

## **IMAGE SENTIMENT CLASSIFICATION USING DEEP CONVOLUTIONAL NEURAL NETWORK MODELS**

**Rahul Subhash Gaikwad**

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

**Dr Sharanabasappa C.Gandage**

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

**Abstract:** Thoughts, feelings, desires, or views can be expressed using words, photographs, or moving pictures. A developing topic of social analytics study is sentiment analysis of web content. Using various social networking sites like Instagram, LinkedIn, Twitter, Messenger, and others, users communicate their feelings online by sharing texts and sharing photographs. There have been little studies on sentiment analysis of visual information; nevertheless, there's been a huge amount of research on sentiment analysis of textual information. The effectiveness of Convolution Neural Networks (CNNs) for object detection and verification utilizing various networks is examined in this research. There are numerous networks available to test how well the various networks operate. Convolution Neural Network results are evaluated using the well-known standard dataset ImageNet as well some real time image dataset. This research focuses on the investigation of three well-known networks, VGG16, VGG19, ResNet50 on ImageNet dataset. These three models are implemented using Keras & Tensorflow for picture categorization. These models are also used to classify and assess the correctness of an arbitrary collection of tagged photos from the web. In comparison to VGG16 and VGG19, it has been found that ResNet50 can detect images more accurately.

**Keywords:** Image Sentiment Analysis, VGG16, VGG19, ResNet50, Convolutional Neural Network

### **1. Introduction**

Text, images, and videos can all be used to convey emotions. There are a ton of published research on text sentiment classification, but not that many on image sentiment classification. In order to get the best outcomes, a lot of study has been conducted in this field over the past few decades due to the increasing popularity of social networking sites to communicate one's feelings. For image sentiment classification, numerous methodologies have been put forth. These methods can be broadly divided into two categories: Machine Learning (ML) based techniques and lexicon-based techniques. Support Vector Machine (SVM), Bayesian Network, Neural Network (NN), Naive Bayes(NB) and Maximum Entropy are examples of ML-based techniques, whereas lexicon-based methods use statistical and semantic methods. Deep Learning (DL) is a branch of ML or a method for making computers intelligent enough

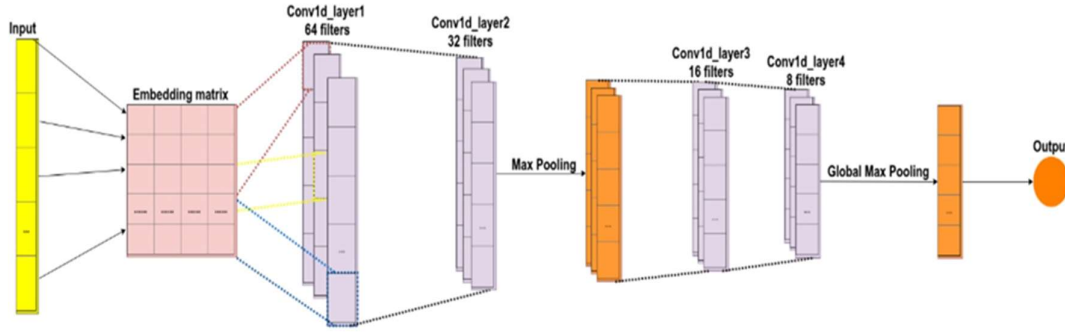
just to learn through experience and comprehend the world of ideas. Without user intervention, computers draw information from their knowledge in the actual world to grasp complex situations and make judgments. Deep learning approaches have been investigated for image sentiment classification and have also produced noteworthy outcomes. The DL method can achieve accuracy or rarely may surpass the cognition of a human and can provide an effective task performance. For picture sentiment classification, DL is crucial.

Currently, CNNs have been proven to beat hand-crafted characteristics on sentiment forecast tasks, delivering end-to-end feature learning methods. The dataset named ImageNet, which contains billions of high-resolution tagged pictures and is publicly accessible develop NN models, is being used in the study work to assess different Deep Convolutional Neural Network (DCNN) models like VGG16, VGG19, ResNet50. The outcomes are viewed in terms of predictive performance. For image identification, the pre-trained classifiers VGG16, VGG19 & ResNet50 have been implemented using Keras & Tensorflow and deployed by using Python.

The paper also addresses the several DCNN methods and their design in part I. The findings of the analysis on image sentiment classification that has been done thus far by using numerous methods are covered in part II. The proposed system architecture and algorithm are explained in detail in part III. Experimental results, observations, and comparative analyses are covered in part IV. The research work is concluded in part V.

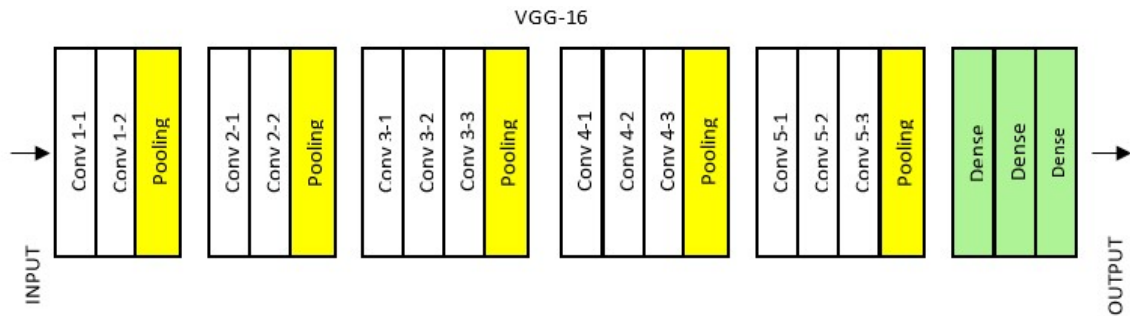
### **1.1 Models of Deep Convolutional Neural Network (DCNN)**

A unique variety of feed-forward NN called CNN was first used in fields including machine vision, recommendation systems, and Natural Language Processing (NLP). Convolutional layers and pooling layers are frequently used in this Deep Neural Network (DNN) architecture to feed data to a fully-connected classification layer. It collects characteristics from their inputs through filtering; the results of various filters can be merged. Layers that are pooled or subsampled have lower feature resolutions, which can make CNNs more resistant to noise and interference. Classification steps are carried out by fully connected layers. Figure 1 shows an illustration of a convolutional neural network model. Pre-processing was done on the raw data to modify it for the embedding matrix. The graphic displays an input embedding matrix that was pushed through 2 max-pooling and 4 convolution layers. A max-pooling layer is used to simplify the result and avoid the problem of over-fitting the information during the initial 2 convolution layers, which include 64 filters and 32 filters respectively that are used to train variety of features. A max-pooling layer is added after the third convolutional layer and fourth convolution layers, which feature 16 filter and 8 filters, respectively. Given that there are 2 classes to be forecasted, the last layer is a fully - connected layer that will decrease the vector of height 8 to an output vector of height 1.



**Figure 1: Convolutional Neural Network Model**

### 1.1.1 VGG16



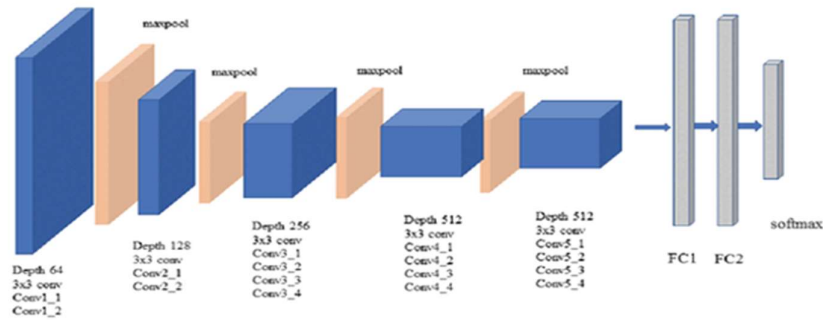
**Figure 2: VGG16 Architecture**

Due to its extremely homogeneous design, VGGNet-16, which has 16 convolutional layers, is quite compelling. It has several filtrations but only 3\*3 convolutions, like AlexNet. On 4 GPUs, it can be trained for 15 to 21 days. The industry now considers it as the best option for collecting characteristics from photos. The weight setup of the VGGNet is openly accessible and it has been utilized as a standard feature extractor in numerous different applications and issues.

### 1.1.2 VGG19

The VGG-19 design has 19 weight layers, and as one move deeper into the structure, the number of filters accessible within every layer increases. It has a stacking of convolutional, max-pooling layer that collects picture characteristics and 3 fully-connected and a Soft-max layer in the categorization section. By blocking the initial 3 convolutional modules (module 1, 2, and 3) and training the following layers (module 4 and 5), which could

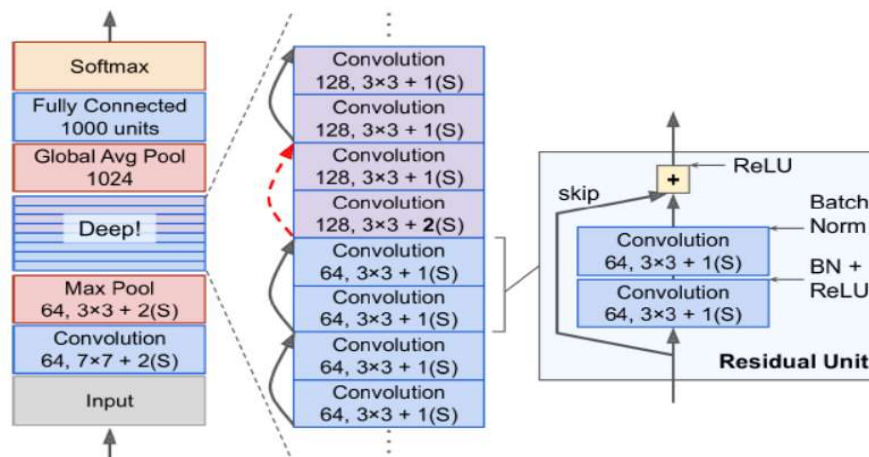
retrieve features relevant to the dataset, we were able to train the dataset using the VGG-19 framework.



**Figure 3: VGG19 Architecture**

### 1.1.3 ResNet50

A CNN with 50 layers is called ResNet-50. A typical NN that serves as the foundation for many machine vision tasks is called Residual Network (ResNet). The main innovation with ResNet was that it enables to train enormously DNN with more than 150 layers. 3\*3 convolutional layer's residual blocks make up the ResNet structure. It has employed a stride of 2 and also has regularly expanded the amount of filters. Resnet50's architecture is depicted in the figure.



**Figure 4: ResNet50 Architecture**

## 2. Literature Survey

Chetanpal Singh et al. [1] describe the creation of a Deep Learning method for classifying the mood of face images in 2021. For object dependent image fragmentation, Recurrent Neural Networks (RNNs) as well as bidirectional long-short-term memory (Bi-LSTM) are utilized, and the CNN-RNN method is being used for non - linear mapping. A number of Deep Neural Networks (DNN) is trained to identify facial movements in 10,000 photos in order to evaluate the usefulness of this suggested methodology. The test findings indicate that the suggested method can categorize the emotions on face images with a correctness of 99.12%. Findings further indicate that the technique improves the categorization model while reducing expense. Image Sentiment Analysis (ISA), which shows how people respond to visualizations like multimedia content, has been an interesting and sort of assumed issue since 2020, according to Papiya Das et al. [2]. In order to categorize, retrieve, and calculate the subjective data in an informative manner, ISA shows how to use the domains of Machine Vision and Natural Language Processing. The development of image processing and machine vision methods is to be recognized for the success of present methods. Most current methods made an effort to address the issue by emphasizing specific visual aspects of the entire images and videos. The most expected inputs are entire picture characteristics. A DL-based approach that includes attention mechanisms for focusing local regions of images and incorporating Support Vector Machines in instead of soft-max categorization layers on Deep Convolution Neural Networks is developed in order to enhance the accuracy of the entire ISA system. Furthermore, it takes into account the pertinent hash tags of a picture to give the Convolutional neural network layer attention weights by semantically linking picture areas and hash tags. The DL-based suggested approach outperforms current state-of-the-art techniques to VSA in its ability to automatically determine the emotional of given photos.

According to Amirhossein Shirzad et al. [3] new research, social media users will progressively use visual media, such as GIFs, videos, and photographs, to share their feelings and emotions in 2020. For various types of twitter posts, a multimodal sentiment classification tool is developed in Python to calculate the sentiment value not only based on the text of the twitter post, but also to take into account GIFs and pictures to enhance the precision of the tweet's entire sentiment value. Researchers employ enhanced convolutional neural network for image sentiment, VADER for textual sentiment, and GIFs both for image sentiment & facial emotion analysis in each frame of the file. In this work [3], it is shown that using both visual and textual features allows for improved outcomes than previous methods that simply use visual or textual characteristics. The output values from every text, picture, and GIF components will be combined to get the final sentiment value for the incoming twitter message. Due to its widespread use and ease of use, textual sentiment classification has grown significantly in importance in social networking sites as of 2017, according to Siqian Chen et al. [4]. In recent years, image sentiment classification has also received a lot of interest. It is clear that these methods—neither text sentiment analysis nor picture sentiment analysis—are insufficient to produce an accurate result on their own. On the other side, their mixture has made the issue

worse. In order to create an advanced approach, Supervised Collective Matrix Factorization (SCMF), this paper demonstrates how to take advantage of the advantages of these techniques. Bag of Glove Vector (BoGV) as well as Alexnet DL network reflects the visual and linguistic features. The suggested method, which draws its inspiration from the graph Laplacian work, factors matrices while taking label data into consideration. To compare the performance of the suggested strategy with other cutting-edge techniques, tests have been carried on two datasets, autonomously labeled and individually labeled sets of data.

Owing to the potential use in personalized recommendation, and other areas, sentiment classification has gained a lot of interest in 2017 according to Xingyue Chen et al. [5]. Since it is extremely difficult to extract enough data from just one type of data—textual or visual—the majority of existing solutions do not produce satisfying results. This paper suggest an end-to-end deep fusion CNN to jointly learn text and image sentiment representations from training instances, which is motivated by the finding that there is substantial semantic association between text and images data in social networks. To forecast the sentiment scores, the data from the two modalities are combined in a pooling layer and supplied into fully connected layers. On two commonly used data sets, the proposed methodology is assessed. Findings shows that the model performs favorably when contrasted to state-of-the-art methods, evidencing its competence.

In 2016, Jie Chen et al. [6] since so many people are able to discuss their emotions and thoughts through visual images on social networking sites; Visual Sentiment Analysis (VSA) has drawn a lot of attention. DNNs are being developed quickly for this purpose as a result of large datasets. Large-scale dataset labeling, therefore, is very time- and money-consuming. In this study, a unique active learning approach is put forth that creates an efficient sentiment classification method using a small number of labeled training data. The conventional convolutional neural network is first extended with a new branch called the "texture component." Analyzing the kernel function of feature maps from several convolutional blocks in this phase will yield the emotional vector. This vector is used by the algorithm to differentiate between emotive images. Secondly, the categorization scores from both the standard Convolutional neural network and the texture module are used to determine the query method. The system is then trained using samples retrieved by applying the query approach. The methodology utilizes a small number of tagged training images to produce positive outcomes for VSA, according to numerous tests on four publicly available emotional datasets.

In 2018, Namita Mittal et al. [7] Feelings, emotions, preferences, or opinions can be expressed using words, photographs, or moving pictures. A developing topic of social analytics study is sentiment classification of online information. Through various social networking platforms like Instagram, LinkedIn, Twitter, Telegram, and others, users communicate their feelings online by trading texts and sharing photographs. There has been little study on sentiment classification of visual information; nevertheless, there has been a great deal of research on

sentiment classification of text data. Adjective Noun Pairs (ANPs), or image feeling automatically identified tags in web photos that are helpful for identifying the feelings or attitudes the picture is trying to express. Predicting or figuring out the emotions in unlabeled photographs is the main challenge. Since DL methods are capable of efficiently learning the picture activity or polarization, sentiment classification employed to address this challenge. Image categorization, image sentiment classification, image sentiment analysis, image identification, and Neural Network effectiveness are some of the areas where DL has demonstrated substantial efficiency. The purpose of this analysis is on some of the notable DL methods, including CNN, Deep Neural Network, Region-based CNN (R-CNN), and Fast R-CNN, as well as the appropriateness of their uses in image sentiment classification and their drawbacks. The report also covers the issues and possibilities facing this developing area.

In 2020, Udit Doshi et al. [8] Social networking sites like Facebook, Twitter, Flickr, and others are increasingly popular, and photographs play a big part in this. People now post specific photographs to these websites to exhibit their sentiments and feelings in the form of pictures on practically each event because it is claimed that "An image is equal to thousands of words." In present era, wherein pictures have mostly taken over everybody's lives, it serves the most significant function. Only a small number of studies have concentrated on evaluating the emotion of visual data, while the majority of current research has concentrated on sentiment analysis of text information. The potential of CNN is investigated in this study in order to forecast the many feelings (for example joy, shock, sorrow, terror, hate, and neutral) expressed by an image. Programs for autonomous tag forecasts of the visual data accessible on social networking sites and for comprehending human attitudes and feelings can both benefit from these kinds of forecasts. On the basis of both visual features and surrounding social network data, Yilin Wang et al. [9] tackled the issue of recognizing human moods from large-scale collections of online photographs in 2015. Despite significant progress in text-based user sentiment classification, sentiment analysis of picture content has mostly been neglected. In doing so, it raises the bar for text-based sentiment forecast problems by extending them to the more difficult goal of guessing the underlying feelings of pictures. It demonstrates that neither text nor visual characteristics by themselves are enough for precise sentiment categorization. As a result, it offers a method for combining the two and proposes the sentiment forecast issue in two different contexts: supervised learning and unsupervised learning. Under the suggested methodology, an optimization technique is created for locating a local-optima solution. The suggested strategy outperforms current state-of-the-art techniques considerably, according to trials on two sizable data. The program plans to incorporate more social media network data in the future and investigate user sentiment on verified social networks.

The proliferation of Internet information in 2021, according to Rui Man et al. [10], makes timely evaluation and assessment of online public sentiment increasingly necessary. Sentiment classification of public opinion occurrences is much more crucial. Words include data that cannot be fully expressed by the word2vec method. In order to fulfil the goal of sentiment classification, it has been suggested to use the BERT method as the article extraction of features

model, the DCNN to retrieve the article's local data, and the fully connected network to categorize the article. CNNs and BERT-based sentiment classification techniques outperform more conventional sentiment analysis methods, according to empirical results using publicly available data sets.

A deep measurement network using heterogeneous semantics, a unique technique for image sentiment classification, is presented by Yun Liang et al. in 2021 [11]. The proposed approach makes significant contributions by incorporating picture captioning into picture sentiment classification to portray a general feeling that cannot be captured by traditional visual data retrieved from images. The suggested strategy creates a new network to incorporate this varied semantic information in order to take into account a sentiment link between perceptual and textual characteristics. The proposed methodology also creates a sentiment latent space by incorporating the center loss for correlations between various sentiments and allows the categorization of picture sentiments while taking into account relations between emotions based on Heterogeneous semantics characteristics. The efficiency boost by deep metric network via heterogeneous semantics is confirmed by the research observations.

According to Jie Xu et al. [12], sentiment classification will gain popularity in 2020 due to its ability to comprehend people's emotions and perceptions in the context of social big data. Conventional sentiment classification techniques concentrate on a single aspect and become useless as massive amounts of data with numerous manifestations appear on social media platforms. To record the relationships among a words and pictures, multi-modal learning algorithms are suggested, however they only go as far as the region level and overlook how strongly connected the pathways are with the semantic data. Additionally, social photographs on social networking sites are related by many different kinds of connections, which are similarly conducive to sentiment analysis but are ignored by the majority of previous works. In order to enhance the effectiveness of multi-modal sentiment classification by including rich social data, an Attention-based Heterogeneous Relational Model (AHRM) is developed in this study. To learn the joint picture-text representation from the standpoint of content data, a gradual dual attention component is specifically suggested to acquire the connections among words and pictures. Furthermore, a channel focus paradigm is suggested to draw attention to semantically rich picture streams, and a region attention model is also created to draw focus to the affective areas involved with the observed networks. In order to learn high quality presentations of social media pictures, it then builds a diverse relation network and expands Graph Convolutional Network to collect the material data from online settings as a complement. Experimental results indicate the efficiency of the suggested model, which has been thoroughly assessed on two standard datasets.

Calligraphy is typically thought of as the discipline of writing lines since calligraphic strokes may convey rich emotive data and spark rich thoughts, according to Yingying Pan et al. in 2020 [13]. One can't possibly know how the general public experience about calligraphic strokes



because conventional calligraphy concept and contemporary aesthetics have explored this issue extensively yet are largely based on individual experience. This finding suggests a research-and-teaching-valued educational study that explores the affective picture of calligraphy strokes while incorporating artistic, scientific, and engineering principles. Each respondent in the study is required to thoroughly feel and envisage using the calligraphic strokes they are given, as well as to freely select a topic and make an article. In order to demonstrate the rich sensory picture of the usual strikes of twelve renowned historical calligraphers, it uses sentiment classification to all of the publications, performs quantitative method, and uses data visualization. In 2016, Junfeng Yao et al. [14] due of the difficulties of determining sentiment explicitly from low-level visual data, sentiment forecasting from visualizations is a challenging task. Adjective Noun Pairs are a common middle level used in latest researches to bridge the gap among perception and emotion. It has become able to implement quite intricate mappings as Convolutional Neural Networks get more sophisticated. In this study, a sentiment-tagged image data is constructed for training. As portion of the study, 15,000 scene photos were trained on 3 distinct Convolutional neural network methods demonstrating how well deep learning can handle a particular sentiment forecast task. In terms of data collecting and technical implementation, Convolutional neural networks methods are likewise more short and simple than those utilizing artificial neural programming.

In 2015, Stuti Jindal et al. [15] on social media sites, pictures are the most accessible form of emotional expression. Users of social media platforms are increasingly sharing their thoughts and concepts through photographs and videos. The forecasting of sentiment from visualizations is complimentary to text sentiment classification since it can help better capture user emotions regarding occurrences or subjects, like those in picture twitter posts. Although this technology has progressed significantly, little study has been undertaken on the image emotions. Convolutional neural networks are used in this work to construct a picture sentiment forecasting model. To execute transfer learning, this architecture is particularly pretrained on extensive object identification data. On a dataset of Flickr images that had been manually categorized, extensive tests were carried. It utilizes a progressive technique of domain-specific deep CNN fine tuning to take advantage of such tagged data. The findings demonstrate that the suggested convolutional neural network training can outperform rival networks in picture sentiment classification.

Convolution Neural Networks (CNNs) are renowned for their amazing results in Machine Vision studies, producing state-of-the-art findings, according to Igor Santos et al. in 2017 [16]. Convolutional neural networks can, nevertheless, function well enough for natural language processing, as per recent research. The entire concept is around compiling the embeddings into an image. This article illustrates how to perform sentiment classification using the just-released Facebook- fast-Text word embeddings as word representations. The introduction of social media and technical advancements have caused the Web to be overrun with viewpoints, are what sparked interest in this research work. The outcomes demonstrate that the suggested strategy surpasses the baseline systems and operates similarly to cutting-edge algorithms.

A study of various deep learning techniques for sentiment classification in Twitter data is presented in 2019 by Sani Kamş et al. [17]. DL methods have grown in favor among academics in this field since they simultaneously assist to the resolution of many different challenges. In addition, convolutional neural networks, which excel in the domain of image analysis, and RNN, which are successfully used in NLP applications, are two classes of NN that are used. The long short-term memory (LSTM) networks and an Recurrent neural network class are both evaluated and compared in this research work. The global vectors for word representation (GloVe) framework and Word2Vec are two further word embedding techniques that are contrasted. One of the most well-known worldwide conferences in the field, the global conference on semantic evaluation (SemEval), supplied information for the assessment of such methodologies. The optimum rating value for every system is evaluated in terms of effectiveness using a multiple tests and variations. By examining the achievements, benefits, and limits of the aforementioned approaches with an assessment process carried out within a single testing method using the same dataset and computational context, this study aims to contribute to the area of sentiment classification. In 2017, Lifang Wu et al. [18] A sizeable dataset must be used for training DL-based graphic sentiment classification. Since some of the photographs gathered in this way were mislabeled, the social network database is both popular and noisy. As a result, the dataset has to be refined. It provides a refinement approach that relies on the emotions of adjective-noun pairs and labels in perspective of analysis to certain datasets. The sentiment discrepancy among the adjective noun pair and tags is used to first identify the photos with unreliable labeling. If there are an equal amount of tags with good and negative attitudes, these photographs are deleted. The leftover pictures are re-labeled using the emotions that received the most votes in the tags. Additionally, by integrating the softmax and Euclidean loss functions, it enhances the conventional DL method. The revised dataset is also used to train the better model. The dataset refinement technique and enhanced DL method both show promise in experiments. The standard outcomes are outperformed by the suggested methods.

In 2020, Selvarajah Thuseethan et al. [19] there are several uses for interpreting people's attitudes from data published on the internet, including understanding the context, predicting election outcomes, and forming opinions about an occurrence. This provides a big research topic. Up until now, text or images have been the main focus of sentiment classification of online data. Furthermore, the widely accessible multiple modal data, such as images and various text formats, when combined, can help to more precisely predict the feelings. Additionally, integrating text and images characteristics indiscriminately makes the model more complicated, which eventually lowers the efficacy of sentiment classification because it frequently misses the proper interconnections across multiple methods. Therefore, in this research, a sentiment classification model is proposed, utilizing the interactions among multimodal online data and salient visual signals and high attention textual cues. To discover the connections among text and learnt prominent visual aspects, a multi-modal deep association classifier is constructed. Additionally, two streams of unimodal deep characteristic extractors are designed to obtain the visual and textual aspects that are most important to the emotions in

order to automatically acquire the discriminative characteristic from the picture and text. The characteristics are then integrated using a late fusion technique to estimate the emotions. In contrast to current unimodal techniques and multimodal techniques that arbitrarily integrate the text and image elements, the thorough assessments show that the suggested approach produced impressive outcomes for sentiment classification utilizing web data.

With the advancement of computer and communication technologies in 2019, Jiajie Tang et al. [20] activity scenario design photos have been processed. This research suggests a multi-featured action scene picture sentiment analysis approach to address the issue of automatically classifying such scene photos. This study created a database of activity scene images and examined the connection between color attributes (artistic style) and affective meanings. It retrieved the global and local color characteristic depending on the scene picture's properties, and then used a DNN classification model to finish the sentimental analysis of the activity scene picture.

In 2016, Haimin Zhang et al. [21] Due to the prevalence of user-generated videos on the web, research efforts on emotion detection in those videos are growing. The majority of current attempts, though, rely on frame-level audiovisual elements, which may not adequately represent temporal features, such as traits gathered over time. It suggested analyzing frequency domain parameters altered by Discrete Fourier Transform (DFT) to extract audiovisual temporal data. A pre-trained DCNN extracts frame-level characteristic initially. After that, time domain characteristics are transformed into discrete fourier transform characteristics and transferred. For sentiment analysis, further encoding and fusion of convolutional neural network and discrete fourier transform features is used. In this manner, temporal data represented by DFT characteristics and static image characteristics collected from a pre-trained deep convolutional neural network are simultaneously taken into account for video sentiment detection. The findings of the experiments show that combining discrete fourier transform characteristics can efficiently collect temporal data and subsequently enhance sentiment detection capabilities. On the biggest video emotional dataset (VideoEmotion8), the method performed at a state-of-the-art level, improving performance from 51.1 percent to 55.6 percent.

In 2009, Jia Deng et al. [22] An increase in robust and advanced methods and techniques for indexing, retrieving, organizing, and interacting with photos and multimedia data may result from the expansion of picture content on the Web. The precise method for organizing and utilizing such data, however, continues to be a major challenge. Here, a brand-new database dubbed "ImageNet," a huge ontology of pictures based on the WordNet structure's framework, is introduced. The most of WordNet's 80,000 synsets will be filled with an aggregate of 500–1000 crisp, full-resolution images owing to ImageNet. As a consequence, WordNet's semantic hierarchy would be used to arrange massive amounts of annotated photos. This article provides a thorough evaluation of ImageNet as it stands today: 12 subtrees totaling 3.2 million pictures

and 5247 synsets. It demonstrates how much more accurate and diverse ImageNet is compared to the existing image datasets. Building a database of this size is a difficult endeavor. It explains the plan for using Amazon Mechanical Turk to gather information. Finally, three straightforward uses in object identification, image analysis, and autonomous object grouping are used to demonstrate the value of ImageNet. The project aimed to provide researchers in the field of machine vision and beyond with unequalled advantages due to the size, precision, diversity, and hierarchy of ImageNet.

In 2014, Maxime Oquab et al. [23] in the major recent visual detection challenge, CNN demonstrated remarkable image categorization accuracy. The advantage of convolutional neural networks over other image classification techniques is their capacity to learn rich middle level visual features as compared to hand-designed limited characteristics. Furthermore, learning convolutional neural networks involves estimating millions of variables and necessitates a huge quantity of tagged image examples. Because of this feature, convolutional neural networks cannot currently be applied to situations with little training data. This study demonstrates how little amounts of training dataset can effectively transfer picture representations learnt using convolutional neural networks on massive labelled data to certain other image detection tasks. It develops a technique for computing medium level image representation for pictures in the PASCAL VOC dataset by reusing layers trained on the ImageNet dataset. It demonstrates that, regardless of changes in picture statistics and tasks between the 2 different datasets, the transferred representations produce noticeably good outcomes for object as well as action categorization on the Pascal VOC 2007 & 2012 datasets than the state of the art. For the localization of objects and actions, it has also produced encouraging results.

According to Ramazan Gokberk Cinbis et al. in 2014, [24] a difficult issue in machine vision is item category localisation. Bounding box labels of object instances are necessary for conventional supervised learning. Weakly supervised learning avoids this time-consuming annotating procedure. The supervised data in this situation is limited to binary tags that show the presence or absence of data objects in the picture without indicating their positions. The method uses multiple-instance learning to continuously train the detector and predict the positions of the objects in the successful training photos. This article's main contribution is a multi-fold instance - based learning method that stops training from picking up on incorrect object placements too early. When using high-dimensional formats like the Fisher vectors, this method is especially crucial. By using the PASCAL VOC 2007 dataset, it provides a thorough experimental assessment. The method effectively localizes items in the training images than state-of-the-art weakly supervised detectors, which leads to greater detection accuracy. Christian Szegedy et al. [25] present a DCNN design dubbed Inception that advances the state of the art for categorization and identification in The ImageNet Large-Scale Visual Detection Competition in the year 2014. This design's primary distinguishing feature is the better exploitation of the network's computational capabilities. It improved the depth as well as width of the network while maintaining the same computing expense through a well-

considered design. The Hebbian principle and the multi-scale processing intuition served as the foundation for the structural choices made to maximize quality. GoogLeNet which is a 22 layers deep network is one specific implementation that was used in the submission for ImageNet Large-Scale Visual Recognition Challenge. Its quality is analyzed in the light of categorization and identification.

In 2015, Maxime Oquab et al. [26] training datasets with a large number of elaborately labeled photos are often necessary for effective visual object identification algorithms. Furthermore, accurate and detailed tagging is both costly and frequently arbitrary, such as by object bounding boxes. It discussed a weakly supervised CNN for classifying things that simply uses image-level labels but can pick up new information from congested situations with many of different objects. On the smaller Pascal VOC 2012 (20 object classes) and much bigger Microsoft COCO (80 object classes) databases, it measures the object categorization and object position forecast performance. It was discovered that the network works similarly to its fully-supervised equivalents by using object bounding box tagging for training, produces efficient image labels and predicts estimated positions of objects (but not their extents). In 2015, Tianjun Xiao et al. [27] since only minute and local differences can distinguish between groups, fine-grained categorization is difficult. Typically, variations in the stance, scale, or orientation make the issue harder. Discovering region of interest or object sections (where) to obtain discriminant information is the process that the majority of fine-grained classification algorithms use (what). In this paper, it is suggested to use deep neural networks to provide visual focus to a fine-grained categorization job. Three different attention categories are integrated by the pipeline: bottom to up attention that suggests potential patches, top to down attention that chooses appropriate patches for a particular object, and part-level top to down attention that automatically detects discriminative portions. These focus areas are combined to train domain-specific deep networks, which are then applied to enhance both where and what factors. Furthermore, it stays away from employing costly tags like bounding boxes or end-to-end part information. The research is simpler to generalize because of the lax supervision restriction. On the subdivisions of the ILSVRC2012 database and the CUB200 2011 database, it has confirmed the efficiency of the strategy. Under the lowest supervision circumstance, the pipeline produced remarkable progress and attained the maximum performance. In comparison to other techniques that depend on more annotations, the effectiveness is competitive.

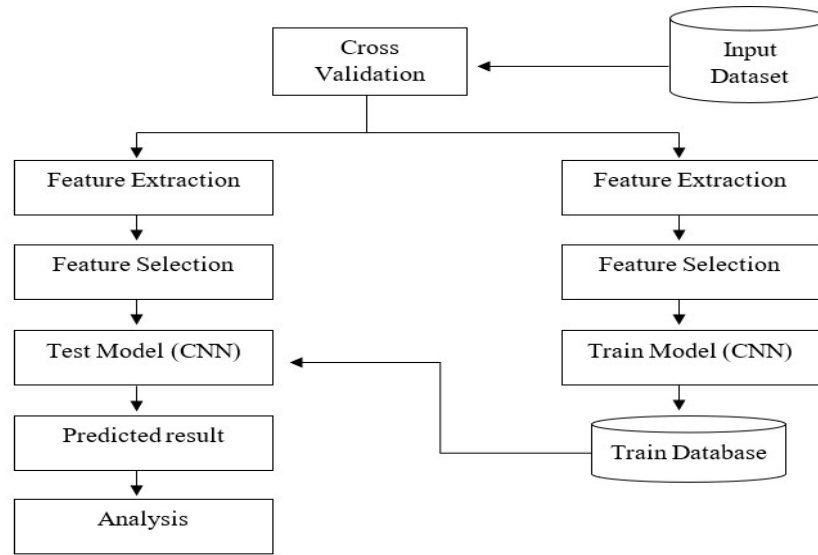
In 2015, Kuan-Chuan Peng et al. [28] explored the two new facets of photographs and feelings. First, it disproves earlier research that only recorded and predicted a single dominating sentiment for every picture by demonstrating through psychovisual investigations that distinct individuals had various sentimental responses to the same picture. The tests also demonstrate that a single image can elicit a variety of emotional responses from the same person. For many uses, it is crucial to forecast feelings as "distributions" rather of a single dominant sentiment. Secondly, it demonstrates that in addition to frequently altering color, texture, and tone-related elements that affect a picture's evoked sentiment, it can also pick which "emotional direction" this change happens by choosing a target picture. It also introduces a brand-new database called

Emotion6 that contains dispersion of sentiments. In 2015, Jiajun Wu et al. [29] recent advancements in deep representation learning have shown how widely applicable they are to typical vision tasks like categorization and recognition. There's been very little research on how to develop a deep learning approach in a weakly supervised environment. This study is an effort to model DL inside an instance - based learning (weakly-supervised learning) paradigm. Every picture in the context maintains a dual multi-instance hypothesis, where potential text tags and object suggestions can be thought of as two instance groups. As a result, it creates efficient technology that uses DL algorithms from both ends to leverage the MIL characteristic. It also attempts to jointly know the connection between object & annotation suggestion. It carries out in-depth tests to demonstrate that the weakly supervised DL paradigm is capable of extracting reasonable region-keyword pairs on broadly used standards like PASCAL VOC as well as MIT Indoor Scene 67 while also achieving brilliant performance in vision applications like categorization and image labels.

DNNs are more challenging to train, according to Kaiming He et al. in 2016 [30]. To make training networks that are significantly deeper than those traditionally utilized easier, it presents a residual learning approach. Rather than learning unreferenced functions, it deliberately reformulates the levels to train residual operations with respect to the layer inputs. It offers in-depth scientific proof that these residual frameworks are simpler to improve and can enhance the accuracy over far more depth. Researchers test residual networks up to 152 layers depth on the Imagenet database, which is 8 layers deeper than VGG but still less sophisticated. On the ImageNet test dataset, an ensemble of such residual network obtains a failure rate of 3.57 percent. This outcome took first place in the 2015 ILSVRC classification task. Additionally, evaluation of CIFAR-10 with 100 layers & 1000 layers is presented. For many image identification tasks, the complexity of representations is crucial. It achieves a 28 % significant enhancement on the COCO object identification dataset only as a result of the extraordinarily deep representations. The contributions to the ILSVRC & COCO 2015 contests were built on deep residual networks, which also took first place on the challenges of ImageNet identification or localization, COCO recognition and fragmentation.

### **3. Proposed system**

DL is used in the proposed research project on image sentiment analysis. This study essentially demonstrates different extraction of features as well as selection processes from image objects and creates the train information with those techniques. Numerous feature selection techniques have been used to retrieve the various characteristics. Text metadata has rarely also been employed to determine the appropriate image sentiment. To improve classification performance, normalizing the data set should have the greatest influence.



**Figure 5: Proposed system design**

A training model is developed in accordance with the characteristics that have been taken from the training data set during the training stage. On the test data set, a similar extraction of features approach has been used to capture every picture attribute appropriately. Finding similarities between training and testing characteristics is done throughout the weight calculation procedure. It is a step in the subprocess of comparing two feature sets. With the desire threshold levels, the weight factor is assessed, and emotion tags are defined as a result. Initial weight is 0, and threshold is user-definable. Since some photographs contain distortion or a certain input picture already includes a particular type of noise, we first examine every picture's height and breadth and adjust as necessary. We remove noise from photos using the noise filtration. The proposed DL method can recognize these characteristics by utilizing the imageNet library. The complete test dataset's emotion class was identified using the Deep Convolutional Neural Network classification method. The system may use the flicker picture dataset for sentiment analysis using a supervised learning strategy, and we divided the data into cross-validation sets of five, ten, and fifteen folds, accordingly. The system immediately creates a train unit for each scaled image because the information has been pre-processed in the trained component. Assessment of the testing dataset has been carried out using several test cases, and the confusion metrics have been computed.

#### 4. Experiments and results

The pre-trained classifiers VGG16, VGG19, and ResNet50 are realized using Keras, a DL-API built in Python and TensorFlow. Massive amounts of photos from the ImageNet dataset, which has been labeled by humans, were used to pre-train these classifiers. For picture classification, several photographs from various classifications have been taken from the web. For bringing together the evaluation of prediction using various networks, we have selected 10 distinct types of photos with human emotions. For a picture, many models have produced varying prediction values.

For instance, when Figure 6 was put through picture prediction, Resnet50, Vgg16, and Vgg19 predicted it to be anger with prediction accuracy of 0.785, 0.680, and 0.714 respectively. In a similar manner, Figure 7 represents boredom with 0.986, 0.758, 0.721 prediction accuracy by ResNet, VGG16 and VGG19 nets respectively. The Resnet50, VGG16, and VGG19 nets were applied to Figure 8, and they correctly identified the object as confusion with prediction accuracy values of 0.999, 0.995, and 0.955, correspondingly.



**Figure 6: Anger**



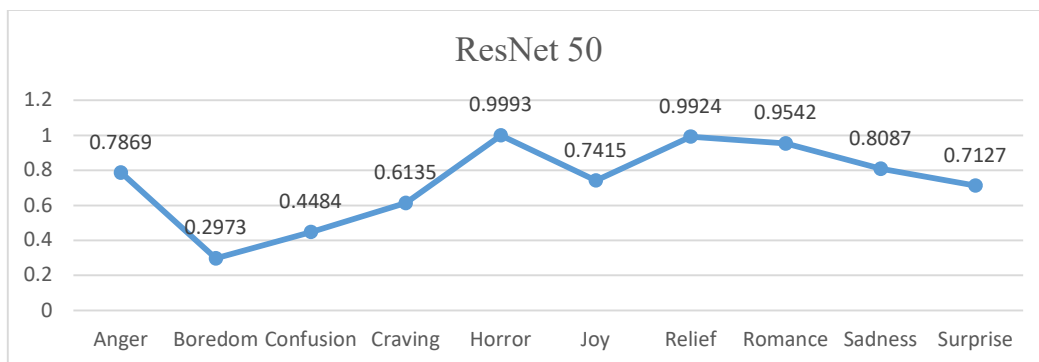
**Figure 7: Boredom**



**Figure 8: Confusion**

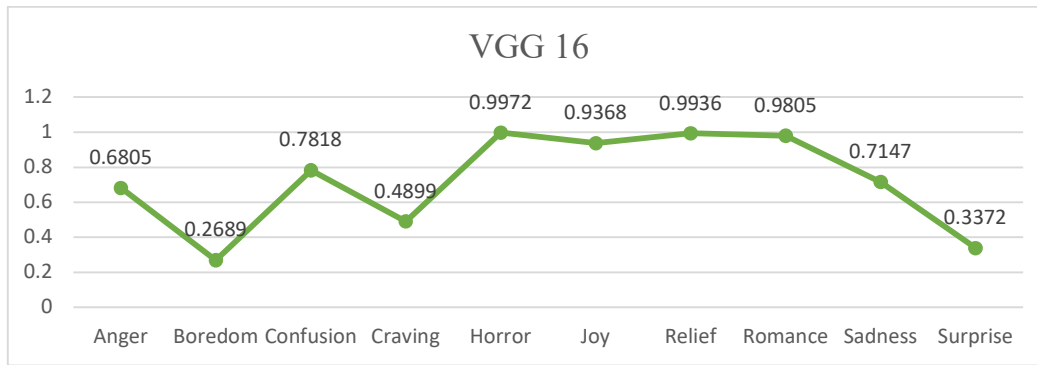
#### 4.1 Image classification using classifiers: RESNET50, VGG16, VGG19

To recognize and categorize pictures, the ImageNet dataset is used to train the ResNet50, VGG16, and VGG19 designs. For our evaluation, we select a random collection of 10 photos and feed them to the ResNet50, Vgg16, and Vgg19 designs. For a picture, many classifiers provide a range of prediction values. Accuracy values for each picture are recorded and contrasted with those of the other networks. The distribution of various accuracy measures for ResNet50, VGG16, and VGG19 for each picture is shown in figures 9, 10, and 11. The chart's x-axis represents the input collection of 10 photos, while the y-axis displays the pictures' predicted values.

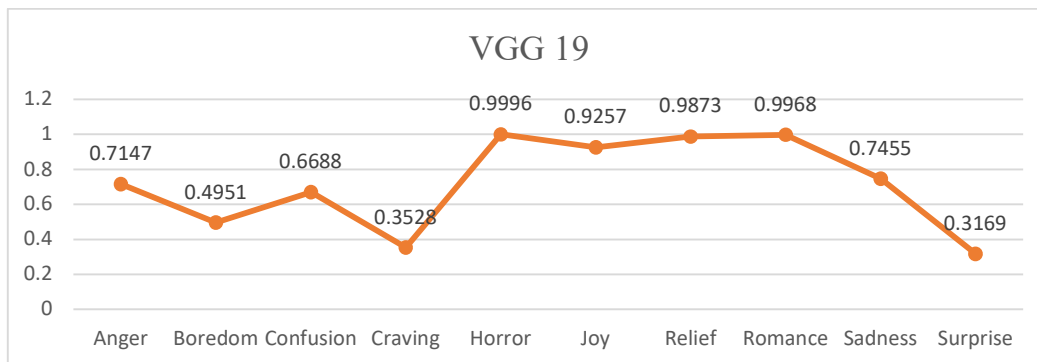


**Figure 9: Images prediction values on Resnet50 framework**





**Figure 10: Images prediction values on VGG16 framework**



**Figure 11: Images prediction values on VGG19 framework.**

#### 4.2 Comparison of images prediction values using classifiers: RESNET50, VGG16, VGG19

The analysis of estimated parameters for each picture is shown in fig. 12. For picture recognition, every picture is submitted to the ResNet50, VGG16, and VGG19 designs. Convolutional Neural Network predicts a specific image and provides the outcome with good predictive accuracy based on the design and dataset. The score of the same picture's forecast changes depending on the design. Figure 13 for three alternative designs illustrates this fluctuation for various accuracy values for all the photos.

Various designs are used to implement convolutional neural network for various image types, and their accuracy values are examined. Based on a study, ResNet50 outperforms VGG16 and VGG19 in terms of accuracy for the provided image. Compared to VGGNet, ResNets have a deeper network and greater precision. They also use less computing power. With the exception of the shortcut link being placed in the residual network, Resnet50's design, which is primarily made up of 3\*3 filters, is identical to that of the VGGNet. In comparison to VGG16 and VGG19, this residual network is more accurate since it solves the degradation issue. A specific image is identified using the pre-trained designs. Depending on how thoroughly the systems are trained using the specific data set, this precision accuracy may vary. For all of the photos, the ResNet50 design yields the best accuracy, as seen in the chart in figure 14.











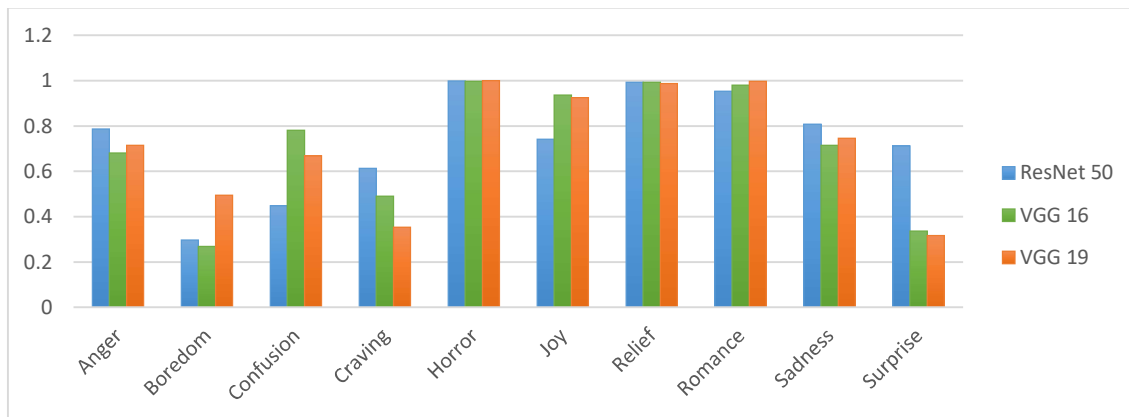
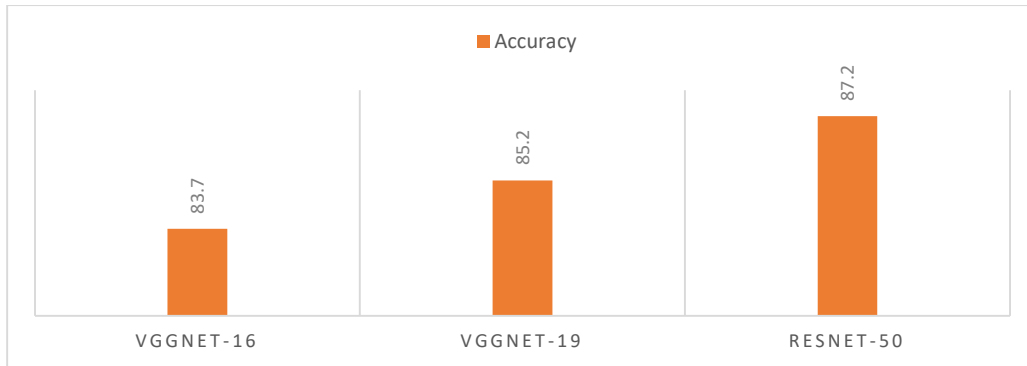
Images	Emotion Name	Prediction Accuracy of 10 images			Deviation in prediction accuracy		
		ResNet 50	VGG 16	VGG 19	(ResNet 50 - VGG16)	(ResNet 50 - VGG 19)	(VGG 16 - VGG19)
	Anger	0.7869	0.6805	0.7147	0.1063	0.0721	0.0342
	Boredom	0.2973	0.2689	0.4951	0.0281	0.1977	0.2259
	Confusion	0.4484	0.7818	0.6688	0.3333	0.2203	0.1129
	Craving	0.6135	0.4899	0.3528	0.1235	0.2606	0.1369
	Horror	0.9993	0.9972	0.9996	0.0020	0.0002	0.0023
	Joy	0.7415	0.9368	0.9257	0.1952	0.1841	0.0110
	Relief	0.9924	0.9936	0.9873	0.0011	0.005	0.0062
	Romance	0.9542	0.9805	0.9968	0.0263	0.0426	0.0162
	Sadness	0.8087	0.7147	0.7455	0.0939	0.0631	0.0306
	Surprise	0.7127	0.3372	0.3169	0.3754	0.3957	0.0202

Figure 12: Image distribution with predicted values



**Figure 13: Comparison of image prediction values using classifiers: RESNET50, VGG16, VGG19**



**Figure 14: Average of image predicted values**

## 5. Conclusion

This study uses Python to deploy a DCNN based on Keras and Tensorflow for picture categorization. This study examined the performance of DCNN networks for prediction accuracy on the ImageNet database. The major goal of this study is to compare the accuracy of several networks on the same dataset and assess how consistently every convolutional neural network makes predictions. For evaluating the efficiency of the networks for various photos, we have included a thorough prediction analysis. The VGG16, VGG19, and ResNet50 nets are applied to the pictures, and the prediction scores are assessed. According to the investigation, ResNet50 can identify photos more accurately than VGG16 and VGG19. Additional high-level networks will be the focus of future development. The use of other pre-trained algorithms for the categorization problem will be the subject of future research.

## References

- [1] Chetanpal Singh, Santoso Wibowo and Srimannarayana Grandhi. "A Deep Learning Approach for Human Face Sentiment Classification", 2021, 21st ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter), IEEE.
- [2] Papiya Das, Anupam Ghosh and Rana Majumdar. "Determining Attention Mechanism for Visual Sentiment Analysis of an Image using SVM Classifier in Deep learning based Architecture", 2020, 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), IEEE.
- [3] Amirhossein Shirzad, Hadi Zare and Hadi Zare. "Deep Learning approach for text, image, and GIF multimodal sentiment analysis", 2020, 10th International Conference on Computer and Knowledge Engineering (ICCKE), IEEE.

- [4] Siqian Chen, Jie Yang, Jia Feng and Yun Gu. “Image Sentiment Analysis using Supervised Collective Matrix Factorization”, 2017, IEEE.
- [5] Xingyue Chen, Yunhong Wang and Qingjie Liu. “Visual and Textual Sentiment Analysis using Deep Fusion Convolutional Neural Networks”, 2017, IEEE.
- [6] Jie Chen, Qirong Mao AND Luoyang Xue. “Visual Sentiment Analysis with Active Learning”, 2016, IEEE.
- [7] Namita Mittal, Divya Sharma and Manju Lata Joshi. “Image Sentiment Analysis using Deep Learning”, 2018, WIC/ACM International Conference on Web Intelligence (WI), IEEE.
- [8] Udit Doshi, Vaibhav Barot and Sachin Gavhane. “Emotion Detection and Sentiment Analysis of Static Images”, 2020, International Conference on Convergence to Digital World – Quo Vadis (ICCDW), IEEE.
- [9] Yilin Wang and Baoxin Li. “Sentiment Analysis for Social Media Images”, 2015, 15th International Conference on Data Mining Workshops, IEEE.
- [10] Rui Man and Ke Lin. “Sentiment Analysis Algorithm Based on BERT and Convolutional Neural Network”, 2021, Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), IEEE.
- [11] 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.
- [12] 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.
- [13] Yingying Pan, Ruimin Lyu, Qinyan Nie and Lei Meng. “Study on the Emotional Image of Calligraphy Strokes based on Sentiment Analysis”, 2020, IEEE.
- [14] 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.

- [15] 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.
- [16] Igor Santos, Nadia Nedjah and Luiza de Macedo Mourelle. “Sentiment Analysis using Convolutional Neural Network with fastText Embeddings”, 2017, IEEE.
- [17] Sani Kaniş 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.
- [18] 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.
- [19] 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.
- [20] 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.
- [21] Haimin Zhang and Min Xu. “Modeling Temporal Information using Discrete Fourier Transform for Recognizing Emotions in User-Generated Videos”, 2016, IEEE.
- [22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei. “ImageNet: A Large-Scale Hierarchical Image Database”, 2009, IEEE.
- [23] Maxime Oquab, Leon Bottou, Ivan Laptev and Josef Sivic. “Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks”, 2014, Conference on Computer Vision and Pattern Recognition, IEEE.
- [24] Ramazan Gokberk Cinbis, Jakob Verbeek and Cordelia Schmid. “Multi-fold MIL Training for Weakly Supervised Object Localization”, 2014, Conference on Computer Vision and Pattern Recognition, IEEE.

- [25] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke and Andrew Rabinovich. “Going Deeper with Convolutions”, 2015, IEEE.
- [26] Maxime Oquab, Leon Bottou, Ivan Laptev and Ivan Laptev. “Is object localization for free? – Weakly-supervised learning with convolutional neural networks”, 2015, IEEE.
- [27] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaying Zhang, Yuxin Peng and Zheng Zhang. “The Application of Two-level Attention Models in Deep Convolutional Neural Network for Fine-grained Image Classification”, 2015, IEEE.
- [28] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik and Andrew Gallagher. “A Mixed Bag of Emotions: Model, Predict, and Transfer Emotion Distributions”, 2015, IEEE.
- [29] Jiajun Wu, Yinan Yu, Chang Huang and Kai Yu. “Deep Multiple Instance Learning for Image Classification and Auto-Annotation”, 2015, IEEE.
- [30] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. “Deep Residual Learning for Image Recognition”, 2016, Conference on Computer Vision and Pattern Recognition, IEEE.