Received: December 13, 2023. Revised: February 7, 2024.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

FPN_GCT: Feature Pyramid Network Gated Channel Transformation for Automatic Image Sentiment Analysis

Chockalingam Arumugam[1]* **Krishnan Nallaperumal[1]** **Saravanan Arumugam[2]**

[1]*Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, 627012, Tamilnadu, India*
[2]*Department of Mechanical and Industrial Engineering, College of Engineering, Northeastern University, Boston, Massachusetts, USA*
* Corresponding author's Email: sasichockalingam5@gmail.com

Abstract: The field of aspect-level sentiment classification has gained significant attention in recent years, but limited image datasets have hindered the effectiveness of neural network models. To address the issue, we present a novel approach, Feature Pyramid Network Gated Channel Transformation (FPN_GCT) that can automatically extract highly fine-grained sentiment information from images. Our approach utilizes Gated Channel Transformation (GCT) for accurate image sentiment analysis and incorporates ResNet and Reduced Layer to enhance the model's novelty. We validate the performance of our model on the Twitter and CrowdFlower sentiment analysis datasets and demonstrate that it outperforms existing techniques including VGG-19, DenseNet121, ResNet50V2, SVM, MemNet, and RAM, with an accuracy of 91.21 for the CrowdFlower dataset and 74.47 for the Twitter dataset. Our approach offers a substantial advancement in the field of image sentiment analysis and has the potential to enhance the accuracy of aspect-level sentiment classification tasks, paving the way for more accurate and efficient sentiment analysis in various applications.

Keywords: Deep learning, Feature pyramid network, Gated channel transformation, Image sentiment analysis, ResNet.

1. Introduction

Automatic sentiment analysis methods are increasingly in demand as there is an increase in the number of evaluations as well as other sentiment-bearing images on the Internet. Images of personal, everyday scenes as well as cartoons and memes that express people's thoughts are frequently shared on social media today. This type of content analysis from social media and data-sharing platforms like Twitter, Flickr, and others can reveal public opinion on topics like presidential elections. Understanding the emotion that an image conveys would also be useful for automatically predicting the emotional tags for that image, such as happiness, terror, and so on. [1] By successively using image sentiment analysis in addition to many other techniques, the system is capable of assessing images and deriving intrinsic sentiments [2,3] from them. Because more people are expressing their thoughts and feelings online, computerised analysis of these sentiments and emotions is becoming more common in the opinion analysis sector. [3] Many industries, including those in education, entertainment, and advertising, have already effectively adopted this trend. [4]

One can communicate their emotions using text, photographs, or videos. [5] Text sentiment analysis has been the subject of many research articles, while image sentiment analysis is less understood. Since individuals are using social media to express their feelings more frequently, this has become one of the most crucial areas of research. A lot of study has been done on this topic over the last several years to perfect the methods and improve the results. Deep learning, a kind of machine learning, is a technique that gives computers the intelligence to understand ideas and use their experiences to learn. A computer can make decisions and retrieve knowledge from actual

experience even without the help of a person. [6] Finding out if an image's sentiments are positive, negative, or neutral is the main objective of sentiment analysis. The type of analysis performed to determine how a user feels about something is called sentiment analysis. For target-dependent sentiment classification, it is still difficult to show how a target and its image are semantically associated.

Image sentiment analysis has made significant strides throughout the last several years, owing to advancements in deep learning algorithm development that have produced remarkable results. Deep learning technology has the ability to complete tasks quickly and accurately, and in some cases, even outperform human intelligence. When discussing neural networks, the word "deep" refers to the total number of hidden layers, which allow for the creation of complex models capable of handling large amounts of data. To train deep learning models for image sentiment analysis, a large quantity of labelled data and a neural network design are both required. Unlike traditional machine learning algorithms that rely on manually extracted features, deep learning architectures learn features and parameters directly from the data. This has been particularly advantageous for image sentiment analysis algorithms, including, but not limited to, Deep Belief Networks (DBNs), Deep Neural Networks (DNNs), Region Neural Networks (RNNs), and Convolutional Neural Networks (CNNs), which have all shown great promise in this field.

One of the main advantages of deep learning in image sentiment analysis is its ability to capture complex features that are difficult to identify using traditional methods. For instance, in the case of facial expressions, deep learning models can detect subtle changes in the face that indicate different emotions, which can be challenging for human analysts. Additionally, deep learning models can be trained to recognize patterns in images that are indicative of a particular sentiment, such as the color scheme, texture, or composition. Despite the advantages of deep learning in image sentiment analysis, there are still obstacles that must be overcome. It may be time-consuming and costly to obtain a large quantity of labelled data to train the models, which is one of the key issues. Additionally, deep learning models can be prone to overfitting, which occurs when the model does well on the training data but not on test data. Therefore, careful model selection, regularization techniques, and data augmentation are important factors to consider when designing a deep learning model for image sentiment analysis.

CNNs have shown promising results in the field of sentiment analysis in images because of their ability to learn in-depth features through deep learning. They can process large datasets and achieve high performance, making automatic image sentiment analysis a practical application for various industries. [7,8] Opinion analysis is frequently viewed as a challenging task because of the significant time and money investment required. [9] It is also important to remember that emotional items cannot sustain themselves, making it challenging to employ them as research resources.

To improve contextual information modelling for automatic image sentiment analysis and to maximise the efficacy of our channel-based interactions, we make use of a technique called "Gated Channel Transformation" (GCT). The normalisation process of GCT creates cooperative or competitive interactions among channels without any additional parameters. To enable learning, we developed a global context-integrating operator that integrates global context and adjusts channel weight before normalization. Every input channel is manipulated by a gated adaptability operator in response to the output of normalization. GCT can be widely deployed with only a small number of lightweight and powerful trainable parameters. To clarify its behaviour, we display the characteristics of the gating adaptation activator. In conclusion, the GCT architecture based on normalisation is a powerful and portable tool for modelling channel interactions in image sentiment analysis. [10]

This paper presents a new model, which we call the “Feature Pyramid Network Gated Channel Transformation” (FPN_GCT), which utilises deep learning techniques to improve the efficacy of sentiment analysis of images. FPN_GCT enables us to better comprehend the emotions and semantics conveyed by images. With deep learning networks, we can train the model in both supervised and unsupervised environments, saving time and effort by automating the feature extraction process without the need for manual intervention. This approach offers a powerful solution to the challenge of sentiment analysis, paving the way for more accurate and efficient image analysis in a wide range of industries.

The following are our primary contributions:

- To address the difficulties associated with image sentiment analysis, we devised a novel strategy that makes use of Gated Channel Transformation (GCT) named FPN_GCT.
- We used layers, such as reduced layers and completely connected layers, to reduce overfitting. By adding a novel technique that makes use of ResNet18, RLF, and essential layers, our model became capable of determining the values that give the model its best accuracy.

- Our findings demonstrate that the proposed model for image sentiment analysis outperforms state-of-the-art methods on a prominent dataset.

This paper proposes Feature Pyramid Network Gated Channel Transformation (FPN_GCT) for automated image sentiment analysis. The paper is organised as follows: The existing research on sentiment analysis of images using various techniques is covered in Section 2. Section 3 proposes the FPN_GCT along with a schematic diagram that outlines the process for detecting image polarity. The dataset details and the experimental results are outlined in Section 4. A comprehensive comparison of the numerous existing methods is also covered. Section 5 presents the conclusion of our paper.

2. Related work

In this section, we looked at a few earlier studies on image sentiment analysis. Several techniques for image sentiment analysis include the collection of low-level characteristics, semantic features, machine and deep learning models. This paper also examines the rapidly changing state of sentiment classification research and other techniques involved in our proposed approach.

A multimodal sentiment analysis approach based on interactive transformers and soft mapping is presented in. [11] After data fusion, their model is able to fully account for the link between data from various modalities, which is useful for sentiment analysis. It has given equivalent findings on the CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset; however, it still has significant issues. Only the data from the textual and auditory modalities are used, and the data from the visual modalities is not fully utilized. When visual modalities are attempted to be added, the results are not sufficient. In the following stage, they continued looking for a technique to incorporate visual data; therefore, the facial expressions and body language of characters often have complicated feelings that are particularly useful for emotion recognition. The performance evaluation for the Multimodal EmotionLines Dataset (MELD) dataset also reveals issues. They ignore the reality that people's emotions interact in multi-person discourse contexts. As an illustration, when someone externally shows bad feelings, other people's emotional reactions will also alter negatively.

Cluster Correlation Mining (CCM) and Active Sample Refinement (ASR) methods were introduced in [12] to perform picture sentiment analysis. The model was tested for efficacy and reliability on two standard datasets. The findings showed that while the ASR technique improved the performance, the addition of CCM, which facilitates efficient and resilient feature learning, has the potential to further enhance overall performance. The ASR approach also proved to be a valuable complement to the current data augmentation techniques, significantly improving the baseline. However, there are two drawbacks to this approach. Firstly, the model's foundation relies on machine learning, which although highly effective, is not end-to-end, limiting its practical utility. Secondly, further research is needed to improve the performance of the Twitter dataset.

[13] Introduces a multimodal sentiment analysis that makes use of social media photos and textual data to categorise polarity into three classes. The independent classifications of the textual and image are combined into a final classification using automated machine learning (AutoML). To do the individual classifications, deep neural network capabilities were used. They analysed the performance of numerous networks for this and chose the top one for the text and image components. In order to construct the merging approach, they utilised AutoML to conduct a random check to choose the ideal model for the ultimate categorization. Using a dataset of more than 4,70,000 tweets, each of which contains both textual and visual content, they assessed the technique against the backdrop of the multimodal sentiment analysis task.

Peng Chen [14] suggested a novel paradigm called Recurrent Attention on Memory (RAM) to address the aforementioned issues in target sentiment analysis. To build the memory (i.e., the states of time steps generated by Long Short-Term Memory (LSTM)) from the input, the framework first implements a Bidirectional LSTM (BiLSTM), as bidirectional Recurrent Neural Networks (RNNs) were proven to be successful for a similar reason in machine translation. So that distinct targets from the same phrase have their own custom-made memories, the memory slices are then weighted based on their relative positions to the target. The position-weighted memory is then the focus of many attentions, and the attention outcomes are nonlinearly combined with a recurrent network, or Gated Recurrent Units (GRUs). Finally, softmax was used to forecast the sentiment towards the target using the GRU network's output. Using an improved transfer learning model, the algorithm's capacity to correctly anticipate attitudes is demonstrated. [15] The Visual Geometry Group-19 (VGG-19) model is an option given that it is a

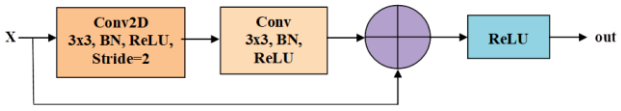

Figure. 1 Residual Block

much enhanced variant of the widely used Visual Geometry Group-16 (VGG-16) model for image classification tasks. [16]

In order to efficiently design the intra-modality interactions, such as entity-text/image alignments, in addition to the inter-modality discussions, like text-image alignments, in the inquiry of entity-level multimodal sentiment classification, an entity-sensitive attention and fusion network was proposed. [17] It outperforms numerous unimodal and multimodal techniques in terms of efficiency, as per the experimental data on two multimodal datasets and one unimodal benchmark dataset. The fundamental flaw in that strategy is the assumption that named entity recognition techniques have been used to provide or retrieve the entities in each phrase.

Plutchik, R. et al. [18] gave a summary of the state-of-the-art model and reviewed pertinent works on image sentiment analysis. The three main views of emotional models, defining datasets, and designing features are used to show and debate the design concepts of image sentiment analysis systems. They identify and look at the components that can affect how individuals feel in different ways about an image, and they also present a discussion of recent challenges. The identification of fundamental emotions has been the subject of numerous studies. The most current deep learning approaches and machine learning techniques can yield excellent outcomes as long as these systems are taught using very large-scale datasets like Visual Sentiment Ontology (VSO). [19] It is straightforward to collect such datasets by using social media platforms where people regularly share pictures. These datasets enabled the development of machine learning systems that require a significant quantity of data to converge. There isn't a proven approach for choosing visual features yet that takes care of the problem. Recent findings in image sentiment analysis point to the need for more research into representational learning methods like CNN and multi-modal embedding.

By using our proposed Feature Pyramid Network Gated Channel Transformation (FPN_GCT) model that was trained on a huge collection of data, the shortcomings discovered in prior research on image sentiment analysis can be removed. The performance of deep learning is reliant on ample data support. The learning method performs less effectively when getting a large amount of labelled data samples is difficult. The constructed deep learning network topology is prone to overfitting when there is not enough data. In image sentiment analysis, our system fared better than competing systems.

3. Proposed feature pyramid network gated channel transformation (FPN_GCT)

The proposed Feature Pyramid Network Gated Channel Transformation (FPN_GCT) is described in detail in this section. ResNet18, Reduced Layer, GCT Attention, Pooling Layer, and Fully Connected Layer are the main components of our proposed methodology.

3.1 Proposed feature pyramid network gated channel transformation (FPN_GCT)

ResNet18 has been used as a backbone in our proposed approach. The residual layer is depicted in Fig. 1.

In the residual layer, if x is the input, it will be sent into the convolutional layer 3×3, BN, and ReLU with stride 2. The result of this convolutional layer will be fed into another convolutional layer with a 3×3 filter, BN, and ReLU and concatenated with ReLU to yield the outcome.

3.2 Proposed feature pyramid network gated channel transformation (FPN_GCT)

Architecture of the GCT is depicted in Fig. 2. A normalizing technique is used by the Gated Channel Transformation (GCT) to establish rivalry or cooperation between channels. Contextual information modelling that is highly effective channel-wise is done using GCT. [20]

In a convolutional neural network, let an activation feature be denoted by the notation $X \in \mathbb{R}^{C\times H\times W}$, where H and W represent the spatial height and width, respectively, and C represents the number of channels. The following is a typical GCT transformation:

$$\hat{X} = F(X|\omega,\sigma,\lambda), \omega,\sigma,\lambda \in \mathbb{R}^{c} \quad (1)$$

The trainable parameters in this example are ω, σ and λ.The embedding outputs are adjusted by embedding weight α. The gating weight σ and bias λ regulate the gate's activation.

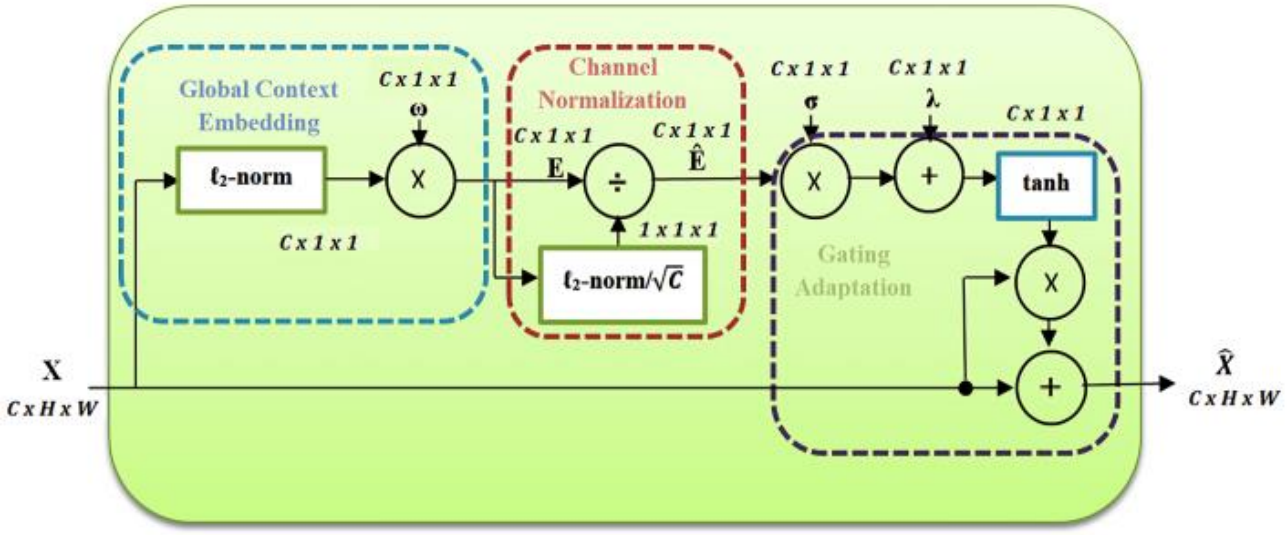

Figure. 2 GCT Architecture

3.2.1 Global context embedding

Each channel contains global context information. Outside of the convolutional layers' limited receptive fields, the module can make use of global contextual information. The module is defined as follows, given the embedding weight α = [α1, ..., αC]:

$$s_c = \omega_c \left|\left|x_c\right|\right|2 = \omega_c\{[\sum_{i=1}^{H}\sum_{j=1}^{W}\left(x_c^{i,j}\right)^2]+\in\}^{\frac{1}{2}} \quad (2)$$

where ∈ is a tiny constant used to get around the issue with zero-point derivation. GCT does not aggregate channel context using global average pooling (GAP), in contrast to SE. In some rare circumstances, GAP might not work. The ℓ1-norm is more computationally effective in this situation. Additionally, as each channel should have a different significance, we employ trainable parameters ω_c to modify the weight of every channel. If ω_c is near 0, the channel c won't be considered in the channel normalizations. To put it another way, the gating weight ω, enables GCT to discover the conditions under which one channel differs from other channels.

3.2.2 Channel normalization

With minimal computational resources and robust training results, normalization techniques can establish a competitive relationship between neurons (or channels). [21,22] We apply a ℓ2 normalization, called channel normalization, to operate across channels in a manner akin to Local Response Normalization (LRN). The formula for channel normalizations is: $s = [s_1, \ldots, s_c]$,

$$\widehat{s_c} = \frac{\sqrt{Cs_c}}{||s||_2} = \frac{\sqrt{Cs_c}}{[(\sum_{c=1}^{C} s_c^2)+\in]^{\frac{1}{2}}} \quad (3)$$

where ∈ is a small constant value. The scalar $\sqrt{C}$ is used to normalize the scale of $\widehat{s_c}$, avoiding a too small scale of $\widehat{s_c}$ when C is large.

3.2.3 Gating adaptation

We used a gating mechanism known as "gating adaptation" to modify the real characteristic. Our GCT can promote competitiveness and teamwork during the training process by introducing the gating mechanism. We create the gating function shown below by setting the gating weight sigma = $\sigma = [\sigma_1, \ldots, \sigma_c]$ and the gating biases $\lambda = [\lambda_1, \ldots, \lambda_c]$.

$$\widehat{x_c} = x_c[1 + \tanh(\sigma_c \widehat{s_c} + \lambda_c)] \quad (4)$$

Every original channel's scale will be adjusted by the gate that corresponds to it, $1 + \tanh(\sigma_c \widehat{s_c} + \lambda_c)$. We created the trainable weight and bias, σ and λ, for learning to carefully manage the activation of gate channels because the channel normalization is parameter-free. Only the rivalries between the neurons are beneficial to LRN.[22] However, by merging normalization techniques and gating mechanisms, more different forms of connections between various channels can be described using GCT. GCT stimulates a channel to interact with some other channels, like LRN, when that channel's gating weight (sigma c) is positively active. While the gating weight is negatively engaged, GCT stimulates this channel to collaborate with the

other channels. In Section 3.4, we evaluate these interactions between adaptive channels.

Additionally, when the gating weight and bias are zero, this gate function permits the original features to pass to the subsequent layer, which is

$$\hat{x} = F(x|\omega, 0,0) = 1x = x \quad (5)$$

Identity mapping can enhance the capacity of deep networks to improve their robustness against degradation problems. This notion is also advantageous to ResNets. Therefore, we suggest setting bias to 0 when the GCT layers are initialized. This will improve the end performance of GCT and make the first phases of the training process more stable.

3.2 Feature pyramid

A feature pyramid is made to mix the characteristics of many convolutional network levels in order to more effectively detect objects of various sizes. Essentially, it means that we aggregate feature maps from various network levels and use those feature maps to produce a higher-level feature map. We accomplish this by building robust semantic feature maps at each size. FPN is used in both single-stage and dual-stage architectures for object detection. To generate high-level semantic feature maps of any size, a top-down framework with lateral links is required. This procedure, also referred to as a "Feature Pyramid Network" (FPN), significantly improves as a general feature extractor in many applications.

Using ConvNet's hierarchical feature pyramid, which includes low-level to high-level semantics, we aim to construct a feature pyramid with consistent high-level semantics across all levels. With a single-scale picture of any size, our approach produces proportionately scaled feature maps at numerous layers, fully convolutional. In this study, we describe findings utilising ResNets18, which are used as the backbone. As explained in the following, our pyramid is built using lateral connections, a top-down pathway method.

3.3 Feature pyramid network gated channel transformation (FPN_GCT)

In Fig. 3, the entire architecture of the proposed FPN_GCT is depicted.

ResNet18 is the backbone of the proposed architecture, and the last four features such as F1, F2, F3, and F4 from ResNet are taken into account.

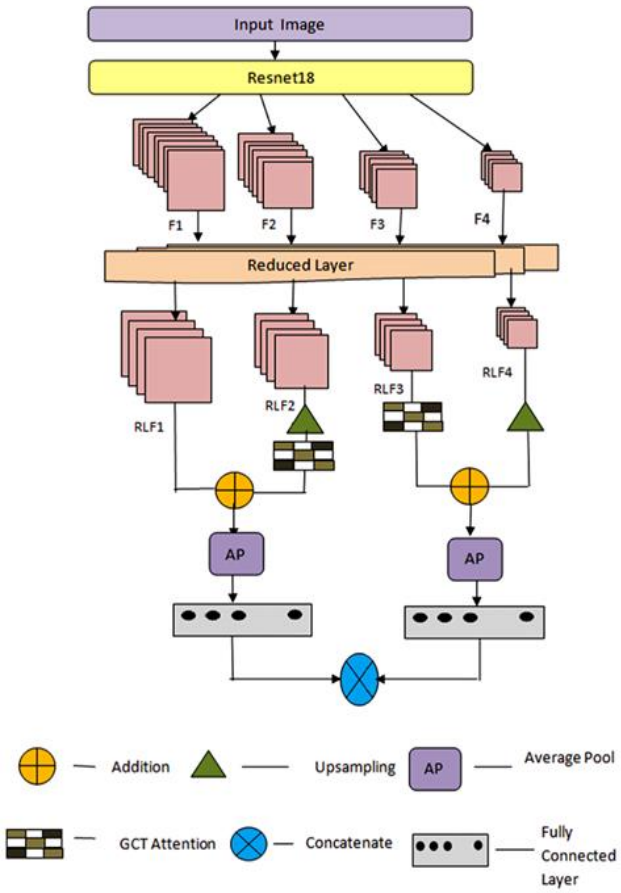

Figure. 3 Proposed FPN_GCT Architecture

ResNet18 will receive the input image with a filter size of 256×256×3. The ResNet feature F1 has a filter size of 64×64×64, F2 has 128×32×32, F3 has 256×16×16 and F4 has 512×8×8.

$$F1_{64} = ResNet\ (input\ image) \quad (6)$$

$$F2_{128} = ResNet\ (input\ image) \quad (7)$$

$$F3_{256} = ResNet\ (input\ image) \quad (8)$$

$$F4_{512} = ResNet\ (input\ image) \quad (9)$$

Since the sizes of these features are varied, we used a reduced layer to reduce their size to be the same. From RLF1 to RLF4, the size will decrease. This process of reduced layer is indicated in the equation as follows:

$$RLF1_{64} = ReduceLayer\ (F1) \quad (10)$$

$$RLF2_{32} = ReduceLayer\ (F2) \quad (11)$$

$$RLF3_{16} = ReduceLayer\ (F3) \quad (12)$$

$$RLF4_8 = ReduceLayer\ (F4) \quad (13)$$

The output from the reduced layer RLF2 is sent into the upsampling UP1 with a size of 32 and 64, and then fed into the GCT attention layer with size 64, 16, 16. After combining the output from the GCT layer *gct1* with the RLF1, the resulting *add1* is transmitted into the average pool layer, *AP1*. On the other hand, the GCT attention layer *gct2* receives the output from the reduced layer RLF3, and their result is combined with the upsampling UP2 of RLF4. The results of the addition, *add2*, are subsequently transmitted to the average pool layer, *AP2*. The results of these two average pool layers, having sizes 64, 32, 32 and 64, 8, 8 were combined after being fed into separate fully connected layers, *FL1* and *FL2*. Finally, the outcome will be sending it into the softmax layer, *SoftM*. This process has been presented in the following equations:

$$UP1 = upsampling\ (RLF2) \quad (14)$$

$$GCT132 = gatechtr\ (UP1) \quad (15)$$

$$Add1 = RLF1 + GCT1 \quad (16)$$

$$GCT216 = gatechtr\ (RLF3) \quad (17)$$

$$UP2 = upsampling\ (RLF4) \quad (18)$$

$$Add2 = GCT2 + UP2 \quad (19)$$

$$AP1 = avgpool\ (Add1) \quad (20)$$

$$AP2 = avgpool\ (Add2) \quad (21)$$

$$FL1 = fullyconnected\ (AP1) \quad (22)$$

$$FL2 = fullyconnected\ (AP2) \quad (23)$$

$$Concat = concatenate\ (FL1, FL2) \quad (24)$$

$$SoftM\ =\ softmax\ (Concat) \quad (25)$$

4. Experimental result and analysis

4.1 Evaluation metrics

In this research, the proposed method was evaluated based on various metrics to ensure its effectiveness in interpreting sentiment in images. The metrics used for evaluation include precision, recall, accuracy and F1-score. These metrics were calculated based on the datasets available for image sentiment analysis. The results of the evaluation metrics were used to compare the proposed method with existing models and to demonstrate its superiority in sentiment analysis of images. Overall, the evaluation metrics used in this study provide a comprehensive and reliable assessment of the efficacy of the proposed method. Below are the formulas used for evaluation metrics:

$$Accuracy = (TP+TN) / (TP+TN+FP+FN) \quad (26)$$

$$Precision = TP / (TP + FP) \quad (27)$$

$$Recall = TP / (TP + FN) \quad (28)$$

$$F1\text{-}Score = (2 \times Precision \times Recall) / (Precision + Recall) \quad (29)$$

In binary classification, a false positive (FP) mistake happens when an outcome of the test falsely suggests the availability of a situation when one is not there. In contrast, a false negative (FN) error happens when a condition is present but the test result incorrectly fails to signal it. For a test result to be considered true positive (TP), it must accurately detect the existence of a condition or situation, whereas for a test result to be considered true negative (TN), it must accurately detect the absence of the condition or situation.

4.2 Linking research survey and comparative analysis

We have conducted a detailed research survey on Visual Sentiment Analysis, forming the basis for our comparative analysis with our FPN_GCT model. This section aims to clearly link the insights from the survey with our choice of comparison targets, such as VGG-19, DenseNet121, ResNet50V2, SVM, MemNet, and RAM. These models were selected for their relevance in terms of methodologies and datasets, offering a meaningful comparison to showcase the advancements of FPN_GCT.

Our FPN_GCT model addresses specific challenges highlighted in the survey. The comparison with selected models goes beyond performance metrics; it explores how FPN_GCT advances the field, particularly in analysing social media data from Twitter and CrowdFlower datasets.

4.3 Dataset description

The Twitter Dataset and CrowdFlower Dataset are the two datasets utilized in this work for analysing the performance of the proposed work. The description of these datasets is explained below.

4.3.1 Twitter dataset

Twitter posts from the TWITTER, [8] that is, TWITTER17, and TWITTER15 image datasets, were used. Datasets that contain contradictory polarity-based information are removed. The information about the datasets is shown in Table 1. This displays the number of training and test sets handy for each dataset in positive, negative, and neutral samples.

Setting parameters: the initial learning rate for each model is specifically set to 0.001, and the Adam optimizer is used to plan the learning rate. Additionally, 32 and 50, respectively, are specified for the batch size and num epoch. Each model was put into practise using PyTorch, and testing was conducted using an NVIDIA RTX 2060 GPU with 6GB of RAM.

Table 1. Statistics of Twitter Datasets

Data set	Category	Train	Development	Test
TWITTER17	Positive Samples	1508	515	493
	Negative Samples	416	144	168
	Neutral Samples	1638	517	573
TWITTER15	Positive Samples	928	303	317
	Negative Samples	368	149	113
	Neutral Samples	1883	670	607

Table 2. Performance analysis of Twitter17 dataset in terms of Precision, Recall, F1-Score and Accuracy

Models	Precision	Recall	F1 Scor	Acc.
SVM [14]	NA	NA	62.11	62.25
MemNet [14]	NA	NA	68.21	69.62
RAM [14]	NA	NA	69.80	70.52
VGG-19 [23]	56.30	58.82	57.53	60.29
DenseNet121 [23]	61.04	63.58	62.28	65.55
ResNet50V2 [23]	58.98	61.33	60.13	63.53
FPN_GCT	**70.57**	**74.57**	**72.52**	**74.47**

Table 3. Performance analysis of Twitter15 dataset in terms of Precision, Recall, F1-Score and Accuracy

Models	Precision	Recall	F1 Scor	Acc.
SVM [14]	NA	NA	63.30	63.40
MemNet [14]	NA	NA	66.91	68.50
RAM [14]	NA	NA	67.30	69.36
VGG-19 [23]	50.09	51.47	50.77	61.13
DenseNet121 [23]	58.01	60.47	59.21	65.95
ResNet50V2 [23]	55.13	57.92	56.49	64.03
FPN_GCT	**66.96**	**70.64**	**68.75**	**72.22**

The performance analysis of the proposed FPN_GCT for the twitter datasets in terms of precision, recall, accuracy and F1-Score are shown in Table 2 and 3.

The FPN_GCT model shows a significant improvement in the Twitter17 dataset over methods like SVM [14], MemNet [14], RAM [14], VGG-19 [23], DenseNet121 [23], and ResNet50V2 [23]. In precision, it leads VGG-19 [23] by 14.27, DenseNet121 [23] by 9.56, and ResNet50V2 [23] by 11.59. Similarly, its accuracy surpasses SVM [14] by 12.22, MemNet [14] by 4.85, RAM [14] by 3.95, VGG-19 [23] by 14.18, DenseNet121 [23] by 8.92, and ResNet50V2 [23] by 10.94. This pattern of enhanced performance extends to recall and F1-score, underscoring FPN_GCT's robustness in sentiment analysis.

In the Twitter15 dataset, the FPN_GCT model notably outperforms the above mentioned methods. It leads in precision by 16.87 over VGG-19 [23], 8.95 over DenseNet121 [23], and 11.83 over ResNet50V2 [23]. Its accuracy exceeds SVM [14] by 8.82, MemNet [14] by 3.72, RAM [14] by 2.86, VGG-19 [23] by 11.09, DenseNet121 [23] by 6.27, and ResNet50V2 [23] by 8.19. This consistent superior performance is also reflected in recall and F1-score metrics, further affirming FPN_GCT's effectiveness in image sentiment analysis. This marked improvement over established models underscores the effectiveness of the FPN_GCT approach in handling complex social media data for sentiment analysis.

4.3.2 CrowdFlower dataset

The CrowdFlower image sentiment dataset [23] was used to evaluate the proposed FPN_GCT model.

Table 4. Statistics of CrowdFlower Datasets

Dataset	Category	Train	Test	Total
Crowd-Flower	Positive Samples	514	128	642
	Negative Samples	286	72	358

Table 5. Performance analysis of CrowdFlower dataset in terms of Precision, Recall, F1-Score and Accuracy

Models	Precision	Recall	F1 Score	Acc.
VGG-19 [23]	73.26	73.89	73.58	73
DenseNet121 [23]	88.86	89.11	88.99	89
ResNet50V2 [23]	74.77	73.25	74.00	75
FPN_GCT	**90.86**	**91.12**	**90.99**	**91.21**

The dataset includes a sizable collection of URLs for images that human experts have annotated to indicate and label their emotional tone, which ranges from neutral to very positive or negative. To begin, we selected a total of 1000 photos from this dataset, all of which fell into either the positive or negative categories, as shown in Table 4.

The performance analysis of the proposed FPN_GCT for the CrowdFlower dataset in terms of precision, recall, accuracy and F1-Score are shown in Table 5.

Compared to previous methods such as the VGG-19 Model, [23] DenseNet121, [23] and ResNet50V2, [23] the proposed FPN_GCT model achieves better results. Specifically, in terms of precision, FPN_GCT outperforms VGG-19 by +17.60, DenseNet121 by +2.00, and ResNet50v2 by +16.09.

Compared to previous methods such as the VGG-19 Model, [23] DenseNet121, [23] and ResNet50V2,

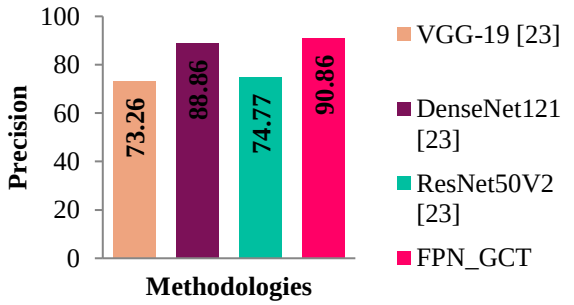

Figure. 4 Performance analysis in terms of precision for CrowdFlower dataset

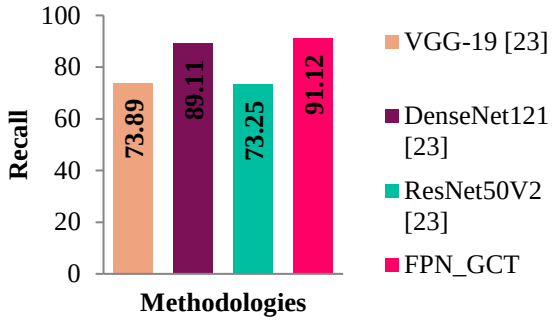

Figure. 5 Performance analysis in terms of recall for CrowdFlower dataset

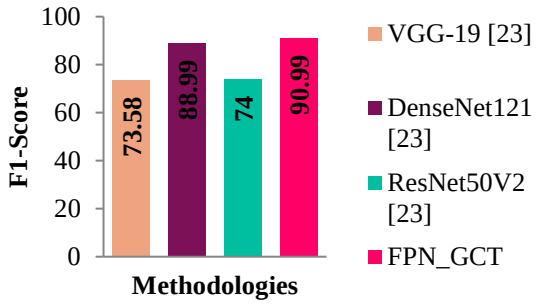

Figure. 6 Performance analysis in terms of F1-Score for CrowdFlower dataset

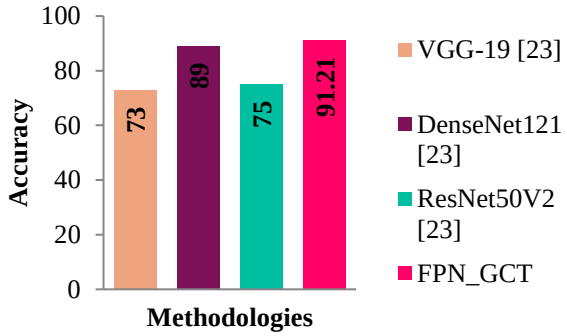

Figure. 7 Performance analysis in terms of accuracy for CrowdFlower dataset

[23] the proposed FPN_GCT model achieves better results. Specifically, in terms of recall, FPN_GCT outperforms VGG-19 [23] by 17.23 points, DenseNet121 [23] by 2.01, and ResNet50v2 [23] by 17.87 points.

The proposed FPN_GCT gives better result than the previous methods like VGG-19 Model, [23] DenseNet121, [23] ResNet50V2. [23] In terms of F1-Score, FPN_GCT outperforms VGG-19 [23] by +17.41 points, DenseNet121 [23] by 2, and ResNet50v2 [23] by +16.99 points.

The proposed FPN_GCT model outperforms the previous methods, including VGG-19 Model, [23] DenseNet121, [23] and ResNet50V2, [23] with better results. More specifically, concerning accuracy,

FPN_GCT achieves a significant improvement of +18.21 points compared to VGG-19, [23] +2.21 points compared to DenseNet121, and +16.21 points compared to ResNet50v2. [23]

Figs. 4-7 shows the performance analysis of the proposed method for the CrowdFlower dataset in terms of precision, recall, F1-Score and accuracy respectively.

4. Conclusion

This paper presents the use of the Feature Pyramid Network Gated Channel Transformation (FPN_GCT) as a powerful tool for sentiment classification in images. The scientific contribution of this research lies primarily in the development and validation of a novel deep learning model, FPN_GCT. Our model employs Feature Pyramid Network with Gated Channel Transformation to enhance image sentiment analysis.

Key contributions of our model include:

- Utilization of Gated Channel Transformation, leading to improved accuracy in sentiment analysis. Specifically, our model achieved an accuracy of 74.47% on the Twitter dataset and 91.21% on the CrowdFlower dataset, surpassing the benchmarks set by existing models.
- Adoption of ResNet as the backbone architecture, which enhances feature extraction capabilities.
- Implementation of innovative layering techniques to effectively reduce overfitting.

Our experimental findings on the Twitter dataset and the CrowdFlower dataset, particularly the comparative analysis with existing techniques like VGG-19, DenseNet121, ResNet50V2, SVM, MemNet, and RAM, demonstrate the superior performance of our approach. The FPN_GCT model outperformed these established techniques in terms of precision, recall, F1-score, and accuracy, showcasing its robustness in automatic image sentiment analysis.

The significant improvements in performance metrics highlight the practical impact of our research in the field of visual sentiment analysis. These advancements, backed by concrete data from our experiments, reinforce the FPN_GCT model's potential in handling diverse and complex image datasets, making a substantial contribution to the domain of deep learning-based image sentiment analysis.

The superior performance of our approach can be attributed to the use of the FPN architecture and the Gated Channel Transformation technique. Overall, this research offers substantial advancements in automatic image sentiment analysis.

Conflicts of Interest

The authors declare that they have no conflict of interest. It is worth noting that the 1st author (corresponding author) and the 3rd author are brothers. However, this familial relationship does not constitute a conflict of interest that would influence the objectivity or integrity of the research conducted and reported herein.

Author Contributions

The contributions of authors are as follows: Conceptualization, Chockalingam Arumugam; Methodology, Chockalingam Arumugam; Software, Saravanan Arumugam; Validation, Krishnan Nallaperumal; Formal analysis, Chockalingam Arumugam; investigation, Chockalingam Arumugam; resources, Saravanan Arumugam; Data Curation, Chockalingam Arumugam; Writing-original draft preparation, Chockalingam Arumugam and Krishnan Nallaperumal; writing—review and editing, Chockalingam Arumugam; visualization, Chockalingam Arumugam; supervision, Krishnan Nallaperumal.

Acknowledgments

No funds, grants, or other support was received.

References

[1] V. Gajarla, and A. Gupta, "Emotion Detection and Sentiment Analysis of Images", *Georgia Institute of Technology*, Vol. 1, pp. 1-4, 2015.

[2] P. K. Chaubey, T. K. Arora, K. B. Raj, G. R. Asha, G. Mishra, S. C. Guptav, M. Altuwairiqi, and M. Alhassan, "Sentiment Analysis of Image with Text Caption using Deep Learning Techniques", *Computational Intelligence and Neuroscience*, Vol. 2022, pp. 1-11, 2022.

[3] L. Wu, M. Qi, M. Jian, and H. Zhang, "Visual Sentiment Analysis by Combining Global and Local Information", *Neural Processing Letters*, Vol. 51, No. 3, pp. 2063-2075, 2019.

[4] E. Chu., and D. Roy., "Audio-visual sentiment analysis for learning emotional arcs in movies", In: *Proc. of the IEEE International Conference on Data Mining (ICDM)*, New Orleans, LA, USA, pp. 829-834, 2017.

[5] R. Singh, P. Shukla, P. Rawat, and P. K. Shukla, "Invisible Medical Image Watermarking using Edge Detection And Discrete Wavelet Transform Coefficients", *International Journal of*

Innovative Technology and Exploring Engineering, Vol. 9, No. 1, pp. 5074-5080, 2019.

[6] A. Ortis, G. M. Farinella, and S. Battiato, "Survey on visual sentiment analysis", *IET Image Processing*, Vol. 14, No. 8, pp. 1440-1456, 2020.

[7] N. K. Rathore, N. K. Jain, P. K. Shukla, and U. R. Rawat, "Image forgery detection using singular value decomposition with some attacks", *National Academy Science Letters*, Vol. 44, No. 4, pp. 331-338, 2021.

[8] F. Al-Turjman, A. Nayyar, A. Devi, and P. K. Shukla, *Intelligence of Bings: AI-IoT Based Critical-Applications and Innovations*, Springer, New York, 2021.

[9] Y. Vijay, S. Goyal, R. Sharma, and U. Mamodiya, "Green building design and security system", *Journal of Web Engineering & Technology*, Vol. 6, No. 2, pp. 10-14, 2019.

[10] PP. N. Srinivasu, A. Kumar Bhoi, R. Jhaveri, G. T. Reddy, and M. Bilal, "Probabilistic deep Q network for real-time path planning in censorious robotic procedures using force sensors", *Journal of Real-Time Image Processing*, Vol. 18, pp. 1773-1785, 2021.

[11] S. Stalin, P. Maheshwary, and P. K. Shukla, "Nonlinear 2D chaotic map and DNA (NL2DCM-DNA) sequences-based fast and secure block image encryption", *Emerging Technologies in Data Mining and Information Security*, Vol. 1300, pp. 69-76, 2021.

[12] G. Khambra, and P. Shukla, "Novel machine learning applications on fly ash based concrete: an overview", In: *MaterialsToday: Proceedings*, Vol.80, No. 3, pp. 3411-3417, 2023.

[13] G. Kaur, K. S. Saini, D. Singh, and M. Kaur, "A comprehensive study on computational pansharpening techniques for remote sensing images", *Archives of Computational Methods in Engineering*, Vol. 28, No. 7, pp. 4961-4978, 2021.

[14] P. Chen, Z. Sun, L. Bing, and W. Yang., "Recurrent Attention Network on Memory for Aspect Sentiment Analysis", In: *Proc. of the Conference on Empirical Methods in Natural Language Processing, pp.* 452-461, 2017.

[15] R. Socher, J. Pennington, E. H. Huang, Y. Ng. Andrew, and C. D. Manning., "Semi-supervised recursive auto encoders for predicting sentiment distributions", In: *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pp. 151-161, 2011.

[16] O. Appel, F. Chiclana, J. Carter, and H. Fujita, "A hybrid approach to the sentiment analysis problem at the sentence level", *Knowledge-Based Systems*, Vol. 108, pp. 110-124, 2016.

[17] J. Yu, J. Jiang, and R. Xia, "Entity-Sensitive Attention and Fusion Network for Entity-Level Multimodal Sentiment Classification", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 429-439, 2020.

[18] K. Schouten, and F. Frasincar, "Survey on Aspect-Level Sentiment Analysis", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, pp. 813-830, 2016.

[19] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter sentiment classification", In: *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 151-160, 2011.

[20] Z. Yang, L. Zhu, Y. Wu, and Y. Yang., "Gated Channel Transformation for Visual Recognition", In: *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11791-11800, 2020.

[21] S. Ioffe, and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", In: *Proc. of International Conference on Machine Learning*, Vol. 37, pp. 448-456, 2015.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", *Communications of the ACM*, Vol. 60, pp. 84-90, 2017.

[23] G. Chandrasekaran, N. Antoanela, G. Andrei, C. Monica, and J. Hemanth, "Visual Sentiment Analysis Using Deep Learning Models with Social Media Data", *Applied Sciences*, Vol. 12, pp. 1030, 2022.