Visual Sentiment Analysis Multiplying Deep learning and Vision Transformers

Tetsuya Asakawa[1], Riku Tsuneda[1], Masaki Aono[1]
[1]Toyohashi University of Technology, Japan
asakawa.tetsuya.um@tut.jp, tsuneda.riku.am@tut.jp, masaki.aono.ss@tut.jp

ABSTRACT

Visual sentiment analysis investigates sentiment estimation from images and has been an interesting and challenging research problem. Most studies have focused on estimating a few specific sentiments and their intensities using several complex CNN models. In this paper, we propose multiplying CNN and Vision Transformers method in MediaEval 2021 Visual Sentiment Analysis: A Natural Disaster Use-case. Specifically, we first introduce our proposed model used in subtask1. Then, we also introduce a median-based multi-label prediction algorithm used in Subtask 2 and 3, in which we assume that each emotion has a probability distribution. In other words, after training of our proposed model, we predict the existence of an evoked emotion for a given unknown image if the intensity of the emotion is larger than the median of the corresponding emotion. Experimental results demonstrate that our model outperforms several models in terms of subset Weighted F1-Score.

1 INTRODUCTION

With the spread of SNS and the Internet, a vast number of images are widely available. As a result, there is an urgent requirement for image indexing and retrieval techniques. When viewing an image, we can feel several emotions simultaneously. Different visual images have different emotional triggers. For instance, an image with a snake or a spider may most likely trigger a bad feeling like "disgust" or "fear," whereas an image with a flower may most likely trigger a good feeling like "amusement" or "excitement".

Visual sentiment prediction investigates sentiment estimation from images and has been an interesting and challenging research problem. In this paper, the purpose is to accurately estimate the sentiments as a single-label and multi-label multi-class problem from given images that evoke multiple different emotions simultaneously [1].

We also introduce a new combined neural network model which allows inputs coming from both ViT features and pre-trained CNN features. In addition, existing deep learning had weak classifications, therefore we propose a new fully connected 2 layers. The new contributions of this paper include (1) propose a novel feature considering both ViT and CNN features to predict sentiment of images, unlike most recent research which only concerns adopting CNN features, (2) propose a combined feature method to combine the output of each feature, unlike previous work which focuses on combining feature vectors.

2 APPROACH

We propose single-label (Subtask1) and multi-label (Subtask2 and 3) visual sentiment analysis system to predict multiple emotions. And we will describe our deep neural network model that enables the single-label and multi-label outputs, given images that evoke emotions.

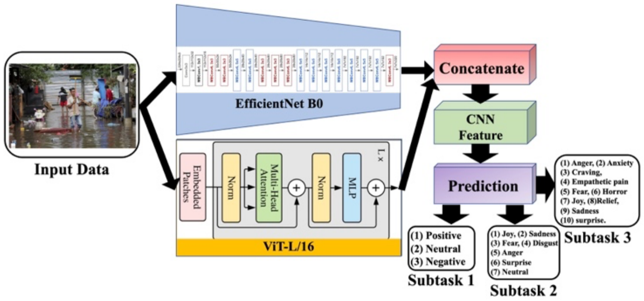


Figure 1: Calculated spatial distribution of the in-plane dynamic magnetization.

2.1 Subtask1

This is a multi-class single label classification task, where the images are arranged in three different classes, namely positive, negative, and neutral. There is a strong imbalance towards the negative class, given the nature of the topic.

To solve our multi-class, single-label classification problem, we propose new combined neural network models which allow inputs coming from both End-to-end (Vision Transformers: ViT and CNN) features.

As illustrated in Figure 1, we adopt ViT-L/16 at ViT and extracted features. On the other hand, CNN features extracted from a pre-trained CNN-based neural network include EfficientNetB0[2].

- Vision Transformers (ViT)

The Vision Transformer is a model for image classification that employs a Transformer-like architecture over patches of the image. This includes the use of Multi-Head Attention, Scaled Dot-Product Attention and other architectural features seen in the Transformer architecture traditionally used for NLP[3].

-CNN

In addition to ViT features described above, our system incorporates CNN features, which can be extracted from pre-

MediaEval'21, December 13-15 2021, Online

trained deep convolutional neural networks with EfficientNetB0. Because of the lack of dataset in visual sentiment analysis, we adopt transfer learning in our feature to prevent over fitting.

We decrease the dimensions of fully-connected layers used in CNN models. Specifically, for EfficientNetB0, we extract a 1280-dimensional vector from 'Global Average Pooling 2D' layer (or the second to the last fully-connected layer), and reduce the vector to 512 dimension by applying a fully-connected layer.

2.2 Subtask2 and 3

In Subtask 2 and 3, this is a multi-class multi-label image classification task, where the participants are provided with multi-labeled images.

To solve our multi-class, multi-label classification problem, we propose new combined neural network models which allow inputs coming from both End-to-end (ViT and CNN) features. We adopt ViT-L/16 at ViT and extracted features. On the other hand, CNN features extracted from a pre-trained CNN-based neural network, include EfficientNetB0.

To deal with the above combined features, we proposed a deep neural network architecture where we allowed multiple inputs and a multi-hot vector output. The combined feature is represented by the following formula:

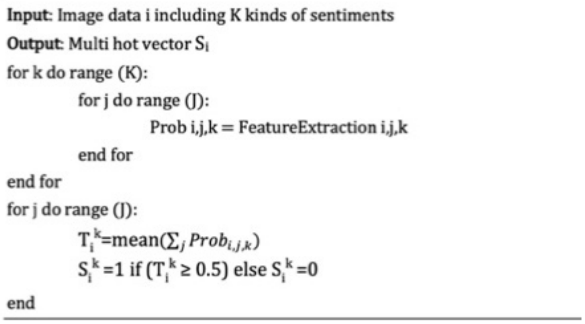
$$combined\,feature = average(\omega_1(ViT) + \omega_2(CNN)) \qquad (1)$$

Based on this formula, after the training process, we allowed our neural network system to predict the visual sentiment of unknown images as a multi-label multi-class classification problem.

-multi-label prediction

To detect a multi-hot vector, we employed a method based on our research [4]. We proposed a method illustrated in Algorithm 1. The input is a collection of features extracted from each image with K kinds of sentiments, while the output is a K-dimensional multi-hot vector.

Algorithm 1: Predicting multi hot vector for an image



```
Input: Image data i including K kinds of sentiments
Output: Multi hot vector $S_i$
for k do range (K):
        for j do range (J):
                Prob i,j,k = FeatureExtraction i,j,k
        end for
end for
for j do range (J):
        $T_i^k$=mean($\sum_j Prob_{i,j,k}$)
        $S_i^k$ =1 if ($T_i^k \geq 0.5$) else $S_i^k$ =0
end
```

In Algorithm 1, we assumed that the extracted features (here ViT and CNN) are represented by their probabilities. For each sentiment, we summed up the features, followed by averaging the result, which is denote by T_i^K in Algorithm 1.

We used a fixed threshold for sentiment, and adopted "0.5" for each feature, which we employed as the threshold of the corresponding emotion evocation. After obtaining all the thresholds dynamically determined based on this threshold, the multi-hot vector of each image is generated such that if T_i^K is equal to or greater than the thresholds, we set S_i^K=1; otherwise S_i^K=0, where S_i^K is the element of K-th sentiment of i-th image. In short, the vector S_i represents the output multi-hot vector. We repeated this computation until all the test (unknown) images were processed.

3 EXPERIMENTAL RESULTS

Here we describe experiments and the evaluations. And, we have divided the training dataset into training and validation data with an 8:2 ratio. We determined the following hyper-parameters; batch size as 256, optimization function as "SGD" with a learning rate of 0.001 and momentum 0.9, and the number of epochs 200. For the evaluations of single-label and multi-label classification, we employed Weighted F1-Score.

Here we compare in terms of Weighted F1-Score. Also, the table includes several base line methods including ViT, EfficientNet B0, and our proposed combined model. The "Dim" column of the table represents the feature dimension. For our proposed combined model, we have tested with one variation, i.e., ViT+ EfficientNet B0. For our proposed combined model, it turns out that ViT+EfficientNet B0 has the best score. It is observed that the proposed method could correctly recognize the images whose emotions are falsely classified by the base CNN.

Table 1: The results of doing experiment

Model	Dim	Weighted F1-Score in Subtask 1	Weighted F1-Score in Subtask 2	Weighted F1-Score in Subtask
ViT-L/16	512	0.692	0.412	0.402
EfficientNet B0	512	0.715	0.534	0.392
Proposed model	1024	**0.753**	**0.585**	**0.415**

4 CONCLUSIONS

We proposed a model for Visual Sentiment Analysis: A Natural Disaster Use-case which accurately estimates single-label and multi-label multi-class problems from given images, evoking multiple different emotions simultaneously.

Our proposed model is simple yet effective and achieves new state-of-the-art performance on multiple datasets.

ACKNOWLEDGMENTS

A part of this research was carried out with the support of the Grant for Toyohashi Heart Center Smart Hospital Joint Research Course and the Grant for Education and Research in Toyohashi University of Technology.

REFERENCES

[1] H, Syed Zohaib and Ahmad, K and Riegler, M and Hicks, S and Conci, N, and Halvorsen, P and Al-Fuqaha, A Al-Fuqaha, 2021, December. Visual Sentiment Analysis: A Natural Disaster Use-case Task at MediaEval 2021. In Proceedings of the MediaEval 2021 Workshop, Online.

[2] Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. *ICML 2019* (05 2019), https://arxiv.org/pdf/1905.11946.pdf.

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., and Houlsby, N. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[4] Asakawa, T., & Aono, M. 2019. Median based Multi-label Prediction by Inflating Emotions with Dyads for Visual Sentiment Analysis. In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (*APSIPA ASC*) (pp. 2008-2014). IEEE.