

HCMUS at MediaEval 2021: Efficient methods of Metadata Embedding and Augmentation for Visual Sentiment Analysis

Bang-Dang Pham^{*1,5}, Nhat-Tan Bui^{*1,4,5}, Minh-Khoi Pham^{*1,5}, Pham Van Ngoan^{2,5},
Truong-Hai Nguyen^{1,5}, Thang-Long Nguyen-Ho^{1,5}, Hai-Dang Nguyen^{1,3,5}, Minh-Triet Tran^{1,3,5}

¹University of Science, VNU-HCM, ²University of Technology, VNU-HCM

³John von Neumann Institute, VNU-HCM, ⁴International Training & Education Center

⁵Vietnam National University, Ho Chi Minh city, Vietnam

pbdang18@apcs.fitus.edu.vn, 1859043@itec.hcmus.edu.vn, pmkhoi@selab.hcmus.edu.vn

1870468@hcmut.edu.vn, nthai18@apcs.fitus.edu.vn, {nhtlong, nhidang}@selab.hcmus.edu.vn, tmtriet@fit.hcmus.edu.vn

ABSTRACT

The Visual Sentiment Analysis Task which is the new task in The Multimedia Evaluation 2021 Challenge concentrates on recognizing emotional responses to natural disaster images. Our HCMUS team performs different approaches based on multiple pretrained models and many techniques to deal with 3 subtasks having a different set of labels for each one. Besides, we also perform the data processing for duplicated images. Based on our experiments, we submit 3 runs. Run 1 introduces a novel way to use meta-features to enhance vision models and combine various effective methods to tackle the class imbalance problems. In run 2, we propose the efficient method to utilize the result of the deep learning model and boost the accuracy in the post-processing step. Finally, run 3 leverages the power of the pretrained model and the new optimizer technique to ameliorate the accuracy. In the final result, our method achieves 0.7716 in task 1, 0.6276 in task 2, and 0.5838 in task 3 - which are the highest scores of our submission.

1 INTRODUCTION

In the Multimedia Evaluation Challenge 2021 (MediaEval2021), Visual Sentiment Analysis is one of the new challenges that has been established to express the emotion or the interior content of images. The aim of Sentiment Analysis is to extract the attitude of people toward a topic or the intended emotional effect the author wishes to have on the readers. By this way, this task might aid in the comprehension of visual materials beyond semantic notions in a variety of application fields such as education, entertainment, advertising, and journalism. We propose 3 different methods which are mainly based on deep learning model to solve the problem in various aspects, and they are would be described in the later sections.

2 RELATED WORK

Visual Sentiment Analysis exists similar objectives to the image classification task. Hence, in our submission, we take advantage of pretrained state-of-the-art CNN models and fine-tune them for this task. We also adapt many improvement methods which are commonly used for computer vision tasks and analyze their effectiveness on this dataset as well.

2.1 Transfer Learning

Pretrained state-of-the-arts models are usually inherited for fine-tuning on small dataset, which helps faster convergence. EfficientNet [9] or the recent EfficientNetv2 [10] and Big Transfer [4] are known to achieve high performance on classification tasks.

2.2 Data Augmentation

Data augmentation has play a central goal to improve performance of training deep vision models. Mixup [12], Cutmix [11] are used to generate combination of images and their corresponding soft labels. Recent popular methods such as RandAugment [2] or AutoAugment [1] automatically choose a sequence of suitable image manipulation operations to increase generalization and robustness to the dataset.

3 METHOD

3.1 Data Processing

We found many duplicated images which are not in the same class. We used a pre-trained EfficientNet to extract feature embedding vectors of every data point and compare them using cosine distance to find identical pairs.

We found 8 pairs of images that are the same but differ in class labels. Table 1 describes some examples. We carefully clean up these conflicts before the training stage.




ImageID	Label	Image
171798f0-6343-41e0-8c51-b96a865bae0b c5a74b3b-dfcb-47ed-afc9-042877a45e1d	Neutral Negative	
78d7325a-9536-428b-a094-f72e1f701220 3a0a8c15-7081-4226-a53e-156a22576467	Negative Positive	
b0e4218a-1cef-4bc9-8257-c6c5a141bf36 3a0a8c15-7081-4226-a53e-156a22576467	Neutral Negative	

Table 1: Samples that have two different labels

3.2 Run 01

In this run, we use various models to tackle the 3 subtasks. In the end, we apply heuristic ensemble method to finalize our results. Experimenting models are Efficientnet, Efficientnetv2 and MetaViT

The Efficientnet and Efficientnetv2 are well-known for its effectiveness in image classification domain. Hence, we use the implementation from timm¹ without much modification. In the MetaViT method, we introduce a more complex pipeline. We use the FaceNet model² as a feature extraction tool, extracting face information from the image. We observed that the majority of the samples were mainly photographs of human activity. Such images are often identified based on the emotions of the main subjects in the scene that the photographer wants to convey. Therefore, we use human face emotions extracted from all samples in the dataset to inform the next inference steps.

Secondly, according to the dataset analysis [3], visible objects in the scene also affect the feelings of the observer, thus we inherit FasterRCNN model³ to extract bottom-up attention features as additional information. Finally, for each sample, two meta features are embedded into the same vector space, then concatenated and forwarded through a pretrained Vision Transformer¹ for learning.

To ensemble all our models, we perform majority votes technique on prediction of these models.

In order to deal with imbalance dataset, which is a huge problem of the task, we apply following methods:

- Use Focal loss [5] as objective function for task 1.
- Strong data augmentation techniques such as RandAugment, CutMix, MixUp [12] to add more training samples.
- Smart sampling technique - split samples into batches so that all classes are evenly distributed for each batch.

3.3 Run 02

In this run, we try to utilize more features of the deep model through post-processing the extracted vector after classifying. The main model of this approach is based on the architecture EfficientNet-B1[8], but with Binary Cross Entropy (BCE) loss and Sigmoid function for optimizing each label's result in all tasks. In task 1, after processing with the model, we set the class of the image based on the max value of the logit vector after going through Sigmoid function. We analyze that the given dataset has the imbalance problem with the gap samples between positive-negative and neutral, that leads to the result after solving almost having a positive and negative label. For improvement, we set a threshold that if the max value we get does not exceed this one, we would assign its label neutral. By this way, the class neutral would easily be emphasized by the deep model, which means the "emotion" of the picture does not gravitate exactly towards positive or negative.

In addition, we analyze that the image's sentiment quite depends on the emotion of people in that one. Hence, if the image in the dataset has the smiling face, then all of them are labeled positive (task 1) and joy (task 2 & 3). By contrast, the image having crying emotion or sadness on the face is labeled negative (task 1) and other negative meaning labels in task 2,3. Based on that, we apply Face

Detection² and Face Emotion Recognition² to boost the result more accurately for some specific label. Moreover, to overcome the imbalance dataset problem, in this approach, we still apply the smart sampling technique and strong augmentation as mentioned above to reduce the bias prediction.

3.4 Run 03

In this run, we make use of different versions of pretrained Big Transfer (BiT) [4] in Tensorflow Hub⁴ to deal with 3 subtasks. Big Transfer is the pretrained model on huge datasets to attain impressive accuracy in any given new dataset.

For task 1, we utilize medium version⁵ of BiT which is composed of a ResNet-152 four times wider (R152x4) pretrained on ImageNet-21k. The optimizer is Rectified Adam [6] with 1e-3 initial learning rate combine with Lookahead [7] technique to ameliorate the learning stability. We empirically find that using MixUp [12] data augmentation technique with the ImageDataGenerator of Keras can increase the F1 score for task 1.

We modify the medium version⁶ of BiT for task 2 to ResNet-101 three times wider (R101x3). The optimizer is also the combination of Rectified Adam and Lookahead technique, however, the learning rate is set to 5e-4. We also use the multi-label Focal Loss to tackle the imbalance dataset.

In the final task, we configure the same model as task 2 except the learning rate is set to 8e-4. Note that all tasks are trained with batch size is 32.

4 EXPERIMENTS AND RESULTS

Table 2 shows the results of our 3 runs in term of Weighted F1-Score.

Team-run	Task-1	Task-2	Task-3
SELAB-HCMUS khoi_submission/run1	0.7224	0.6054	0.5838
SELAB-HCMUS pbdang_submission/run2	0.7716	0.6276	0.5254
SELAB-HCMUS tan_ngoan_result/run3	0.7217	0.6047	0.542

Table 2: HCMUS Team Submission results for Visual Sentiment Analysis Task

5 CONCLUSION AND FUTURE WORKS

In summary, we identify challenges of the dataset and propose different approaches to address the issues. We conclude that the sentiment task is heavily biased towards the viewer feelings, therefore making it difficult for neural networks to learn the true label. In the future, we aim to tackle the challenge in semi/unsupervised manner such as utilizing metric learning to capture the semantic representation of each class.

ACKNOWLEDGMENTS

This work was funded by Gia Lam Urban Development and Investment Company Limited, Vingroup and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2019.DA19

¹<https://github.com/rwightman/pytorch-image-models>

²<https://github.com/timesler/facenet-pytorch>

³<https://github.com/airsplay/py-bottom-up-attention>

⁴<https://tfhub.dev>

⁵<https://tfhub.dev/google/bit/m-r152x4/1>

⁶<https://tfhub.dev/google/bit/m-r101x3/1>

REFERENCES

- [1] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. 2019. AutoAugment: Learning Augmentation Policies from Data. (2019). arXiv:cs.CV/1805.09501
- [2] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. 2019. RandAugment: Practical automated data augmentation with a reduced search space. (2019). arXiv:cs.CV/1909.13719
- [3] Syed Zohaib Hassan, Kashif Ahmad, Steven Hicks, Paal Halvorsen, Ala Al-Fuqaha, Nicola Conci, and Michael Riegler. 2020. Visual Sentiment Analysis from Disaster Images in Social Media. (2020). arXiv:cs.CV/2009.03051
- [4] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2020. Big Transfer (BiT): General Visual Representation Learning. In *European Conference on Computer Vision*. Springer, Cham.
- [5] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal Loss for Dense Object Detection. (2018). arXiv:cs.CV/1708.02002
- [6] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the Variance of the Adaptive Learning Rate and Beyond. In *International Conference on Learning Representations*.
- [7] Michael R. Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. 2019. Lookahead Optimizer: k steps forward, 1 step back. In *Conference on Neural Information Processing Systems*.
- [8] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 6105–6114.
- [9] Mingxing Tan and Quoc V. Le. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. (2020). arXiv:cs.LG/1905.11946
- [10] Mingxing Tan and Quoc V. Le. 2021. EfficientNetV2: Smaller Models and Faster Training. (2021). arXiv:cs.CV/2104.00298
- [11] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. (2019). arXiv:cs.CV/1905.04899
- [12] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.