experiment that is distinct from the typical empirical research and uses NLP to analyse the data in order to test an essential claim in traditional calligraphy theory—the richness of the subjective picture of calligraphy strokes. It has been discovered that regular technical students are capable of creating complex visual associations for individual calligraphy strokes. Only the "horizontal strokes" of running and regular script are included in this paper's experimental samples. The system intends to increase the experimental stimuli in the future in order to better comprehend the picture space of calligraphy strokes. This paper's importance comes from offering a fresh experimental aesthetic approach to classical aesthetics, which entails designing experiments to test aesthetic feelings, expressing those feelings verbally, and employing quantitative analysis techniques for natural language processing. This approach can be expanded to include the aesthetic study of painting, music, and other art forms in addition to calligraphy study. The categorization network of convolutional neural networks was used by Junfeng Yao et al. [4] in 2016 to study sentiment classification. By using supervised training, it may turn the scene image emotion prediction into a more conventional classification forecast. In the 3 additional categorization networks, 15000 photos have indeed been trained and tested. Numerous studies show that deep convolutional networks are effective at directly mapping picture information to very abstract emotional semantics. This strategy also requires the assistance of a sizable dataset, which may be its most important component. It anticipates the release of datasets like ImageNet to encourage additional study in this area.

In 2015, Stuti Jindal et al. [5] It might be difficult and intriguing to perform visual sentiment classification. It propose to concentrate on the evaluation of photos, one of the most common media types used by internet microblogging services, whereas the great majority of prior works of sentiment classification on social web were performed on words. Convolutional Neural Networks is employed in this research to address this issue. To tackle small training sample sizes, it has introduces a unique training methodology. It has created an emotion benchmark using picture uploads on Flickr, which has a huge collection of images and accompanying tags representing users' feelings, to assess the developed model on actual data. Additionally, it added a 7-scale granularity of emotion grading, which is more thorough than the bi-polar labeling method found in the datasets already available and used by other studies. A small number of photos with certain labels have shown significant increases in both Transfer Learning (TL) and gradual training. According to the empirical results, convolutional neural networks that have been adequately trained using the provided methodology can surpass existing solutions for the extremely difficult task of visual sentiment classification. We can explore a number of intriguing possibilities ahead. The first step is to apply semi-supervised learning to adjust convolutional neural network to the emotion photos using Flickr user-tagged information. Additionally, the system seeks to apply the study results to several uses in other fields, like video games and polling.

Igor Santos et al. [6] evaluated a different CNN model in 2017 for sentence sentiment classification. On standard datasets, the suggested technique produces extremely good outcomes. While training on a multiclass dataset like SST-1, the convolutional neural

network technique surpassed all others and observed significant improvements. Despite the positive outcomes, the T-Test reveals that there is no variation in how often word embedding is utilized. Therefore, using dynamic vectors that lengthen training times while fixed ones produce the best results makes absolutely no sense. A superior outcome for a dynamic embedding was only shown in one of the examples. The convolutional neural networks include a number of additional hyperparameters, and tweaking those takes a lot of effort. Since the goal of this experiment was to examine their impact on convolutional neural network effectiveness, it held them consistent while changing the word embeddings. As a result, the system intends to evaluate the effects of various settings of many other hyperparameters in the future.

Various Deep Learning (DL) approach setups based on convolutional neural network and Long short - term memory (LSTM) networks are investigated for sentiment classification in Twitter data by Sani Kamş et al. in 2019 [7]. This analysis produced slightly less accurate results than the most recent methodologies, but it still produced data that were comparable, allowing drawing conclusive results about the various settings. These techniques' comparatively poor performance demonstrated the field constraints of convolutional neural network and long short term memory networks. In terms of its setup, it has been found that combining Convolutional neural network and LSTM networks yields better results than doing so alone. This is because CNNs effectively reduce the number of dimensions, and LSTM networks preserve word connections. Additionally, the system performs better when many CNN and LSTM nets are used. As suggested, having an adequate dataset is the crucial component for boosting the effectiveness of such algorithms, as shown by the differences in predictive performance between various datasets. Therefore, it appears that investing more time and energy in producing high-quality training sets offers more benefits than testing with various CNN & LSTM network configurations. In conclusion, this paper's innovation is that it enabled assessment of several deep neural network setups and experimentation with two distinct word embedding techniques using a single dataset and assessment methodology, allowing for greater understanding of their benefits and drawbacks.

In 2017, Lifang Wu et al. [8] introduced a pre-processing approach to improve the database in accordance with the emotions of ANPs and labels to tackle the concerns of mislabeled train dataset that is obtained from social networks for DL-based visual sentiment classification. By integrating the softmax and Euclidean loss functions, it significantly enhanced DL approach. The test results validated the efficacy of the suggested algorithms. The techniques for dataset refinement were verified. It might be applied to gather a new, more substantial dataset from social networking sites. Since their real emotion labels can't be determined automatically, certain photographs are currently eliminated. Future development on the system hopes to provide more capabilities to remedy the issue, allowing for the inclusion of more photos. A multimodal oriented sentiment classification system that successfully assesses the emotions from text-image online data is presented by Selvarajah Thuseethan et al. in 2020 [9]. To the best of our knowledge, this is the first study to understand the relevant correlations among high attention words and salient picture areas in sentiment classification, in complement to the integrated graphic and text characteristics. Three extracting features streams—VFS, TFS, and AFS are included in the proposed methodology. The results show that this approach is applicable for predicting the feelings from text-image online data. Additionally, this methodology is tested on a newly created multimodal emotion dataset. The integration of VFS, AFS, and TFS considerably enhances the resilience and overall sentiment classification efficiency, according to the comparative examination of different setups. Potential developments of the suggested multimodal based sentiment classification method include Human-Computer Interaction (HCL), psychology, social media tracking, and web user's assessment. This program captures the connections among text and images characteristics in the same manner that people interpret emotions. Additionally, more paradigms and how they interact can be leveraged to improve sentiment analysis effectiveness. Integrating several other paradigms to create a better sentiment classification methodology is thus one possible future approach in this area.

The picture sentiment classification is a significant research topic and a hotspot in the area of computer vision in 2019, according to Jiajie Tang et al. [10]. The activity scene picture is used as the research object in this work, and an activity scene-based sentiment analysis atlas is created. For photos of active scenes, a sentiment categorization technique based on deep neural network is suggested. The deficit in the field of activities scene design emotion categorization has been filled in part by the conceptual gap among low-level qualities and high-level characteristics of pictures being decreased. This research proposes a deep neural network-based sentiment classification approach for active scene photos that has excellent resilience for dynamic scene analysis. Creative media firms can categorize the gathered activity scene graphics psychologically using this classifier, which not only lightens the burden but also addresses the issue that there is no set standard for manual categorization. Consumers

can assess their interests while selecting activity situations offline or online by measuring the type and quantity of photographs, and then suggest analogous activity situations that satisfy their emotional demands. In essence, it addresses the issue of automatically classifying images in active scene design and retrieving active scenes from the viewpoint of people's subjective sentiments, opening up a new area of study for picture emotional classification methods. In 2020, Monika Saini et al. [11] Sentiment classification is the method for examining people's opinions and viewpoints based on the text or photos they publish on social networking sites like Facebook and Instagram. Because it might be difficult to pinpoint the precise thoughts, emotions, and sentiments expressed in the text, sentiment classification is a difficult undertaking. Individuals express their emotions in diverse ways depending on the situation and subject. By integrating the text and existing information, this problem can be addressed. In ability to accomplish sentiment classification on tweets, this investigation suggests a Deep Convolutional Neural Network (DCNN) that employs word- to paragraph level data. Convolutional neural networks can be trained effectively and useful characteristics can be found with the help of a new method for initializing the weights of the network. A DL model is used to further fine-tune the algorithm and lower the performance of the classifier. It makes use of feature extraction using word vector characteristics and a CNN. Additionally, the approach includes soft-max classification development. Three distinct datasets containing 3K, 10K, and 100K tweets each are used in the trials. In compared to current methodologies, the suggested technique significantly improves accuracy, specificity, and recall.

In 2015, Theodoros Giannakopoulos et al. [12] all contemporary business analytics paradigms place a high priority on brand tracking and reputation management. To capture the core attitude for certain brands, contemporary related technologies, though, mainly depend on the linguistic component of web content. In this research, the system breaks the program's text-only hurdle by demonstrating sentiment classification in the context of a business monitoring paradigm. A broad range of visual elements are retrieved to this goal, some of which concentrate on the fundamental semiology and aesthetics of the pictures. Additionally, it makes use of text information that is included in the photos under investigation by using text mining methods that concentrate on sentiment extraction. The classification method for the specific binary task (negative sentiment versus positive sentiment) is evaluated, and a fusion strategy is suggested that integrates the two various paradigms. The assessment process was then applied to two distinct use cases, including (a) a general visual sentiment classification model for brand and marketing pictures and (b) a brand-specific categorization approach where the identity of the input photos is known a-priori. Findings have shown that the categorization of brand as well as advertising data depends on emotion can function better than the corresponding text-based categorization. Additionally, combining the two modalities significantly improves efficiency.

In 2018, Jufeng Yang et al. [13] with the rising trend of people conveying their ideas online through videos and images, autonomous sentiment analysis of visualizations has received much interest. This study looks into the issue of visual sentiment classification, which requires a sophisticated level of abstraction during the identification procedure. The system is inspired by the fact that both the entire picture and local areas communicate substantial sentiment data, while the majority of existing approaches concentrate on enhancing holistic presentations. It suggests a paradigm to utilize emotive areas, where it first generates possibilities using a commercial object category tools and then uses a candidate selection procedure to eliminate duplicate and distracting suggestions. The affective areas are then autonomously found while taking both the sentiment value and the object rating into account after each applicant is linked to a CNN to calculate the sentiment value. In order to create the final forecasts, the CNN outputs from local areas are combined with the complete images. The system merely calls for image-level tags; greatly reducing the strain of annotations that otherwise would have been needed for training. This is crucial for sentiment analysis since emotional zone identification is too subjective and time-consuming and emotion can be abstract. Numerous tests show that the suggested technique beats cutting-edge methods on 8 well-known standard datasets.

In 2021, Munish Mehta et al. [14] the scope of sentiment classification has expanded in recent years. It initially primarily concentrated on the analysis of textual data, with an emphasis toward extending its ranges. Progressions have also been made in various data types over time, such as audio and visual data. Investigators from all over the world have demonstrated a significant knowledge in this topic and are developing new methods to do this work. The many techniques that are frequently used for sentiment classification have been described thoroughly. Here, some of the strategies implemented by different researchers in their investigations serve as examples. The planned study work has also looked into a variety of present uses for sentiment classification. For instance, keeping an eye on social media platforms like Facebook, monitoring customer support, and forecasting election outcomes.

In 2019, Dongyu She et al. [15] with the growing tendency for sharing opinions internet, automated sentiment analysis of visual images have drawn a lot of interest. Due to the huge amount of complexity used in the identification procedure, the system used in this paper resolves the difficult challenge of visual sentiment classification. Regardless of the fact that various image areas may have varying influences on the evoked emotion, current approaches based on CNNs acquire sentiment evaluations from the holistic picture. A poorly supervised coupled convolutional model is presented in this research work. The accompanying contributions are made by this method, which is devoted to autonomously choosing pertinent soft proposals from poor tags (such global picture labels) in order to greatly lessen the annotated load. By training a fully-convolutional net with the cross spatial pooling approach in the identification branch, weakly supervised coupled convolutional network initially identifies a sentiment-specific softness mapping. Secondly, by combining the emotion map with deep characteristics for robust presentation in the categorization branch, both the global and localized data are made use of. The branches for sentiment detection and characterization are combined into a single deep architecture, and the network is optimized from beginning to end. Poorly supervised sentiment categorization and identification gain from one another through this collaborative learning approach. Numerous tests show that the developed weakly supervised coupled convolutional network surpasses the most recent findings on 7 standard datasets.

In 2019, Jing Zhang et al. [16] since it have been researched for so long; the majority of ML techniques treat the image emotion as distinct, independent tags. Feelings are actually produced by simultaneous signal suppression in the brain as a result of numerous hormone interactions. A unique Multi-Subnet Neural Network (MSNN) is developed that models the human mind process for image sentiment analysis. It is influenced by the neurological microcircuit in the brain. In contrast to conventional neural networks, the MSNN expands a new domain network to simulate how images activate various neural circuits in the brain and generate emotional semantic features using a multi-subnet and signal reformation network. Studies reveal that MSNN works better than other multi-class sentiment classifiers and is ideally suited to the task of classifying the emotion of multiple classes of pictures. According to Xingxu Yao et al. [17] in 2020, a picture speaks a million words. Since more users are expressing their feelings through photos and videos online, several academics have carried out in-depth research to investigate visual emotions. The majority of CNN-based approaches currently in use, though, ignore both the hierarchical and complicated character of feelings in favor of retrieving and categorizing emotional pictures in a discrete label space. On the one side, there is a hierarchical link between emotions, unlike tangible and separated object ideas (like cat and dog). However, the majority of commonly employed deep approaches rely on a representation from completely linked layers, which is deficient in the texture data required to recognize emotions. This research uses adaptable deep metric learning to solve the issues. To be more precise, it creates an adaptable sentiment similarity loss that may incorporate emotive images while taking the sentiment polarity into account and flexibly change the boundary among various image pairs. It further suggests the emotion vector, which captures the texture data acquired from several convolutional layers, to differentiate emotional images efficiently. In order to accurately improve retrieval and categorization goals, it builds a unified multi-task deep architecture. The suggested scheme works favourably compared to the state-of-the-art approaches, as shown by rigorous and in-depth assessments on four standard datasets.

In 2019, Jianfei Yu et al. [18] with the aim of predicting the emotion alignments over specific target entities addressed in users' postings, entity-level sentiment classification of social media posts has recently gained significant attention. The majority of current techniques to this problem focus exclusively on text data, failing to take into account other significant data sources (such as photographs, videos, and user information), which may be able to improve these text-based techniques. This research investigates entity-level multimodal sentimental analysis in light of the observation and investigates the potential utility of photos for entity-level sentiment analysis in social platform post. For this objective, researchers specifically suggest an Entity-Sensitive Attention and Fusion Network (ESAFN). It first uses an efficient attention mechanism to create entity-sensitive textual representations that are then aggregated with a text fusion layer to record the intra-modality dynamics. Following the entity-oriented visual attention process and a regulated technique to remove the messy visual environment, ESAFN develops the entity-sensitive visual display. Additionally, ESAFN additionally combines the textual and visual depictions with a bilinear interaction layer to depict the inter-modality patterns. Depending on two currently released multimodal NER datasets, it physically annotates the sentiment polarity across each given object to assess the efficacy of ESAFN. The results demonstrate that ESAFN can dramatically surpass a number of highly aggressive unimodal and multimodal approaches.

In 2019, Pooyan Balouchian et al. [19], Due to the large amount of internet content that is now accessible online,

still picture sentiment identification has gained more interest in recent times. To mention a few, the application fields include opinion mining, visual sentiment classification, exploration, and retrieval. Less research has been done on sentiment classification, which deals with identifying the precise sentiment generated when presented to particular visual stimulation, even if there are publications on the topic that propose ways to identify picture emotion, i.e., identifying the orientation of the picture. This research fills two major gaps: (1) there aren't enough large-scale image collections for DL of visual emotions; and (2) there aren't enough context-sensitive solitary techniques to sentiment classification in the still picture domain. In this study, Labelled UCF Emotion Recognition (LUCFER) which is a dataset with over 3.6 million images and three-dimensional annotations for sentiment, environment, and valence—is introduced. The LUCFER was assembled utilizing a brand-new data collecting process that was suggested and put into practice in this study. Furthermore, by adding a dimension reduction tier to the convolutional neural network, the innovative multinomial classification technique presented here is used to train a context-sensitive DL classification model. Based on the theoretical approach to emotion detection, the framework contends and experimentally demonstrates that adding background to the unified training procedure aids in (1) achieving a more stable precision and recall and (2) enhancing performance, resulting in an overall categorization accuracy of 73.12 percent as opposed to 58.3 percent accomplished in the nearest work in the literature.

Despite the advancements in face detection and object categorization, computational models are still having difficulty effectively simulating human attention in the area of human gaze forecast in 2019, according to Macario O. Cordel et al. [20]. It is challenging to represent visual attention programmatically because it is a complicated human activity affected by many variables, varying from low-level aspects (such as color, brightness) to high-level sensory consciousness (such as objects connections, object emotion). In this study, researchers look into the connection between item emotion and viewer interest. Human fixation opinion is the focus of an updated performance measure (AttI) that is used to measure human attention. A number of experimental data analyses using AttI show that attention is favourably drawn to emotion-evoking items, particularly when they co-occur by emotionally neutral items, and that the degree to which this preference changes with picture intricacy. It creates a DNN for human attention forecast depending on the empirical investigations, allowing the attention biases on emotion-evoking items to be incorporated in its feature map. Specifically on criteria that assess the importance of salient areas, tests on two standard data indicate its enhanced results. This study represents one of the very first attempts to computationally describe this phenomenon and offers the best image to date on how item attitudes affect human focus. In 2019, Ya-Fang Peng et al. [21] Developers are frequently compelled to make educated guesses about color schemes and other specifications based on the vague and occasionally inconsistent specifications provided by clients. As a result, collaboration and alteration take up a lot of time. In order to rapidly transform common observations into color palettes, a technique of fusing color imagery using sentiment classification is described in this study. There are 4 phases in the method. It starts by defining effect words as the foundation for text categorization. Affect keywords are CIS image keywords in this research. Secondly, it compiles pertinent text corpus from Wikipedia and Google. Finally, it uses a word-embedding model (word2vec) to determine the lexicon affinity of impact words and colors via training the model. Additionally, a prototype method is created to show how effective it is at creating color palettes automatically.

In 2016, Hakan Bilen et al. [22] an essential issue in picture interpretation that still hasn't found a good answer is weakly supervised learning of item recognition. In order to solve this issue, this paper makes use of DCNN that have already undergone extensive image-level categorization training. The modification of one such net to work at the level of picture areas, performing concurrently area selection and categorization, is described in a weakly supervised deep recognition framework. On the PASCAL VOC data, the framework implicitly trains object recognizer that are superior to competing weakly supervised recognition methods after being trained as a picture classification model. The method outperforms conventional data augmentation and fine-tuning methods for the goal of image-level categorization as well. It is a straightforward and elegant end-to-end framework. In 2016, Bolei Zhou et al. [23] explore the global mean pooling layer and explain how, despite having trained on image-level tags, the CNN is explicitly able to localize objects with amazing accuracy. The method was initially suggested as a way to regularize training, but the system discovers that it actually creates a general localizable deep depiction that reveals the implicit focus of convolutional neural networks on a picture. Despite training on any bounding box labeling, the method is still able to obtain 37.1% top-5 inaccuracy for object localization on ILSVRC 2014 despite the appearance of simplification in global mean pooling. The tests in this

study show that the network can localize the discriminative image areas even though it has only been taught to solve categorization tasks.

In 2017, Xavier Alameda-Pineda et al. [24]The machine vision group has recently become interested in the automatic detection of virality, the ability of an images and videos to spread fast and extensively through social networks, which is vital in our too linked society. In addition, recent developments in DL frameworks demonstrated that activation mapping, which emphasize the areas of a picture most known to contain cases of a particular class, may be extracted using global pooling algorithms. The learned top-N mean pooling, a pooling layer that learns the dimensions of the support area to be aggregated, expands on this idea in this study. It makes the hypothesis that such a thorough pooling method may be necessary for the latent ideas (feature mapping) explaining virality. By adding the Learned Top-N Average (LENA) layer on top of the convolutional framework, it examines the LENA layer's overall performance of forecasting and locating virality. It describes studies on 2 readily viewable datasets that have been tagged for virality and demonstrates how the strategy surpasses cutting-edge techniques.

In 2017, Ronak Kosti et al. [25] in daily lives, it is crucial to comprehend what someone is going through from their point of view. One could assume that technologies with this kind of capability will connect with people more effectively for this purpose. Furthermore, there are currently no systems available that can comprehend an individual's psychological response in great depth. The majority of prior research on using machine vision to discern sentiments has concentrated on examining the facial expression and categorizing it into the six fundamental feelings. Therefore, the environment is crucial to how emotions are perceived, and when it is taken into account, it can imply more human emotions. The "Emotions in Context Database" (EMOTIC), a dataset of photographs depicting humans in context in unrestricted contexts, is discussed in this work. People are identified in these photographs using 26 different emotional categories as well as the continual parameters of polarity, arousal, & dominance. A CNN model is trained using the EMOTIC database, which analyzes the individual and the entire scenario together to identify rich information about feelings and emotions. This serves as a benchmark for the problem of emotion detection in visual context and demonstrates the significance of taking context into account when identifying feelings and emotions in photos.

In 2017, Ali Diba et al. [26] Identifying objects is a difficult activity in the visual understanding domain, and it becomes considerably more difficult if the supervision is insufficient. Few attempts have recently been made to do the work without costly and difficult annotations, due to DNN. To learn a CNN under such circumstances, a new design of cascaded circuits is developed. It offers two such designs that have either two or three end-to-end pipeline-trained cascade phases. By training a fully-convolutional layer, the very first phase of both frameworks selects the best candidate among class-specific zone suggestions. By using results of the first stage's activation maps, the second step of the 3-stage design gives object fragmentation. Both designs' last step is a CNN component that applies instance based learning to the ideas retrieved in the earlier stage (s).The trials on the PASCAL VOC 2007-10-12 and large size item databases, as well as the ILSVRC2013-14 sets of data, demonstrate gains in the weakly-supervised object recognition, categorization, and localization domains. WILDCAT, a DL technique that simultaneously focuses at aligning picture areas for acquiring spatial invariance and acquiring highly localized characteristics, is introduced by Thibaut Durand et al. in 2017 [27]. The system is focused to 3 main visual detection tasks: picture analysis, poorly supervised point - wise object localization, and semantic fragmentation. It is trained using just global image annotations. Modern CNNs are expanded by WILDCAT on three main levels: the intentional design of local attributes associated to various class modes in the net, the use of fully-convolutional layers or retaining spatial resolution, and a novel method of pooling these characteristics to provide the global picture forecast necessary for poorly supervised training. Rigorous testing demonstrates that the approach greatly exceeds the most recent techniques.

In order to overcome the issues with weakly-supervised semantic fragmentation, Yunchao Wei et al. [28] explore a general method to gradually mine discriminative item sections using categorization nets. In contrast to the need of the fragmentation task, which requires localizing dense, inner, and integral areas for pixel-wise inference, categorization systems are only receptive to tiny and scattered discriminative areas from the interest point. This research suggests a novel adversarial erasing method for gradually localizing and enlarging object areas in order to close this gap. The suggested method directs the categorization network to progressively identify new and complimentary object sections by eliminating the existing mined areas in an aggressive way, beginning with a single isolated object area. For the purpose of learning semantic fragmentation, these localized zones ultimately combine to form a compact and comprehensive object area. An online prohibitive fragmentation learning method is created to work in

conjunction with adversarial deleting by offering auxiliary fragmentation supervision that is modified by the more trustworthy classification rating, further enhance the effectiveness of the discovered areas by adversarial removing. Despite appearing straightforward, the suggested method yields mean Intersection-over-Union (mIoU) ratings on the PASCALVOC2012 val and testing dataset that are the most recent state-of-the-art at 55.0 percent and 55.7 percent, respectively.

A system for categorizing emotions that is taught by empirical data from psychological investigations was proposed by V. Yanulevskaya et al. in 2008 [29]. The International Affective Picture System (IAPS), a common set of emotion-evoking pictures used in psychology, was used to evaluate the affective valences in the dataset. The strategy is based on the evaluation of regional picture statistics that are discovered for each emotional class by SVM. It displays outcomes for the approach on the IAPS dataset as well as a selection of works of art. Even though the findings are tentative, they show that robots have the ability to evoke true emotions when examining masterpieces. The prediction of an Emotion Stimuli Map (ESM), which represents the pixel-by-pixel contribution to evoked sentiments, is a new machine vision issue introduced in 2016 by Kuan-Chuan Peng et al. [30]. Researchers discover that the areas chosen by saliency and item recognition do not accurately forecast the picture areas that trigger emotion after using a new image dataset, emotion region of interest, as a baseline for forecasting the ESM. Parts of the surroundings are just as crucial for generating emotion as the objects themselves. This research suggested fully convolutional layers for ESM prediction based on this feature. The approach can identify the areas that generate sentiment better than either of saliency or object recognition, according to both qualitative as well as quantitative testing data.

3. Propsoed System Design

In this part, proposed methodology for sentiment analysis and emotion predictors is discussed in detail. The dataset is initially pre-processed, as seen in Figure 1, to help remove noise and fragmented data from the vast dataset that is accessible. The convolutional neural network method was subsequently employed for training purposes, allowing us to anticipate the mood and sentiment shown by the photos. The trained model is then used to estimate the sentiment represented by the picture under the areas of happy, shock, sorrow, fear, hate, and neutral using a specific image that has been loaded or a live image that has been captured. This section contains comprehensive information.

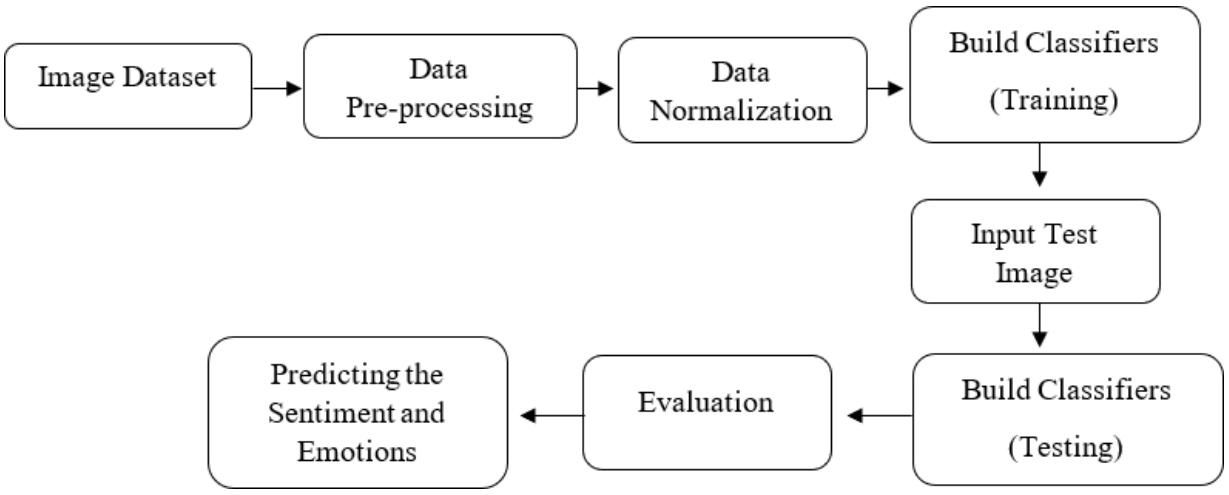

Fig. 1: Proposed system architecture for image sentiment analysis

3.1 Data Gathering

The most essential and crucial component of any emotion forecast is data. To produce a better precision, a lot of data is required. The input for our system is gathered from a variety of social media sites, including Facebook, Twitter, Linkedin, and others, where pictures are frequently shared and data is communicated rather than text. Some of the photographs in the collection were also taken from ImageNet, which offers access to a large set of photos. With this assortment of photographs, we have created a unique dataset that is accessible via the online platform. Figure 2. depicts sample images of ImageNet Dataset.

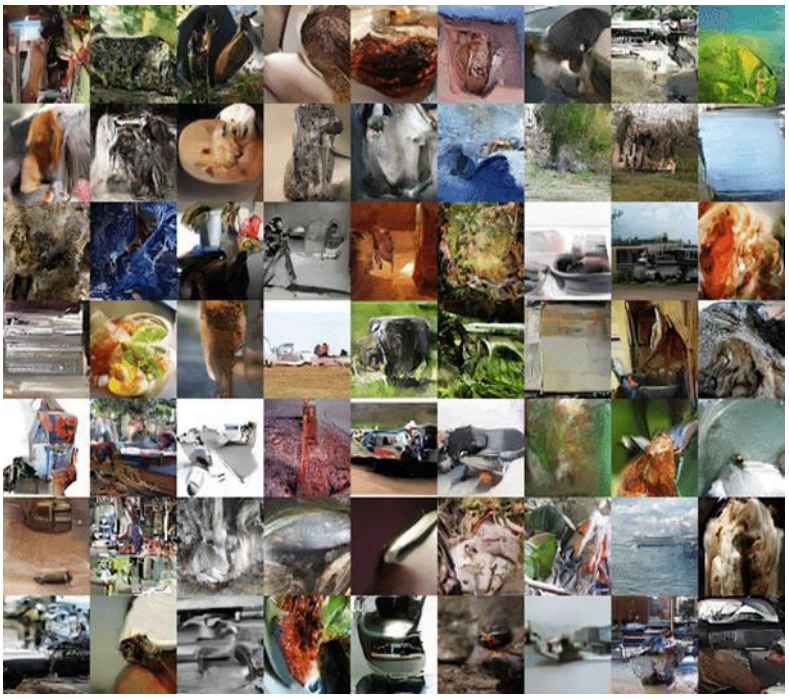

Fig. 2: Sample image of ImageNet Dataset

3.2 Data Pre-processing

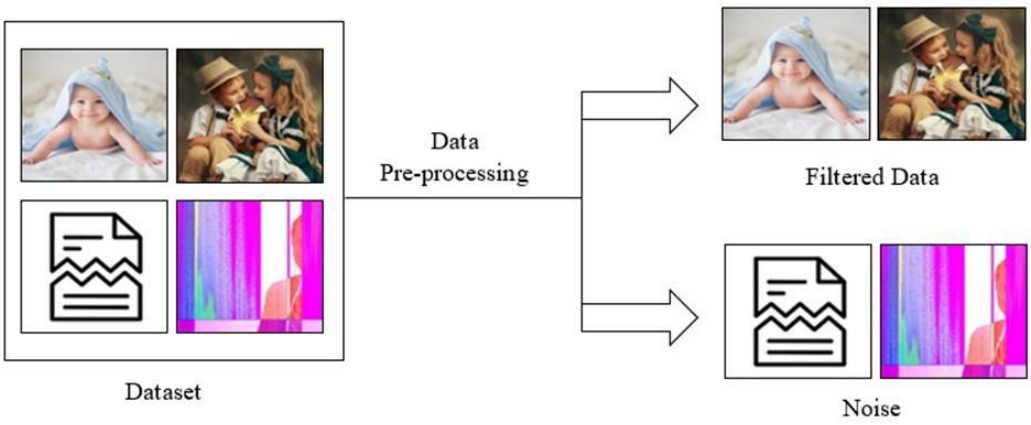

Fig. 3: Data Pre-processing

Before moving on to any further tasks, cleansing the dataset is perhaps the most crucial step. Therefore, before continuing, all of the distortion from the enormous number of datasets that were produced had to be removed. We removed any duplicated or corrupted photographs as well as any pictures that did not meet the necessary criteria. Figure 3 demonstrates how the dataset is pre-processed to eliminate the noise and prepare it for future processing.

For greater precision and quicker outcomes, the photos with various sizes are further translated into a customized size. To enable the model to operate effectively and without interruption, all of the photos in the dataset have been transformed to a certain format.

3.3 Implementation of CNN

We employed a customized convolutional framework trained on our generated dataset together with pre-trained ResNet & VGG16 models for the training procedure. The VGG16 classifier, which is pre-trained using the ImageNet dataset, is displayed in Figure 4. On the other side, we developed our suggested seven-layer customized convolutional neural network model to be trained on our own dataset. Although it was trained on a customized dataset, this model functions similarly to the VGG16 and ResNet models. The reliability of these 3 models is compared, and the one that performs best will be utilized going forward to forecast the emotional state of the visual imagery via the GUI. The more data a deep learning model utilizes for training, the better the results will be which another crucial factor is. We used a unique

dataset for training the customized convolutional neural network framework. In order to improve predictive performance, our proprietary model adds extra layers to the convolutional neural network framework. As a consequence, the dataset produces better outcomes, and the forecast is considerably more precise.

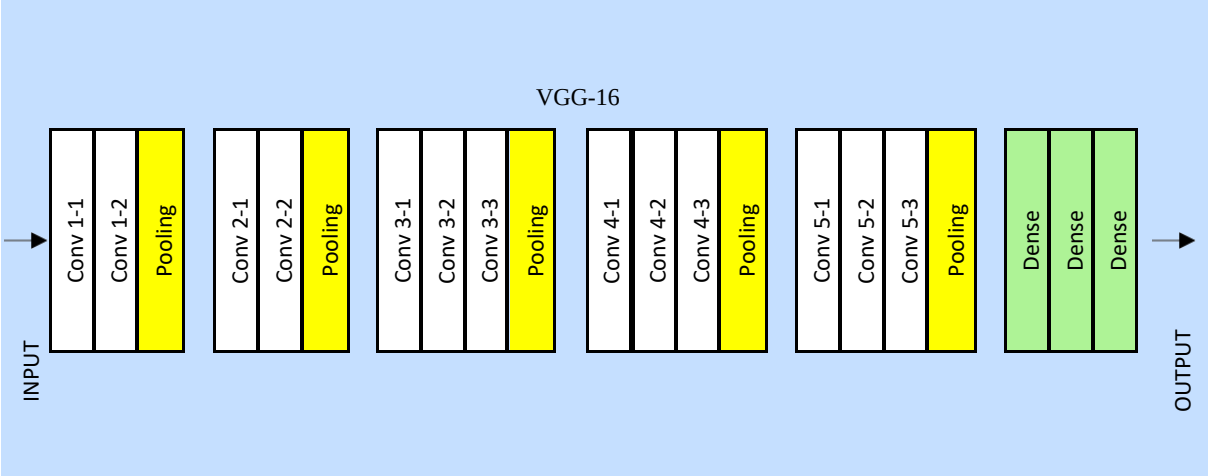

Figure 4: VGG16 Architecture

Algorithm Design

Input: Normalized training dataset *Train_Data[],* Normalized testing dataset *Test_Data[],* defined threshold *qTh*

Output: Result set as output with *{Predicted_class, weight}*

Step 1: Read all test data from *Test_Data[]* using below function for validating to training rules, the data is normalized and transformed according to algorithms requirements

$$\text{test\_Feature(data)} = \sum_{m=1}^{n} (.\ \text{Attribute\_Set[A[m]} \ldots\ldots\ldots \text{A[n]} \leftarrow Test\_Data)$$

Step 2 : select the features from extracted attributes set of $\text{test\_Feature(data)}$ and generate feature map using below function.

$$\text{Test\_FeatureMap}\ [t\ldots\ldots n] = \sum_{x=1}^{n}(t) \leftarrow \text{test\_Feature(x)}$$

$\text{Test\_FeatureMap}\ [x]$ are the selected features in pooling layer. The convolutional layer extracts the features from input and passes to pooling layer and those selected features are stored in *Test_FeatureMap*

Step 3: Now read entire taring dataset to build the hidden layer for classification of entire test data in sense layer,

$$\text{train\_Feature(data)} = \sum_{m=1}^{n} (.\ \text{Attribute\_Set[A[m]} \ldots\ldots\ldots \text{A[n]} \leftarrow Train\_Data)$$

Step 4 : Generate the training map using below function from input dataset

$$\text{Train\_FeatureMap}\ [t\ldots\ldots n] = \sum_{x=1}^{n}(t) \leftarrow \text{train\_Feature(x)}$$

$\text{Train\_FeatureMap[t]}$ is the hidden layer map that generates feature vector for build the hidden layer. That evaluate the entire test instances with train data.

Step 5 : After generating the feature map we calculate similarity weight for all instances in dense layer between selected features in pooling layer

$$\text{Gen\_weight} = \text{CalcWeight}\ (\text{Test\_FeatureMap}\ ||\ \sum_{i=1}^{n} \text{Train\_FeatureMap[i]})$$

Step 6 : Evaluate the current weight with desired threshold

$$\text{if(Gen\_weight} >= \text{qTh)}$$

Step 7 : $\text{Out\_List. add}\ (\text{trainF. class, weight})$

Step 8 : Go to step 1 and continue when $\text{Test\_Data} == \text{null}$

Step 9 : Return Out_List

4. Experimental Analysis And Results

We demonstrate the tests done on our dataset in this section. The training dataset and testing dataset are divided into 80% and 20% in accordance to apply the 80-20 approach. This is performed in order to assess the model's correctness on data that it has never seen before. As a result, we were able to determine whether we were over-fitting the training dataset or whether we needed to train for longer epochs and at a slower learning rate if the accuracy rate was greater than the training accuracy.

4.1 Proposed CNN Model with 7 layers

After the information has been loaded and divided, it is pre-processed by normalizing the values such that they are all within the [0, 1] range and converting the values to the required size as well as what the network requires. We transformed the data into a standard shape because it had previously been different sizes and shapes. It has been found that pre-processing the data improves accuracy. An outcome accuracy of 0.81 was produced by the Convolutional neural network with 7 convolutional layers.

4.2 VGG16 Model

This is a pre-trained classifier that was developed using the ImageNet dataset. For this system, we placed the last layer initially, and for greater fine tuning, we added the softmax layer after that. The model demonstrated test accuracy of 0.35, which is somewhat lower than the suggested convolutional neural network framework because it is pre-trained on a broad variety of classes.

4.3 Model ResNet-50

The ResNet model is a pre-tained classifier as well, making it a useful tool to use while working with photos. In order to compare the levels of accuracy, we also tried out this pre-trained classifier. This model has 50 layers and is particularly deep because it was trained both on scene- and object-centric data (MS COCO). On our unique dataset, this model delivered a testing accuracy of 0.48.

4.4 Results

This section describes the empirical findings of the proposed methodology for sentiment classification and emotion recognition.

Table 1: Accuracy of three models: Proposed, VGG16, ResNet50

Model	Accuracies achieved	
	Validation set	Testing set
Proposed Model	**0.68**	**0.81**
VGG16	0.41	0.36
ResNet50	0.44	0.49

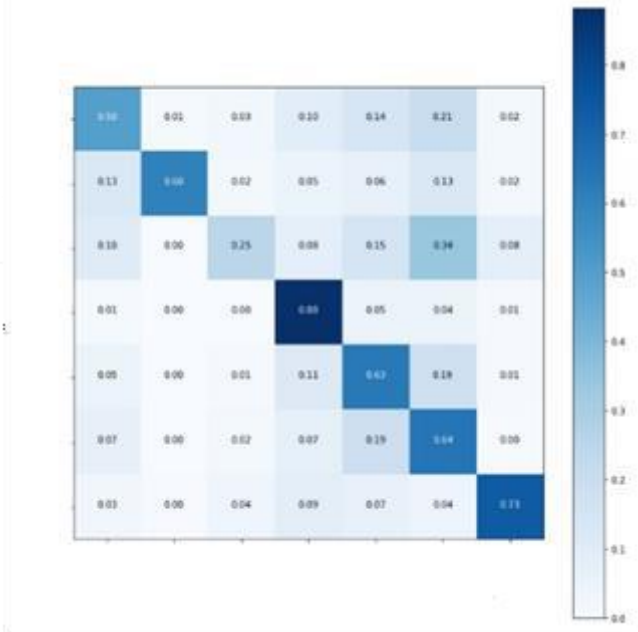

Fig. 5: Confusion Matrix

The accuracy of the suggested model, VGG16 and ResNet, is displayed in Table 1. On validation dataset accuracy achieved by proposed model is 0.68, VGG16 classifier is 0.41and ResNet50 classification model is 0.44. On testing dataset, accuracy achieved by proposed model is 0.81, VGG16 classification model is 0.36 and ResNet50 classifier is 0.49. According to the results of this study, the suggested framework, which was trained on a custom dataset, offers better precision than another two pre-trained models. The confusion matrix found after evaluating the pictures from the test set with various emotions like shock, joy, angry, disgust, fear, neural;, sorrow is shown in Figure 5.

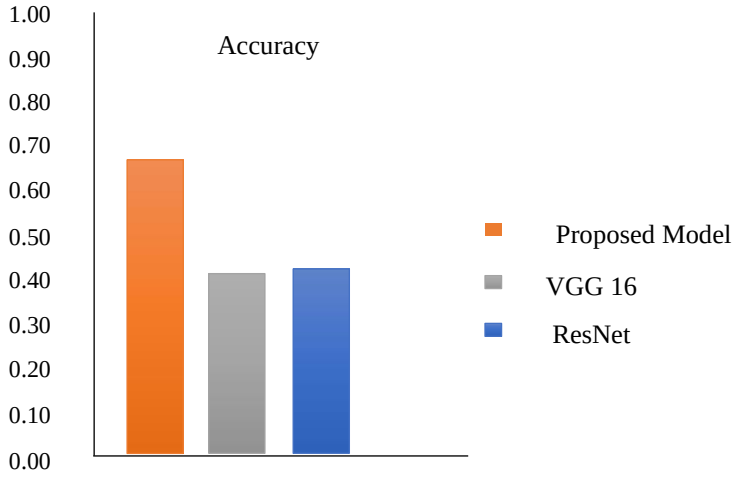

Figure 6: Accuracy of three models: Proposed, VGG16, ResNet50 on Validation Dataset

Figure 6, illustrates the accuracy of proposed model, VGG16 classification model and ResNet50 classifier using Validation dataset.

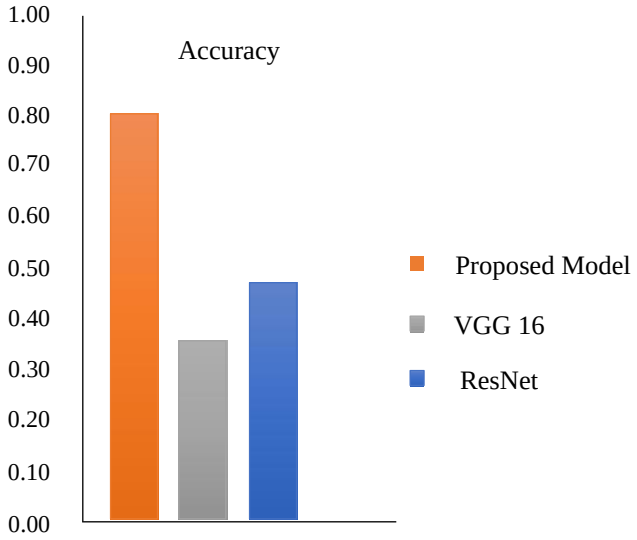

Figure 7: Accuracy of three models: Proposed, VGG16, ResNet50 on Testing Dataset

Figure 7, illustrates the accuracy of proposed model, VGG16 classification model and ResNet50 classifier using test dataset.

5. Conclusion

In this paper we proposed a mCNN for identification of image sentiment with deep learning framework. Forecasting visual feelings has been one of the most active fields of interest in recent decades. Furthermore, the majority of earlier studies used convolutional neural network characteristics for image categorization rather than sentiment prediction. In this study, the accuracy of a proposed model and two pre-trained classifiers is compared in order to determine which model offers the greatest chance of success in the future. Additionally, the suggested model performs better than existing models

when trained on a unique dataset and delivers results that are more accurate and promising. Future trends worth looking at range in many different directions. In this work, our initial focus has been on identifying the sentiment of pictures in order to anticipate the emotion they are intended to convey, such as joy, shock, sorrow, rage, fear, or no feeling at all. On social networking sites where the use of photographs is expanding quickly, this technique may be helpful. Users might save time by using this instead of putting in or looking up sentiment tags. Secondly, it can be advanced to employ dynamic images or videos for safety reasons, using real time video surveillance to identify an individual's actions and determine whether or not they are questionable.

References

[1] Yun Liang, Keisuke Maeda, Takahiro Ogawa and Miki Haseyama. "Deep Metric Network via Heterogeneous Semantics for image Sentiment Analysis", 2021, International Conference on Image Processing (ICIP), IEEE.

[2] Jie Xu, Zhoujun Li, Feiran Huang, Chaozhuo Li, and Philip S.Yu. "Social Image Sentiment Analysis by Exploiting Multimodal Content and Heterogeneous Relations", 2020, IEEE.

[3] Yingying Pan, Ruimin Lyu, Qinyan Nie and Lei Meng. "Study on the Emotional Image of Calligraphy Strokes based on Sentiment Analysis", 2020, IEEE.

[4] Junfeng Yao, Yao Yu, and Xiaoling Xue. "Sentiment Prediction in Scene Images via Convolutional Neural Networks", 2016, 31st Youth Academic Annual Conference of Chinese Association of Automation, IEEE.

[5] Stuti Jindal and Sanjay Singh. "Image Sentiment Analysis using Deep Convolutional Neural Networks with Domain Specific Fine Tuning", 2015, International Conference on Information Processing (ICIP), IEEE.

[6] Igor Santos, Nadia Nedjah and Luiza de Macedo Mourelle. "Sentiment Analysis using Convolutional Neural Network with fastText Embeddings", 2017, IEEE.

[7] Sani Kamış and Dionysis Goularas. "Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data", 2019, International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), IEEE.

[8] Lifang wu, Shuang Liu, Meng Jian, Jiebo Luo, Xiuzhen Zhang and Mingchao Qi. "Reducing Noisy Labels in Weakly Labeled Data for Visual Sentiment Analysis", 2017, IEEE.

[9] Selvarajah Thuseethan, Sivasubramaniam Janarthan, Sutharshan Rajasegarar, Priya Kumari and John Yearwood. "Multimodal Deep Learning Framework for Sentiment Analysis from Text-Image Web Data", 2020, WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE.

[10] Jiajie Tang, Liandong Fu, Chong Tan and Mingjun Peng. "Research on Sentiment Classification of Active Scene Images Based on DNN", 2019, International Conference on Virtual Reality and Intelligent Systems (ICVRIS), IEEE.

[11] Monika Saini and Mala Kalra. "An Enhanced Convolution Neural Network Based Approach for Classification of Sentiments", 2020, International Conference for Emerging Technology (INCET), IEEE.

[12] Theodoros Giannakopoulos, Michalis Papakostas, Stavros Perantonis and Vangelis Karkaletsis. "Visual sentiment analysis for brand monitoring enhancement", 2015, 9th International Symposium on Image and Signal Processing and Analysis (ISPA), IEEE.

[13] Jufeng Yang, Dongyu She, Ming Sun, Ming-Ming Cheng, Paul L. Rosin and Liang Wang. "Visual Sentiment Prediction based on Automatic Discovery of Affective Regions", 2018, IEEE.

[14] Dr. Munish Mehta, Kanhav Gupta, Shubhangi Tiwari and Anamika. "A Review on Sentiment Analysis of Text, Image and Audio Data", 2021, 5th International Conference on Computing Methodologies and Communication (ICCMC), IEEE.

[15] Dongyu She, Jufeng Yang, Ming-Ming Cheng, Yu-Kun Lai, Paul L. Rosin and Liang Wang. "WSCNet: Weakly Supervised Coupled Networks for Visual Sentiment Classification and Detection", 2019, IEEE.

[16] Jing Zhang, Han Sun, Zhe Wang and Tong Ruan. "Another Dimension: Towards Multi-Subnet Neural Network for Image Sentiment Analysis", 2019, international Conference on Multimedia and Expo (ICME), IEEE.

[17] Xingxu Yao, Dongyu She, Haiwei Zhang, Jufeng Yang, Ming-Ming Cheng and Liang Wang. "Adaptive Deep Metric Learning for Affective Image Retrieval and Classification", 2020, IEEE.

[18] Jianfei Yu, Jing Jiang and Rui Xia. "Entity-Sensitive Attention and Fusion Network for Entity-Level Multimodal Sentiment Classification", 2019, IEEE.

[19] Pooyan Balouchian, Marjaneh Safaei and Hassan Foroosh. "LUCFER: A Large-Scale Context-Sensitive Image Dataset for Deep Learning of Visual Emotions", 2019, Winter Conference on Applications of Computer Vision, IEEE.

[20] Macario O. Cordel, Shaojing Fan, Zhiqi Shen and Mohan S. Kankanhalli. "Emotion-Aware Human Attention Prediction", 2019, CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE.

[21] Ya-Fang Peng and Tzren-Ru Chou. "Automatic Color Palette Design Using Color Image and Sentiment Analysis", 2019, 4th International Conference on Cloud Computing and Big Data Analytics, IEEE.

[22] Hakan Bilen and Andrea Vedaldi. "Weakly Supervised Deep Detection Networks", 2016, Conference on Computer Vision and Pattern Recognition, IEEE.

[23] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva and Antonio Torralba. "Learning Deep Features for Discriminative Localization", 2016, Conference on Computer Vision and Pattern Recognition, IEEE.

[24] Xavier Alameda-Pineda, Andrea Pilzer, Dan Xu, Nicu Sebe and Elisa Ricci. "Viraliency: Pooling Local Virality", 2017, Conference on Computer Vision and Pattern Recognition, IEEE.

[25] Ronak Kosti, Jose M. Alvarez, Adria Recasens and Agata Lapedriza. "Emotion Recognition in Context", 2017, Conference on Computer Vision and Pattern Recognition, IEEE.

[26] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash and Luc Van Gool. "Weakly Supervised Cascaded Convolutional Networks", 2017, Conference on Computer Vision and Pattern Recognition, IEEE.

[27] Thibaut Durand, Taylor Mordan, Nicolas Thome and Matthieu Cord. "WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Pointwise Localization and Segmentation", 2017, Conference on Computer Vision and Pattern Recognition, IEEE.

[28] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao and Shuicheng Yan. "Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach", 2017, Conference on Computer Vision and Pattern Recognition, IEEE.

[29] V. Yanulevskaya, J.C. van Gemert, K. Roth, A.K. Herbold, N. Sebe, and J.M. Geusebroek. "Emotional Valence Categorization using Holistic image Features", 2008, IEEE.

[30] Kuan-Chuan Peng, Amir Sadovnik, Andrew Gallagher and Tsuhan Chen. "Where do Emotions come from? Predicting the Emotion Stimuli Map", 2016, IEEE.

[31] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. "Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning", 2015, Transactions on Pattern Analysis and Machine Intelligence, IEEE.

[32] Kaiqi Huang, Weiqiang Ren, Dacheng Tao and Tieniu Tan. "On Combining Multiple Instance Linear SVM and Bag Splitting for High Performance Visual Object Localization", 2015, Transactions on Pattern Analysis and Machine Intelligence, IEEE.

[33] Bogdan Alexe, Thomas Deselaers and Vittorio Ferrari. "Measuring the Objectness of Image Windows", 2012, Transactions on Pattern Analysis and Machine Intelligence, IEEE.

[34] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji and Guiguang Ding. "Continuous Probability Distribution Prediction of Image Emotions via Multi-Task Shared Sparse Regression", 2016, Transactions on Multimedia, IEEE.