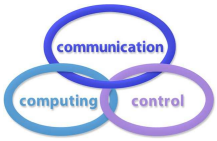
INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 19, Issue: 5, Month: October, Year: 2024
Article Number: 6697, https://doi.org/10.15837/ijccc.2024.5.6697



CCC Publications

EfficientNet Convolutional Neural Network with Gram Matrices Modules for Predicting Sadness Emotion

M. Motiejauskas, G. Dzemyda

Modestas Motiejauskas*
Institute of Data Science and Digital Technologies
Vilnius University, Lithuania
Akademijos str., Vilnius, LT-08412, Lithuania
*Corresponding author: modestas.motiejauskas@mif.stud.vu.lt

Gintautas Dzemyda
Institute of Data Science and Digital Technologies
Vilnius University, Lithuania
Akademijos str., Vilnius, LT-08412, Lithuania
gintautas.dzemyda@mif.vu.lt

Abstract

Images are becoming an attractive area of emotional analysis. Recognising emotions in the images of general nature is gaining more and more research attention. Such emotion recognition is more sophisticated and different from conventional computer tasks. Due to human subjectivity, ambiguous judgments, cultural and personal differences, there is no an unambiguous model for such emotion assessment. In this paper, we have chosen sadness as the main emotion, which has significant impact to the richness of human experience and the depth of personal meaning. The main hypothesis of our research is that by extending the capabilities of convolutional neural networks to integrate both deep and shallow layer feature maps, it is possible to improve the detection of sadness emotion in images. We have suggested integration of the different convolutional layers by taking the learned features from the selected layers and applying a pairwise operation to compute the Gram matrices of feature sub-maps. Our findings show that this approach improves the network's ability to recognize sadness in the context of binary classification, resulting in a higher emotion recognition accuracy. We experimentally evaluated the proposed network for the stated binary classification problem under different parameters and datasets. The results demonstrate that the improved network achieves improved accuracy as compared to the baseline (EfficientNetV2) and the previous state-of-the-art model.

Keywords: EfficientNetV2, Gram matrix, emotion prediction, images of general nature, sadness emotion.

1 Introduction

With the current spread of the Internet of Things and smart devices, the amount of data, much of it visual, has increased significantly. Vision is the main source of information that people receive from the outside. Images can also express meaning or information, but it is also possible to analyse these images as containing and expressing certain emotions. People viewing the images may be affected emotionally, with positive or negative emotions.

Studies in psychology identify the presence of six distinct, basic, and universal emotion categories: happiness, anger, sadness, surprise, disgust, and fear. Various researchers in their studies [17, 20] tackle the problem of emotion recognition by identifying and selecting the following basic emotion categories: joy, sadness, surprise, disgust, anger, fear, and neutral. Facial emotion recognition (FER) is important aspect in various disciplines, as facial expressions play a crucial role in non-verbal communication. The standard FER process includes image pre-processing, face detection, feature extraction, and expression classification, where approaches and their accuracy can significantly vary. Despite its challenges, recent advances in deep learning have improved FER methods, demonstrating that these modern approaches generally outperform conventional techniques in recognition accuracy.

In this paper, we investigate the sadness emotion recognition in images of general nature. Sadness is one of the most common expression and is heavily linked to the other negative emotions. Detecting it in the images of the environment would allow decisions to be taken to modify the environment or adapt it for special needs. Based on psychological insights, our study views sadness as an essential emotion that contributes to the richness of human experience and the depth of personal meaning. With advanced neural networks, this research aims to accurately recognize sadness in images of general nature, thus combining emotional awareness and adaptation to the environment to support human well-being.

A lot of research has been carried out on images of people's faces to identify the emotions of the person in the picture [7, 14, 22]. However, we often deal with images of a general nature, where there is no textual information and faces are not the main focus of the image. In our study, we consider cases where faces are not the main focus of the image to be non-faces. In this case, it is not the emotion of the person that we need to recognise, but the conveyed emotion of the image. Unlike facial emotion recognition problems, the recognition of emotions expressed in the general image remains a challenging task. When analysing images of people's faces, the researchers first manually extracted low-level features including colour, shape and texture. These features are also relevant and used in the analysis of general images. Advancements in deep neural networks, specifically convolutional neural networks, have allowed researchers to capture features and recognize image emotion more reliably. Also, some researchers propose various network fusion strategies [11, 33], through multi-modality. However, relationships between hand-crafted feature fusions are difficult to evaluate objectively. Recognizing the emotion conveyed through non-facial or general images is important for various industries such as marketing, architecture, arts, and design. To achieve this, we leverage deep neural networks.

The aim of our study is to address previous research [37] suggesting that convolutional neural networks (CNNs) tend to favour deep semantic information at the expense of shallower layer features, which are crucial for recognising visual emotions. Given that these shallower layer features play a significant role in conveying emotional content in images, we propose a novel approach that leverages the strengths of EfficientNetV2 CNNs while addressing their limitations.

The main hypothesis of our research is that by extending the capabilities of convolutional neural networks (CNNs) to integrate both deep and shallow layer feature maps, it is possible to improve the detection of sadness emotion in images. Some attempt has been done in [37], where the suggested deep neural network provides a description from the deep semantic representation to shallow visual representation. We extend these ideas. The convolutional neural network (CNN) layer uses a set of filters. The result is a three-dimensional feature map composed of a fixed number of two-dimensional feature sub-maps. Filters produce a feature map that represents different specific detected features, such as edges or textures, contributing to the network's ability to understand diverse characteristics of the visual data. Applications of convolutional neural networks lie at image and video recognition, image classification, image segmentation, natural language processing. Typically, fully connected neural network has many connected weights, which usually may lead to overfitting, but CNNs share receptive

fields through learnable filters. EfficientNet is the CNN-based image classification model family. It was first described in [30]. One of the newest convolutional neural network family EfficientNetV2 [31] was published in 2021. EfficientNet was chosen in our research because we formulate the problem of sadness recognition as an image classification problem. Generally, the term backbone CNN refers to the feature-extracting network that processes input data into a certain feature representation. These feature extraction networks usually perform well as stand-alone networks on simpler tasks, but also researchers utilize them as a feature-extracting part in the more complicated models.

Our contribution is in the integration of different convolutional layers by taking the learned features from the selected layers and applying a pairwise operation to compute the Gram matrices of feature sub-maps, which quantify the correlations between the groups of features in the convolutional layer. The Gram matrix is a mathematical construct that represents the inner product of vectors. By fusing these Gram matrices in the penultimate layer of the network, we transfer additional knowledge from the shallow layer to the deep one. Our findings show that this approach improves the network's ability to recognize sadness in the context of binary classification, resulting in more accurate emotion recognition overall. By computing the Gram matrix for a set of feature sub-maps extracted from an image or feature map of the previous convolutional layer, we can generalize the valuable information that is present. This is particularly important for our goal of recognizing sadness, as this emotion can be conveyed through the textures and visual patterns in an image, besides just facial expressions.

In Section 2, we discuss emotion recognition in images: studies and common architectures. In Section 3, we discuss in detail the motivation, purpose, and strategy of the possible fusion of the Gram matrix module with EfficientNet. This makes a basis of the proposed new architecture, i.e. some extension of EfficientNet. Section 4 describes the experimental setup used for carrying our proposed network training. In Section 5, results of our improved model are described and compared to the baseline network.

2 Related work

Xu et al. [11] demonstrated a visual emotion recognition system using CNN architecture. CNN architecture-based model was trained to recognize objects and then the problem was transferred to sentiment recognition. Chen et. al. [6] used medium-level representations as adjective-noun pairs (ANPs) labelled images. Authors managed by manipulating the strength of the sentiment upon adjectives and nouns to obtain statistical hints for the emotion classification. However, these mentioned works demonstrate how these models solve only binary emotion classification problems. You et al. [35] constructed a large-scale visual emotion dataset named Flick and Instagram set. This dataset was formulated according to the psychology studies and contains 8 labelled emotion categories – amusement, awe, anger, contentment, disgust, excitement, fear, and sadness. This dataset was collected from freely available sources, obtained 90000 weakly labelled emotion images, and using Amazon Mechanical Turk system workers manually labeled emotion images. Using the manual labelling approach final Flickr and Instagram dataset has 23308 visual emotion images.

Other researchers are describing multi-layered network models in order to recognize and classify possible visual emotion [34]. These authors demonstrate the possibility of fusing visual semantic and visual-stream models for predicting emotions. Their proposed visual-semantic model produces possible visual-emotional embedding merging alongside the visual-stream model. Their Visual-semantic model is based on the DeepSentiBank structure [2], which produces conceptual emotion expression, e.g. small beetle, which is expressed as the disgust expression. These expressions are formed as graph embedding in the 2-dimensional space. For the visual stream emotion recognition model authors use ResNet50 [10] model architecture. The final fused model is the multiplication of these 2 different model architectures and in the result, the visual emotion predictions are obtained. A similar approach and study was being done by Zhang et al. [37], where a multi-level representation model with side branches named Gram matrices for shallow features is proposed. The authors in [37] are trying to integrate feature maps from different layers by applying a Gram matrix for further sentiment analysis – i.e. for negative and positive emotion classification.

The training of deep neural networks needs many computing resources. They also tend to have

vanishing or exploding gradient problems. Batch normalization helps here, but, with the increase in depth, the problems above remain. One solution was proposed in *Deep Residual Learning for Image Recognition* by [10] to use Resnet blocks, which connect the output of one layer with the input of an earlier layer. These skip connections are also commonly known as residual connections. The team of *Deep Residual Learning for Image Recognition* won the ImageNet 2015 competition using these deep residual layers, applying skip connections. The authors used ResNet-152 CNN architecture, consisting of 152 layers, This ResNet model surpassed the previously top-performing model on the ImageNet task, named VGGNet16 [28]. Residual connections applicability and usage has been proven widely in various state-of-the-art convolutional neural network architectures such as – Xception, MobileNetV2, DenseNet, EfficientNets [4, 13, 25, 30, 31]. Skip connections are also widely used in other tasks of domain applications – U-Net [23] and DeepLabV3 [3] for various image segmentation tasks.

Previously mentioned studies set the stage for our study, which proposes a novel extension based upon [37] work. Our study addresses EfficientNetV2 improvements by taking advantage of Gram matrices and fusing them onto the network. There are also insufficient recent results regarding specific emotion recognition, which, in our case, is sadness emotion. Furthermore, previous studies employ deep neural networks, which usually need lots of training data. Current investigations on emotion recognition and analysis use ResNet, VGG, DenseNet type networks [14, 34, 37]. There is a lack of visual emotion recognition studies using recent EfficientNet-type networks.

3 Sadness emotion recognition in images of general nature

The recognition of emotions in general images can be considered and evaluated at a higher, more abstract level. In our paper, general images are images that do not contain textual information and where faces are not the main focus of the image. Some authors in their reviews [38], for emotion recognition claim that colors, textures, shapes, and contours as essential (defining) features, determining visual emotion in the given image. This statement (consideration) may lead us to state that emotion recognition in images of a general nature is a different problem as compared to commonly known facial emotion recognition problems. In our case, the expression of emotion detected and recognised may not be of physical origin, as the features describing the emotion can be broad and diverse. For example, the emotion of sadness can be associated with darker colours, textures or a broader subjective emotional pain, which in turn can be associated with many emotional feelings.

Sadness emotion recognition in images of a general nature is being constructed as a binary classification problem – answering whether an image expresses sadness emotion. We chose a convolutional neural network as the means for such classification. Convolutional neural networks still remain one of the main and promising tools for image analysis.

However, for our stated (given) problem we are going to need highly performant and well-structured CNN. Moreover, the convolutional neural networks are known for their need for a large set of images for training [16, 18, 32].

3.1 Modifications in the EfficientNet convolutional neural network

In addressing our stated problem, we will need to handle very large amounts of data. In order to efficiently take advantage of such large datasets, including the aim of reducing computation overhead, it is appropriate to use the EfficientNet convolutional neural network. This network is going to serve as our backbone in striving to maximize the quality of the results, i.e., our goal is to investigate CNN (particularly EfficientNet) improvement areas for better recognition of sadness emotion.

According to the authors of convolutional neural network EfficientNetV2 [31], this network has demonstrated the best results in the ImageNet [24] classification challenge. The used in this challenge ImageNet ILSVRC2012 dataset consists of 1281167 training images, 50000 validation images, and 100000 test images and aims to classify 1000 categories from the mentioned set. EfficientNetV2 family models achieve better results than previous solutions because they incorporate more efficient blocks called MBConv and Fused-MBConv. The authors also conducted a neural network architecture search to find optimal network parameters, using their older EfficientNet [30] B4 version as a base, resulting in a model dubbed EfficientNetV2-S [31].

Model optimization for the EfficientNetV2 was based on these objectives: accuracy, training speed, and number of parameters. EfficientNetV2B0 and EfficientNetV2B2 are scaled-down versions of the original EfficientNetV2, with fewer parameters, fewer convolutional layers, and trained with lower resolution images. The authors also introduced progressive image resolution changing combined with adaptive regularization training methods, which significantly reduced the time required for training not only for their presented model but also for existing older models. The novelty and main idea of progressive training are to divide the training phase into several smaller steps – initially training the network using lower-resolution images with weaker regularization and, in later stages, increasing the image resolution and incorporating stronger regularization using mixing [36] (blending images into one and outputting a probabilistic category), random augmentation [5], and stochastic dropout [29].

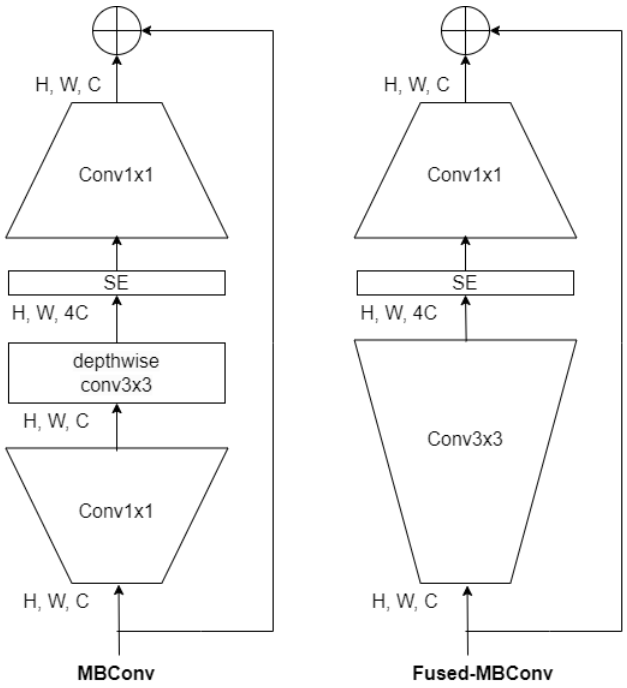


Figure 1: Structure of MBConv and Fused-MBConv blocks. Source: [31]

Figure 1 shows the structural blocks (modules) of EfficientNetV2 architecture. By using combinations from these blocks, the authors determined the entire network structure. Here H and W are the height and width of the input, and C is the number of channels. The MBConv block, also known as the inverted residual block, is understood as a variation of the residual block aimed at achieving higher efficiency. The inverted residual block was first introduced in the MobileNetV2 convolutional neural network architecture [25]. Initially, a 1x1 convolution expands the number of layer channels, followed by a special 3x3 depthwise convolution that reduces the number of parameters, and finally, a 1x1 convolution is applied to normalize the dimensions of the output and input. This normalization is necessary to combine them using a residual connection (skip connection). The authors of EfficientNetV2 also enhanced this block with a so-called squeeze and excitation (SE) layer, which was first introduced by the authors of [12]. This layer, essentially a module, consists of a global average pooling, fully connected, ReLU activation, subsequent fully connected, sigmoid activation, and multiplication operations. Such a block helps achieve better results in benchmark solutions with a minimal increase in computational cost. The essential difference between MBConv and Fused-MBConv is that Fused-MBConv replaces the first two layers with a conventional 3x3 convolution.

Table 1 illustrates the optimized structure of EfficientNetV2S detailing its components and blocks. The structure was optimized using reinforcement learning on the basis of the ImageNet dataset [24]. Stride refers to the convolution operation's step size. Channels No. indicates the number of output channels from a particular block or operation. Layers No. specifies the count of particular block repetitions within a certain stage. For example, the number of layers in the fourth stage, 6, indicates the number of MBConv block repetitions. MBConv[n] denotes the module MBConv with an expansion

Stage	Operation	Stride	Channels No.	Layers No.
0	Conv3x3	2	24	1
1	Fused-MBConv1, k3x3	1	24	2
2	Fused-MBConv4, k3x3	2	48	4
3	Fused-MBConv4, k3x3	2	64	4
4	MBConv4, k3x3, SE0.25	2	128	6
5	MBConv6, k3x3, SE0.25	1	160	9
6	MBConv6, k3x3, SE0.25	2	256	15
7	Conv lxl & Pooling & FC	-	1280	1

Table 1: Structure and parameters of EfficientNetV2S. MBConv and Fused-MBConv blocks are described in Figure 1. Source: [31]

factor of n – the initial 1x1 convolution receives C channels and expands the output to n*C channels. SE0.25 refers to the reduction ratio of the squeeze and excitation block used to model channel-specific relations.

Stage	Operation	Stride	Channels No.	Layers No.
0	Conv3x3	2	32	1
1	Fused-MBConv1, k3x3	1	16	2
2	MBConv4, k3x3	2	32	3
3	MBConv4, k3x3	1	56	3
4	MBConv4, k3x3, SE0.25	2	104	4
5	MBConv6, k3x3, SE0.25	1	120	6
6	MBConv6, k3x3, SE0.25	2	208	10
7	Global Average Pooling & FC	-	1408	1

Table 2: Structure and parameters of EfficientNetV2B2 model variant used in this research. MBConv and Fused-MBConv blocks are described in Figure 1

Table 2 presents the structure of the EfficientNetV2B2 model (stages 0-6), which we primarily utilized as our backbone. For simplicity of notation in this paper, we will use B2 in parallel with EfficientNetV2B2. EfficientNetV2B2 is suggested in [31] as a scaled-down alternative to EfficientNetV2S. In contrast, the EfficientNetV2S model, shown in Table 1, requires twice as much training time and is prone to overfitting, that do not grant any significant advantages over the smaller and more streamlined version of the model. At Stage 7 we have placed a global average pooling layer resulting in a vector of 1408 fully connected units.

Zhang et al. [37] have introduced a multi-level representation model consisting of several Gram matrices of shallow features and a backbone CNN model. They are claiming that deep neural networks namely CNNs mainly rely on deep semantic information making the learned shallow features less important. However, according to Zhang et al. [37], these shallow features are essential for detecting emotions, too.

Therefore, their introduced shallow visual representation model adopts a Gram matrix to extract the correlation between feature sub-maps. They state that the Gram matrix succeeds in capturing low-level visual features retaining the color and texture details of the image, in addition to eliminating the interference of the image content. They are also suggesting that their proposal allows better representation of the low-level features, which supposedly are significant for detecting visual emotion. Their backbone is the ResNet-50 network, which is treated as a high-level semantic feature representation extractor.

The network part that extends the ResNet-50 consists of visual feature representation extraction by computing Gram matrices, which are transformed into one-dimensional vectors and fused with a 1x1 convolutional result. They have chosen to connect and aggregate the set of visual representations by fully connected layers. Through experiments, the authors have determined that the best result is obtained when the chosen number of fully connected units for visual feature representation aggregation

matches the dimension of the semantic feature annotation of ResNet which would be 2048.

Thus, inspired by the research of Zhang et al. [37], we propose improvements to the Gram matrix module, which we then integrate with the backbone CNN model. Our main contributions and improvements upon the previous work of Zhang et al. [37] are as follows:

- We propose enhanced Gram matrix modules that include additional activation functions, thus improving the robustness of feature extraction.
- We extend the capabilities of the convolutional neural network, especially EfficientNetV2, to leverage both deep and shallow layer feature maps.

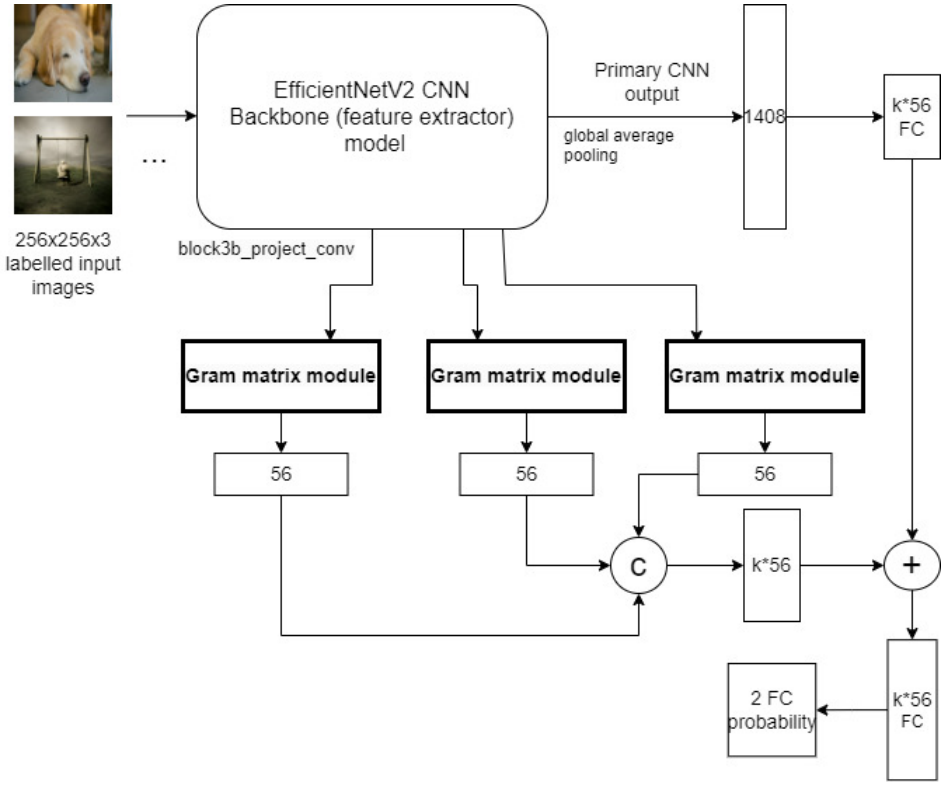


Figure 2: General schema of the proposed network, $k = 3$

Figure 2 shows the general scheme of the model we propose. EfficientNetV2 CNN network was chosen as the backbone due to the smaller number of parameters and competitive results on the ImageNet [24] challenge as compared to the ResNet, VGGNet 16, Xception [4, 10, 28]. The final output of the key backbone EfficientNetV2 CNN model consists of 1408 neurons – a layer of 1408 fully connected units. The subsequent layer compresses a vector of 1408 elements to a vector of $k * 56$ elements by applying compression, where k corresponds to the number of Gram matrix modules, which is $k = 3$ as shown in Figure 2. EfficientNetV2B2 network has 91 convolutional layers, forming 6 convolutional stages (see Table 2). In our extension of EfficientNetV2B2, we choose a feature map that is an output of the selected convolutional layer of the backbone model, in our case that is *block3b project conv*. The name of this layer in the overall backbone network corresponds to the Stage 3 and second repeating block, which performs convolutional projection – reduces dimensions of feature sub-maps (see Table 2). The output of the mentioned layer is passed to several Gram matrix modules. Each Gram matrix module outputs one-dimensional sized output. These outputs are concatenated alongside each Gram matrix module and then finally fused by addition operation with $k*56$ fully connected units outputting layer from the backbone CNN output branch. The final output of the network corresponds to the predicted class *softmax* summed probability of 2 units ranging between 0 and 1, where each unit represents prediction to what class given emotion image belongs to – image expressing sadness emotion, or the image does not contain sadness expression. Our proposed structure allows adding

more Gram modules, granting flexibility in terms of modeling and determining the effectiveness of modules.

We have chosen one feature map extracting layer named *block3b project conv*. This feature map is passed in parallel as input to three Gram matrix modules. Each Gram matrix module produces vectors of equal size of 56 elements and these vectors are fused by applying a concatenation fusion strategy granting one vector of 168 elements. Backbone CNN and auxiliary Gram matrix modules were fused by applying summation operation. Other fusion options were considered such as addition, and concatenation.

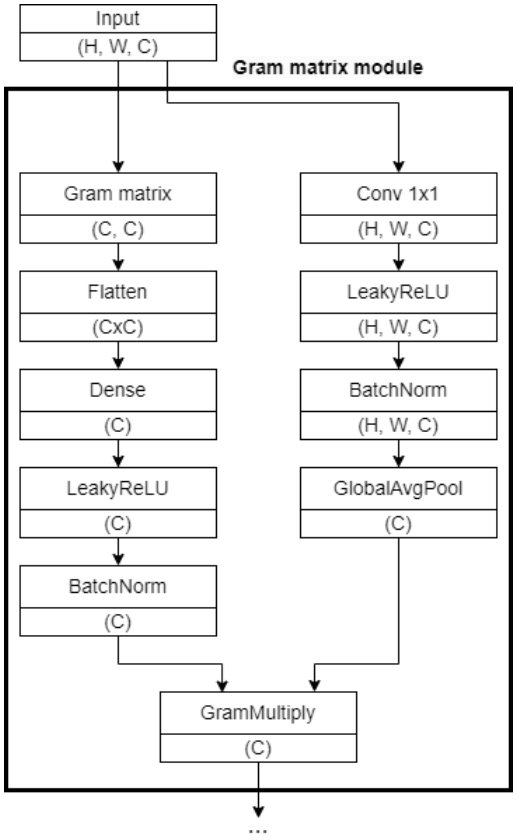


Figure 3: Proposed Gram matrix module schema with detailed flow

Figure 3 shows the proposed Gram matrix module structure. Each module obtains input, whose shape is $\mathbb{R}^{H\times W\times C}$ corresponding to the extracted layer's feature map, consisting of C feature sub-maps. Feature sub-maps are defined by height and width $H \times W$ spatial dimension. It should be noted that the output of the Gram matrix is in quadratic form and is expressed as $C \times C$ squared matrix. The gram matrix is further flattened into 1 1-dimensional vector consisting of $C \times C$ units, which is further compressed by a dense layer resulting in C units, which then are applied by activation function and batch normalized. Accordingly, the other side of our proposed module consists of a 1x1 convolution operation, corresponding activation function – for each feature sub-map among C ones, a singular average value is computed from all $H \times W$ values of the sub-map. As a result, we obtain a vector of C length that contains average values of all C sub-maps. The final result of the Gram matrix module is fused by multiplication from each side of the branch as shown in Figure 3. We also considered the concatenation, addition, and average fusion strategies, but those options yielded no gains.

Gram matrix $G \in \mathbb{R}^{c\times c}$ can be written as:

$$G = FF^T \text{ and } \mathrm{F} \in \mathbb{R}^{C\times HW}, F^T \in \mathbb{R}^{HW\times C}, \tag{1}$$

where in equation (1) C refers to the convolutional layer channels (filters), H and W to the height and width accordingly. $F \in \mathbb{R}^{C\times HW}$ refers to the flattening into C rows and HW column matrix, and $F^T \in \mathbb{R}^{HW\times C}$ is accordingly consists of HW rows and C column matrix.

The question is why repeating the Gram matrix module three times yields different output results even though the input to the module is the same for all instances. This can be explained by the fact that the dense layer, when preceded by the flattened layer, possesses trainable weights. Similarly, the 1×1 convolution also involves trainable filters.

We might also consider the rationality for selecting a specific layer in the backbone network for the Gram matrix module. Following the approach of Zhang et al. [37], we advocate for the extraction of features from shallow layers for the computation of the Gram matrix and further application in the training process. Utilizing shallow layer features brings several advantages: whereas the backbone model outputs a high number of fully connected units — 1408, the aggregated Gram matrices modules yield only 168 units. The output units from the backbone neural network are compressed by applying a subsequent standard dense layer, and both outputs are fused by a summation operation. This method helps to reduce the overfitting of the model.

4 Experimental setup

4.1 Data

WEBEmo [20] may serve as a set for our stated problem. WEBEmo dataset contains about 268000 images. It is a large-scale weakly-labeled image emotion dataset for possible training of convolutional neural networks. We have downloaded a part of the WEBEmo dataset and managed to retrieve 220854 images. This dataset contains images of a general nature, however, part of the images have some text. Textual data may carry some emotion and influences the emotion of the picture. In our case, we should discard these mentioned images. Finally, after the additional undersampling, we have obtained 61074 filtered images dataset, where about 46 % of images represent sadness emotion. Our constructed dataset has been divided into 80 % training, 10 % validation and 10 % testing subset splits. WEBEmo training subset contains 26445 images expressing no sadness emotion, and 22413 images conveying sadness emotion. Similarly, the WEBEmo validation subset consists of 3284 images expressing no sadness emotion, and 2823 images expressing sadness emotion. Finally, the testing subset split is divided into the same ratio as validation subset, consisting of the same number of images in each class.

4.2 Methodology

Like the authors of [37], we took such a pre-trained network EfficientNetV2B2 that recognizes 1000 objects as an initial state for further its training for the sadness emotion classification.

We experienced a strong over-fitting problem and inability to generalize over unseen images using EfficientNetV2B2 (see Stages 0-6 in Table 2), so a fine-tuning improvement approach was done. We suggested including additional layers to reduce over-fitting (see Stage 7 in Table 2). Dropout and fully connected layer combinations were applied after the primary pre-trained model outputs. After the subsequent Stage 7, we applied dropout with rate $p = 0.5$, used leaky Rectified Linear Unit ReLU activation [19] with $a = 0.2$ value, and compressed result to a vector sized of 168 fully connected units as shown in Figure 2. This allows to reduce over-fitting and ensures that output shapes from the backbone branch match with the Gram matrix modules side.

Training process of both networks (EfficientNetV2B2 variant given in Table 2 (Stages 0-7) and its generalization using Gram matrix modules given in Figure 2) was carried out as follows: 61074 total filtered images dataset (without images with text) was split into 80 % subset for training, 10 % subset for validation and remaining 10 % for testing. The validation set in general helps in tuning hyperparameters of the model and detecting whether a trained network is overfitting, thus allowing to estimate of the ability of the model to generalize on the unseen data. In our case, the most relevant and important hyperparameters for fine-tuning were the following ones: learning rate, batch size, which directly influences the number of training steps, and selected learning rate feasibility [26]. Adagrad optimizer [8] for training was used with a 0.002 learning rate, which was scaled (multiplied) with the number of used GPUs in training – in our case, it was 2. Also, we have chosen a batch size equal to 64 per GPU. The general loss function is expressed as below:

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log(p_{y_i}),$$

where L is the average loss for the entire subset (training or validation), i.e. it is a sparse categorical cross-entropy. In our case, number of classes is $K = 2$. N corresponds to the number of training or validation samples depending on the phase of the learning process. p_{y_i} represents the predicted probability of the true class for the i-th sample and y_i is the index of the true class for the i-th sample. The negative logarithmic function is used because probabilities are calculated between 0 and 1.

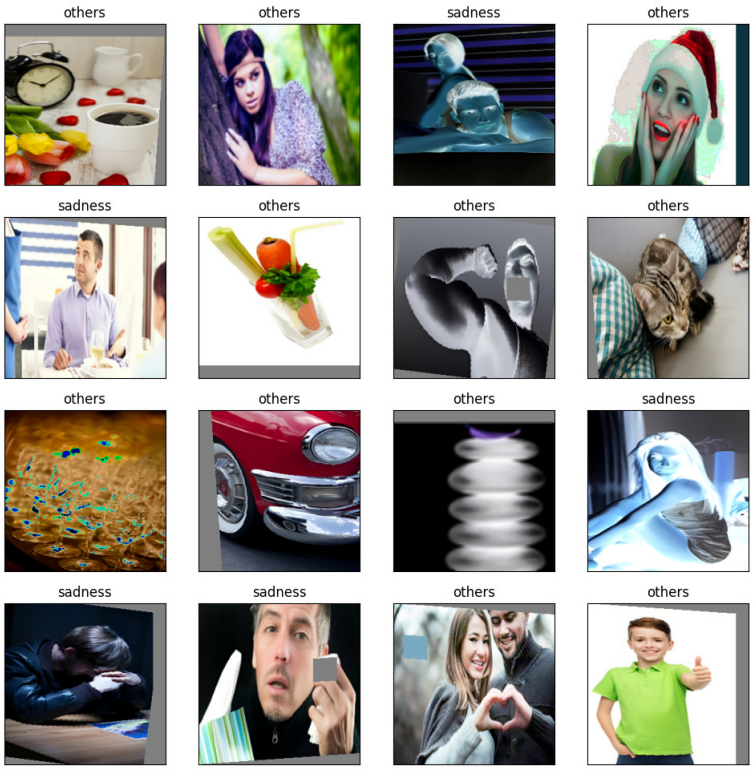


Figure 4: Filtered out text from WEBEmo dataset example

Input images were provided as 256x256 colored images. We have used augmentations in order to reduce the network over-fitting and to improve the generalization capabilities of the model. Such an augmentation does not increase the amount of data. In Figure 4, an example of augmented images with *RandAugment* [5] is shown. This augmentation procedure has parameters, where **N** refers to the number of transformations to apply and **M** refers to the augmentation magnitude and strength. We have chosen $\mathbf{N} = 3$ and $\mathbf{M} = 7$. Our reasoning for these given values is that we have conducted several tests on the baseline model and we have determined those values as the most appropriate experimentally. Note that this augmentation process does not create additional samples in the dataset. Each image is augmented during the training phase, with different random distortions applied in every training epoch. Interestingly, from the authors of *RandAugment* [5], we have also found out that relatively low distortion magnitude gave us the best performance. Therefore, our tests confirm their statement. In this example, we see images of a really general nature despite the fact that there are human faces in the images. However, particular details of the pictures can convey a certain emotion.

In order to evaluate model performance and efficiency in predicting sadness emotion, we take advantage of overall accuracy, $F1$, precision, and recall metrics. These metrics have been defined as

follows [1]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$
$$Precision = \frac{TP}{TP + FP},$$
$$Recall = \frac{TP}{TP + FN},$$
$$F1 = \frac{2 * Precision * Recall}{Precision + Recall},$$

where TP is a number of true positive classification instances – in our case it is correct sadness emotion class prediction, TN is similarly representing to the true negative classification instances – correct no sadness emotion classification labeled as Others class. FP is a model prediction, where it incorrectly predicted the positive class – i.e. incorrect prediction of the sadness emotion in a given image, while the true class of image was named as others. Accordingly, FN is a model prediction, where it incorrectly predicted the negative class – incorrect prediction of no sadness emotion class, which means that the true class of the given image was named as expressing sadness emotion. For our problem, *Precision* measures the percentage of images predicted as expressing sadness emotion, that were correctly classified. Conversely, *Recall* measures the percentage of actual sadness emotion images, which are correctly classified. $F1$ score is the most appropriate metric, because of the imbalance of our classes in the emotion image sets. In addition, $F1$ score is a metric that balances *Recall* and *Precision* metrics, since emotion image datasets are inherently imbalanced. In this case, it is worth exploring the prediction of sadness emotion using other metrics. Often, *Accuracy* is expressed in percent for easier its interpretation.

It follows from the above that *Accuracy* and *Precision* are the most relevant metrics for evaluating proposed network performance. *Accuracy* metric is suitable, because the classes in the datasets are fairly balanced and it is sufficient to estimate how well model performs across all the classes. *Precision* focuses on the accuracy of positive predictions, meaning a ratio of correct positive predictions.

5 Results

The aim of the experimental study is to compare the proposed new model with the backbone using the above metrics and a dataset of images where emotions are annotated. It is also necessary to determine the appropriate number of Gram modules to be connected in parallel.

Network	*Accuracy* (%)	*SD*
Baseline B2	81.368	0.174
Zhang et al. [37]	81.313	0.186
Improved B2, 2 Gram modules	81.772	0.118
Improved B2, 3 Gram modules	81.806	0.177
Improved B2, 4 Gram modules	81.826	0.086
Improved B2, 5 Gram modules	81.816	0.235

Table 3: *Accuracy* of versions of B2 averaged over 5 runs and compared to the baseline. B2 corresponds to the backbone EfficientNetV2B2

In Table 3, the averaged accuracy with standard deviations SD are shown. Our proposed network outperforms the baseline network with around 0.4 % higher *Accuracy*. Using two Gram matrix feature extraction modules ($k = 2$ case) gives us a slightly worse result as compared with the case of three modules ($k = 3$). The main reason for increased *Accuracy* after the inclusion of more Gram modules might be that more Gram modules provide greater generalization capability, as computations (learning) are performed in parallel for different initial values of respective parameters of Gram modules, and then the results are combined. Using four parallel Gram matrix modules produces encouraging results. Trained network with $k = 5$ Gram matrix modules obtains slightly lower *Accuracy* and higher standard deviation suggesting unstable and unreliable performance of this model choice. We have

also evaluated the performance of Zhang et al. [37] model. The total number of training epochs and training parameters were set as used in [37]. Here number of training epochs was equal to 60. Note, that baseline B2 and our improvements of B2 use 25 epochs, only. We see that EfficientNetV2B2 outperforms Zhang et al. [37]. As mentioned above, applying *RandAugment* augmentation procedure [5] allowed us to improve our trained network overall accuracy at around 3 % – 4 % on the test sets. Note, that the augmentation improves the accuracy of the backbone network similarly. Let us note, that the standard deviation is smallest when $k = 4$. Here, we can assume that a greater number of Gram modules produces some stability of the results and of the network in general. However, case $k = 3$ gives almost the same standard deviation as baseline B2, but *Accuracy* is better.

Network	Others *Precision*	Others *Recall*	Others *F*1-*score*	Sadness *Precision*	Sadness *Recall*	Sadness *F*1-*score*
Baseline B2	0.8374	0.8088	0.8229	0.7862	0.8174	0.8015
Improved B2, 2 Gram modules	0.8257	0.8368	0.8312	0.8071	0.7946	0.8008
Improved B2, 3 Gram modules	0.8253	0.8392	0.8322	0.8092	0.7932	0.8011
Improved B2, 4 Gram modules	0.8252	0.8427	0.8339	0.8125	0.7924	0.8022
Improved B2, 5 Gram modules	0.8242	0.8414	0,8327	0.8109	0.7911	0.8009

Table 4: *Precision*, *Recall* and *F*1-*score* results of versions of B2 averaged over 5 runs and compared to the baseline. B2 corresponds to the backbone EfficientNetV2B2

In Table 4, the averaged over 5 runs classification results are shown. Our proposed network consisting of $k = 3$ Gram matrix modules shows a stronger performance on the other class *Recall* and sadness class *Precision* metrics. However, the baseline network, when compared against proposed networks, has better performance when evaluating others class *Precision* and sadness emotion *Recall* metrics. Proposed networks also have better *F*1-*score* in the other class, and no significant difference for the sadness class. Trained network with $k = 5$ Gram matrix yields no gains in any metric when compared to the network consisting of $k = 4$ Gram matrix modules. This means that there is an optimal number of modules, and achieves the best performance for $k = 3$ or $k = 4$. Networks, which were trained with Gram matrices, resulted in a lower precision, but higher *Recall* scores for others class, when compared to baseline network. However, proposed networks also produced slightly higher *F*1-*score* against baseline network. For sadness class at Table 4, we can notice such trends: proposed networks provide higher *Precision*, lower *Recall* and similar *F*1-*score* when compared with the trained baseline network.

5.1 Applying the trained networks on other datasets

In this section, we present results of using our suggested trained networks by WEBEmo data on other emotion image datasets. In addition, the trained baseline network performance is presented for comparison, too. Our proposed network consists of $k = 4$ Gram matrix modules. The networks were trained using WEBEmo dataset as described in Section 4. Then, we use UnbiasedEmo [20] and Emotion-6 [20] subsets for analysis using the trained networks. The goal of experiment is to estimate the generalization capability of trained networks on other unseen datasets.

Class	*Precision*	*Recall*	*F*1-*score*	Support
Others	0.8607	0.8025	0.8307	467
Sadness	0.7603	0.8283	0.7928	353

Table 5: Testing report of the trained network of 5 averaged runs with $k = 4$ Gram matrix modules using UnbiasedEmo subset

When looking at Tables 5 and 6, where a trained model with four Gram modules was used, it can be noted that the proposed model on the particular cases has the highest *F*1-*score* for sadness emotion class. On the UnbiasedEmo testing set, the proposed network performs comparatively well across both classes. Support in the given tables is the number of emotional images of the particular

Class	*Precision*	*Recall*	$F1$-*score*	Support
Others	0.8365	0.7198	0.7737	888
Sadness	0.5907	0.7417	0.6576	484

Table 6: Classification report of Emotion6 testing subset using trained network of 5 averaged runs with $k = 4$ Gram matrix modules

class. The network demonstrates reliable results in terms of sadness class precision and $F1$ values. However, on the Emotion6 testing set, the proposed network performs worse in terms of discerning sad image emotions. The reason might be that there is a slight class imbalance, where the majority group is highly favored. Interestingly, the trained networks on the UnbiasedEmo and Emotion-6 as shown in Tables 5 and 6, have even better precision performance for the others class than in WEBEmo dataset testing subset as described in Table 4.

In Figure 5, the experimental results with test data from subsets of the UnbiasedEmo and Emotion-6 datasets are presented, with an evaluation of the area under the receiver operating characteristic curve (AUC-ROC).

We use the Area Under the Receiver Operating Characteristic curve (AUC-ROC) as an additional measure to evaluate the performance of our proposed model. This measure typically means the area under the curve. The area varies between 0.5, indicating no discriminative ability, and 1, meaning perfect classifier. Since it is possible to calculate predictions at different classification thresholds – not just at the highest probability, we can evaluate the proposed network effectiveness using the ROC curve. The threshold is a value that serves as a decision boundary (of classification probability) choosing to what class a given image emotion belongs. When the probability output of the model is above the chosen threshold, the image emotion is classified as expressing sadness emotion; if it is below, the image instance is classified as not expressing sadness emotion. The true positive rate (TPR) is the proportion of actual positive cases that are correctly identified, while the false positive rate (FPR) is the proportion of actual negative cases that are incorrectly identified as positive. Furthermore, AUC-ROC is also valuable for the following reasons: summarizes binary classifier under a single value, handles class imbalances, and sadness emotion in images is usually a minority category. AUC-ROC metric is commonly interpreted and used in clinical research due to its inherent applications of false-positives, however, this metric is only a measure of model predictive capability [15].

In Figure 6, baseline network testing results on UnbiasedEmo and Emotion-6 subsets are presented. It can be said that the baseline has slightly higher sadness emotion discriminative capability on the UnbiasedEmo testing subset as compared to the network with three Gram matrix modules. The baseline network performs worse on the Emotion-6 testing subset as compared with our baseline network shown on the right side in Figure 5.

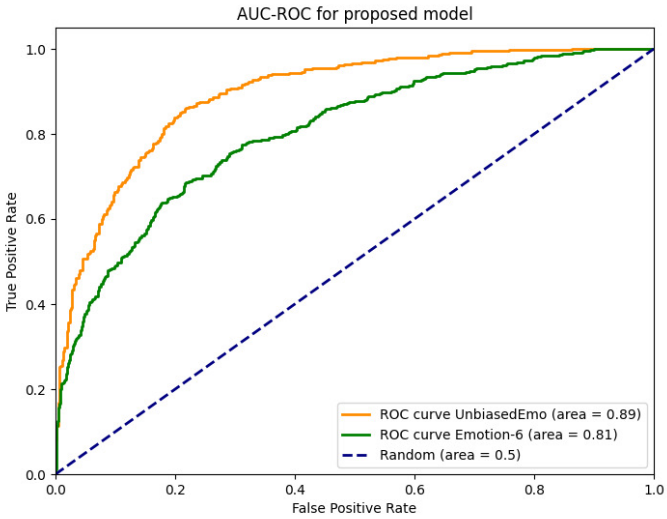


Figure 5: Proposed network AUC-ROC curves for identifying sadness image emotion on unseen test image sets

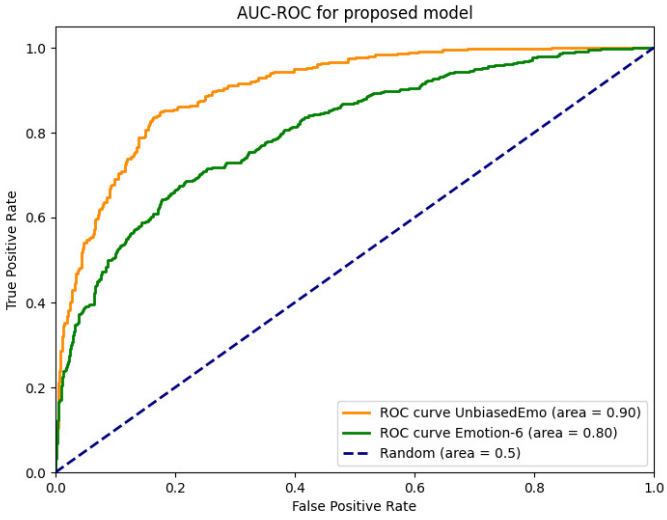


Figure 6: Baseline network AUC-ROC curves for identifying sadness image emotion on unseen test image sets

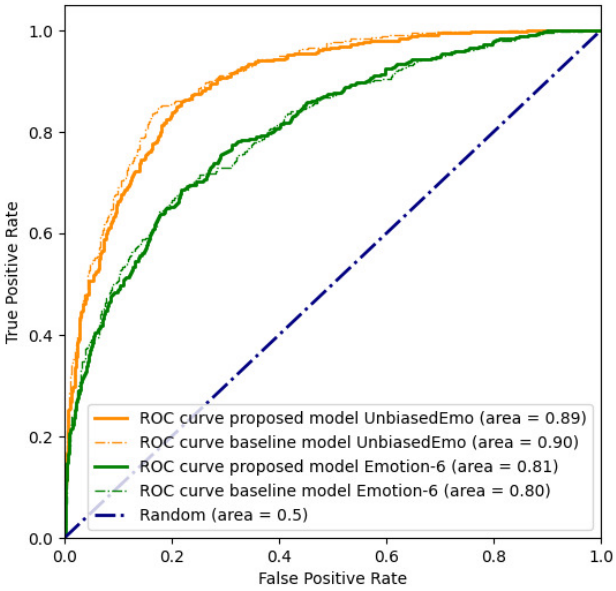


Figure 7: AUC-ROC curves of baseline and proposed networks for identifying sadness image emotion on unseen test image sets

In Figure 7, aggregated baseline and proposed network testing results on UnbiasedEmo and Emotion-6 subsets are presented. AUC-ROc value differences compared our proposed model against baseline model are comparatively small. Proposed network has slightly higher area under the curve on the Emotion-6 image set. However, the proposed network has slightly worse area curve compared to the baseline model on the UnbiasedEmo image set.

5.2 Practical case study on the artwork images

Figure 8: Sadness emotion recognition on the artwork images

The aim of this section is to illustrate the recognition of emotions in the images of general nature. We do not intend to compare our solution with a baseline, but we want to illustrate the possibilities of the analysis.

In Figure 8, several image artworks conveying emotion are shown. The first three artworks are by Vincent Van Gogh. The remaining two are by 8-year-old child. In our given example, we are classifying images using a network trained with $k = 4$ Gram matrix modules. Above the pictures, we present the predicted class and probability of dependence of the picture to the predicted class. It can be noted that the network recognizes some of Van Gogh's artworks as expressing the emotion of sadness. Interestingly, the trained network confidently predicts no sadness in the first image, which is a painting of a flower bouquet. We might wonder that the trained network recognizes emotion on these features: colors, textures, physical expressions. From the given example, it can be observed

that there are common feature details, such as darker colors, texture, and physical shapes, among the images that are recognized as expressing the sadness emotion.

6 Conclusion

The proposed model, which fuses Gram matrix modules, offers the competitive sadness emotion recognition ability. The research shows an improvement at recognizing sadness emotion in images of general nature. The proposed extensions to the EfficientNet network show a new way to increase the quality of recognition by connecting multiple Gram-matrix modules.

In this paper, we have demonstrated improvements to the EfficientNetV2B2 convolutional neural network backbone to address our problem. We have successfully demonstrated the potential of the Gram matrix module as a means to compute feature sub-map correlations and improve visual emotion recognition. Our proposed network architecture allows us to improve the feature extraction capabilities, as shown in Figure 2. In the results section, we have provided results showing improvement of the performance across the metrics. We used EfficientNetV2B2 fine-tuned network as baseline model for evaluating and comparing the efficiency of our proposed network. It can be noted, that accuracy of our proposed models is superior to the baseline model. Our model variants do not significantly improve accuracy compared to the baseline. However, computational costs with the new modules are almost the same as in the baseline case. In addition, the computational cost of training the network is not essential for good classification. This is the main argument to use our model.

Possible new approaches to feature map extraction using the Gram matrix module need to be further explored. One of the main reasons for this may be that sophisticated design of choosing layers of the baseline network can be used as inputs to the Gram matrix modules. This means that there is an unexplored area of research that could lead to an even better emotion image recognition network. Also, the research could be extended to other image datasets.

Acknowledgment

The authors are grateful to the reviewers for their invaluable comments, which have improved the quality of the paper, as well as for their ideas for possible further research.

Funding

The APC was funded by Vilnius University, Lithuania.

Author contributions

The authors contributed equally to this work.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*, Springer Series in *Information Science and Statistics*. Springer, 2006. ISBN: 978-0387310732.

[2] Chen, T.; Borth, D.; Darrell, T.; and Chang, S.F. (2014). Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks, *arXiv Preprint* arXiv:1410.8586, 2014.

[3] Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation, *arXiv Preprint* arXiv:1706.05587, 2017.

[4] Chollet, F. (2016). Xception: Deep learning with depthwise separable convolutions, arXiv Preprint arXiv:1610.02357, 2016. Presented in: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1800–1807, 2017.

[5] Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. (2020). Randaugment: Practical automated data augmentation with a reduced search space, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 702–703, 2020.

[6] Dellandrea, E.; Liu, N.; and Chen, L. (2010). Classification of affective semantics in images based on discrete and dimensional models of emotions, In: *2010 International Workshop on Content Based Multimedia Indexing (CBMI)*, IEEE, https://doi.org/10.1109/CBMI.2010.5529906, 2010.

[7] Deshmukh, R.S.; Jagtap, V.; Paygude, S. (2017). Facial emotion recognition system through machine learning approach, In: *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 272–277, 2017. https://doi.org/10.1109/ICCONS.2017.8250725.

[8] Duchi, J.D.; Hazan, E.; Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, *Journal of Machine Learning Research*, 12, 2121–2159, 2011.

[9] Goutte, C.; Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation, In: *Advances in Information Retrieval*, Springer Berlin Heidelberg, 345–359, 2005.

[10] He, K.; Zhang, X.; Ren, S.; Sun, J. (2016). Deep residual learning for image recognition, In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778, 2016.

[11] He, X.; Zhang, W. (2018). Emotion recognition by assisted learning with convolutional neural networks, *Neurocomputing*, 291, 187–194, 2018. https://doi.org/10.1016/j.neucom.2018.02.073.

[12] Hu, J.; Shen, L.; Sun, G. (2018). Squeeze-and-excitation networks, In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141, 2018.

[13] Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. (2017). Densely connected convolutional networks, In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708, 2017.

[14] Iqbal, J. L. M.; Kumar, M. S.; Mishra, G.; Asha, G. R.; Saritha, A. N.; Karthik, A.; Kotaiah, N. B. (2023). Facial emotion recognition using geometrical features based deep learning techniques, *International Journal of Computers, Communications and Control*, 18(4). https://doi.org/10.15837/ijccc.2023.4.4644.

[15] Janssens, A.C.J.W.; Martens, F.K. (2020). Reflection on modern methods: Revisiting the area under the ROC Curve, *International Journal of Epidemiology*, 49(4), 1397–1403, 2020. https://doi.org/10.1093/ije/dyz274.

[16] Johnson, J.; Karpathy, A.; Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning, In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 4565–4574, 2016.

[17] Karbauskaite, R.; Sakalauskas, L.; Dzemyda, G. (2020). Kriging predictor for facial emotion recognition using numerical proximities of human emotions, *Informatica*, 31(2), 249–275, 2020. https://doi.org/10.15388/20-INFOR419.

[18] Levine, S.; Pastor, P.; Krizhevsky, A.; Ibarz, J.; Quillen, D. (2018). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection, *The International Journal of Robotics Research*, 37(4–5), 421–436, 2018.

[19] Maas, A.L.; Hannun, A.Y.; Ng, A.Y.; et al. (2013). Rectifier nonlinearities improve neural network acoustic models, In: *Proc. International Conference on Machine Learning (ICML)*, 2013.

[20] Panda, R.; Zhang, J.; Li, H.; Lee, J.-Y.; Lu, X.; Roy-Chowdhury, A.K. (2018). Contemplating visual emotions: Understanding and overcoming dataset bias, *arXiv Preprint* `arXiv:1807.03797`, 2018.

[21] Polycarpou, M.M. (2008); Editorial: A new era for the IEEE Transactions on Neural Networks, *IEEE Transactions on Neural Networks*, 19(1), 1–2, 2008. `https://doi.org/10.1109/TNN.2007.915293`.

[22] Revina, I.M.; Emmanuel, W.R.S. (2018). A survey on human face expression recognition techniques, *Journal of King Saud University – Computer and Information Sciences*, (2018). `https://doi.org/10.1016/j.jksuci.2018.09.002`.

[23] Ronneberger, O.; Fischer, P.; Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation, In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*, Munich, Germany, October 5–9, 2015, Proceedings, Part III, Springer, 234–241, 2015.

[24] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A.C.; Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge, *International Journal of Computer Vision*, 115, 211–252, 2015. `https://doi.org/10.1007/s11263-015-0816-y`.

[25] Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks, In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520, 2018.

[26] Smith, S. L.; Kindermans, P.-J.; Ying, C.; Le, Q. V. (2017). Don't decay the learning rate, increase the batch size, *arXiv preprint arXiv:1711.00489*,

[27] Shao, J.; Yongsheng, Q. (2019). Three convolutional neural network models for facial expression recognition in the wild, *Neurocomputing*, 355, (2019). `https://doi.org/10.1016/j.neucom.2019.05.005`.

[28] Simonyan, K.; Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition, *arXiv Preprint* `arXiv:1409.1556`, 2014.

[29] Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting, *Journal of Machine Learning Research*, 15(1), 1929–1958, 2014.

[30] Tan, M.; Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks, In: *International Conference on Machine Learning*, 6105–6114, PMLR, 2019.

[31] Tan, M.; Le, Q.V. (2021). EfficientNetV2: Smaller Models and Faster Training, *arXiv Preprint* `arXiv:2104.00298`, 2021.

[32] Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification, In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1701–1708, 2014.

[33] Yang, J.; She, D.; Sun, M. (2017). Joint image emotion classification and distribution learning via deep convolutional neural network, In: *IJCAI'17: Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 3266–3272, 2017.

[34] Yang, H.; Fan, Y.; Lv, G.; Liu, S.; Guo, Z. (2022). Exploiting Emotional Concepts for Image Emotion Recognition, *Visual Computer*, (2022). `https://doi.org/10.1007/s00371-022-02472-8`.

[35] You, Q.; Luo, J.; Jin, H.; Yang, J. (2016). Building a large scale dataset for image emotion recognition: The fine print and the benchmark, In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), 2016.

[36] Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization, *arXiv Preprint* `arXiv:1710.09412`, 2017.

[37] Zhang, H.; Liu, Y.; Xu, D.; He, K.; Peng, G.; Yue, Y.; Liu, R. (2022). Learning multi-level representations for image emotion recognition in the deep convolutional network, *SPIE-Intl Soc Optical Eng*, 91, (2022). `https://doi.org/10.1117/12.2623414`.

[38] Zhao, G.; Yang, H.; Tu, B.; Zhang, L. (2021). A survey on image emotion recognition, *Journal of Information Processing Systems*, 17(6), (2021).

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Motiejauskas, M.; Dzemyda, G. (2024). EfficientNet Convolutional Neural Network with Gram Matrices Modules for Predicting Sadness Emotion, *International Journal of Computers Communications & Control*, 19(5), 6697, 2024.
https://doi.org/10.15837/ijccc.2024.5.6697