WHERE IS THE EMOTION? DISSECTING A MULTI-GAP NETWORK FOR IMAGE EMOTION CLASSIFICATION

Lucinda Lim *Huai-Qian Khor* *Phatcharawat Chaemchoy* *John See* *Lai-Kuan Wong*

Visual Processing Lab, Faculty of Computing and Informatics,
Multimedia University, Malaysia

ABSTRACT

Image emotion recognition has become an increasingly popular research domain in the area of image processing and affective computing. Despite fast-improving classification performance in this task, the understanding and interpretability of its performance are still lacking as there are limited studies on which part of an image would invoke a particular emotion. In this work, we propose a Multi-GAP deep neural network for image emotion classification, which is extensible to accommodate multiple streams of information. We also incorporate feature dependency into our network blocks by adding a bi-directional GRU network to learn transitional features. We report extensive results on the variants of our proposed network and provide valuable perspectives into the class-activated regions via Grad-CAM, and network depth contributions by truncation strategy.

***Index Terms*—** Image emotion classification, multi-GAP, CNN, visualizations, class activation maps

1. INTRODUCTION

In the booming era of social media and photo-sharing networks, millions of images are used everyday and it is fast becoming a preferred medium over text in terms of expression. Psychological studies [1][2] have demonstrated that various human emotions can be evoked through visual stimuli and a few studies [3, 4] have classified the hierarchy and layout of human emotions into numerous classes which are used in computational analysis of image emotion. With the increasing attention on analyzing image emotion, two large labelled datasets, FI [5] and WebEMO [6] have been released to facilitate more robust learning of image emotion recognition.

Earlier works mostly revolved around prior knowledge and handcrafted features. An early work by Machajdik & Hanbury [7] exploited concepts from psychology and art theory to extract a list of features based on color, texture, composition and content. Another work [8] crafted the relation between artistic principles (e.g. balance, emphasis and harmony) and image emotion. Another work [9] created a visual sentiment ontology by forming Adjective-Noun pairs (ANP) with text data, to be popularly known as SentiBank. Several other works [10][11] leveraged on image handcrafted features such as edges, blobs and Histogram of Oriented Gradients (HOG). However, the content of image emotion samples come in various forms such as text, human faces, common objects and gestures. The high variability may hinder the performance of handcrafted features as well as complicate the process of formulating prior knowledge.

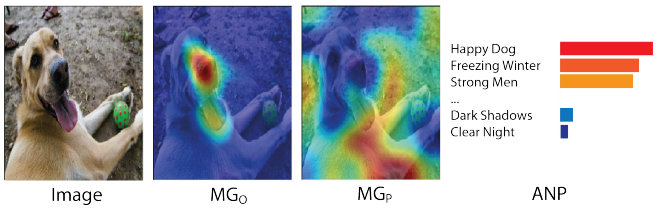

Fig. 1: The class activation maps for respective network on "contentment" as well as its ANP feature vector.

Motivated by the robustness of deep neural networks along with the availability of large datasets, a few works [12, 5] attempted transfer learning to extract deep features including a recent multi-stream lightweight model [13] with similar methodology. Apart from that, He et al. [14] introduced assisted learning where features of a binary CNN that is trained to identify positive or negative emotions in image, are concatenated with features from a finer-grained CNN before classification. Besides, several works [15] [16] emphasized on multi-level feature extraction and learning of feature dependencies between each block which yielded excellent recognition performance. However, these works provide limited insights into what goes on under the hood.

Despite the ever-growing performance of image emotion recognition, most existing works come with limited qualitative analysis, which is as equally crucial as quantitative analysis to better understanding the latent emotional content within images. Moreover, image emotion data comes with a broad variety of images which can consist of human faces, overlaid text as well as daily objects, hence it is also important to analyze whether the algorithm has succinctly captured relevant areas when classifying emotions.

The contributions of this work are as follows:

1. We propose a two-way fusion approach that first inte-

grates different hierarchical levels of information, followed by combining multiple streams of information.

2. We demonstrate the spatial contributions of the streams via gradient-based class activation visualization of the proposed deep neural network.

3. We investigate the impact of different network depths by applying a truncation strategy on the GAP features of each network.

2. METHODS

Motivated by the idea of harnessing features from each depth of a network [15] and the effective use of global average pooling (GAP) [17] operations, we propose a network named Multi-GAP (MG) network. To further extend the idea of MG, we utilize the MG-derived features in two distinct ways: (1) learning feature dependency through a bi-directional GRU, and (2) exploiting early and late fusion schemes, together with high-level Adjective-Noun Pairs (ANP) features to create a richer pool of features for classification. The general framework is shown in Figure 2. In our work, two networks representing the Object and Places information streams are utilized: the Object Network uses a MobileNet [18] pre-trained on ImageNet [19], while the Places Network uses a VGG16 [20] pre-trained on the Places dataset [21].

2.1. Multi-GAP (MG) Feature Extraction

In deep neural networks, earlier layers capture generic features such as edges and contours while later layers enclose domain specific features such as face appearance or detailed contours of certain objects. Considering that the emotion invoked from an image can be quite subjective, and that relying on classical object classification may seemed insufficient, we propose to extract features from earlier to later layers and merge them together. Figure 2(a) outlines the proposed Multi-GAP (MG) base network: each block consists of a convolutional (conv) layer $\mathcal{C}$ (shown in yellow with the number of channels), followed by a 1×1 conv layer $\hat{\mathcal{C}}$ (shown in purple) which acts to reduce the channel dimensions to 64. Then, it passes through a Global Average Pooling (GAP) layer G which is a resource-saving alternative to a fully-connected layer. The output of the MG network can be formally denoted by a concatenation of the GAP layer outputs $\mathbb{MG} = \{MG_i, \ldots, MG_N\}$, where N is the number of blocks:

$$MG_i = G(\hat{\mathcal{C}}_i(\mathcal{C}_i(\cdot))), \quad \forall i \in \{1, N\} \tag{1}$$

2.2. MG with Feature Dependency

Given that early features play an important role in learning better representations, we propose to learn the feature dependency between each adjacent MG block using a single-layer bi-directional GRU (Bi-GRU) network. As shown in Figure 2(b), the outputs of the MG network are now treated as T sequential timesteps in a recurrent network ($T = N$) and hence, the transition of features from early to late layers can be learned via the Bi-GRU network which has M recurrent units. As an additional improvement, we attach a learnable soft attention mask, Att to accentuate important recurrent features while de-emphasizing the less crucial ones. The attention mask Att (shown in Equation 2) is a feature vector containing the attention values for each t-th timestep, where W are the learnable weights, b is the bias and $x_t \in \mathcal{R}^M$ is the recurrent output.

$$Att_t = (tanh(W_t \cdot x_t + b_t))^2 \tag{2}$$

Following the computation of Att, the Hadamard product is then computed between the recurrent output, x and Att to obtain the attention-weighted recurrent output.

2.3. Fusing MG features

Fusion schemes are common strategies that can be exploited to accumulate information from multiple feature descriptors or classifiers to further create more robust frameworks. In our work, we explore both early (feature-level) and late (output-level) fusion. In contrast to the work of [15], we extend our MG network into multiple streams so that the fusion schemes can be exploited to take advantage of information learned through different networks. Specifically, for early fusion, we use features derived from the Object MG network and Places MG network, together with high-level Adjective-Noun Pair (ANP) features (shown in Figure 2(c)). This process merges the bottlenecked features (without the classifier head) of the networks (except the ANP which are merely features). Meanwhile, late fusion is performed by averaging the predicted softmax values of all classifiers; the class with the highest averaged value is chosen as the predicted class. Note that the ANP features are put through a MLP network with 2 hidden layers (1024 and 512 nodes) and a 8-node output layer.

3. EXPERIMENTAL RESULTS

3.1. Datasets

For fair benchmarking, the **FI** dataset [5], a large scale in-the-wild labelled dataset, is chosen for a majority of our experiments. The dataset consists of eight emotion labels defined by a psychological study [4] also known as Mikels' emotion wheel. The total number of images downloaded from the dataset is reduced due to unavailable links. Filtering are applied to obtain only images labeled with three or more agreements. A total of 22,260 samples (after cleaning) are split into training (80%), test (15%) and validation (5%) sets following partitioning ratio used in [5].

On the other hand, a recent and larger image emotion dataset, namely **WebEMO** [6] was established with 235,327

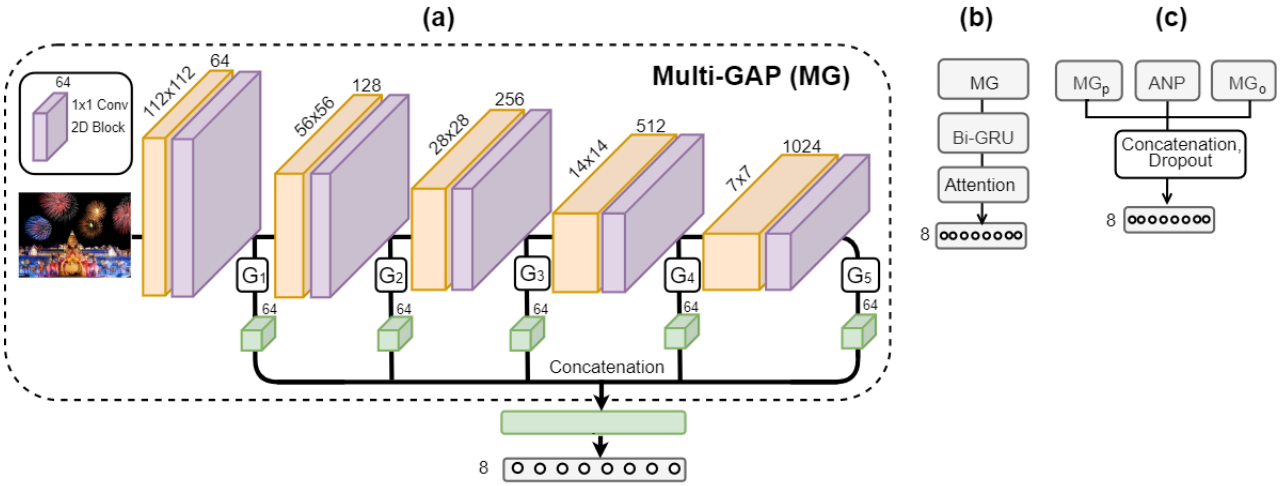

Fig. 2: (a) The proposed Multi-GAP (MG) base network which extracts features from each network block. The GAP layers (G) reduce the overall network parameters, making it less susceptible to over-fitting. Two flavours are explored: (b) MG with feature dependency (Bi-GRU + Attention). (c) Fusion of MG features (MG_p and MG_o) and other auxilliary features (ANP).

total image samples with minimal class imbalance issue. The labels in this dataset were given following the paradigm of Parrott's wheel of emotion, where the emotions are defined through a tree-like hierarchy from coarse to fine emotions. In our experiments, we choose the intermediate hierarchy of emotions which contains seven emotion classes, *i.e.* 'Anger', 'Fear', 'Joy', 'Love', 'Sadness', 'Surprise', 'Confusion', as it is closer to FI's eight classes. Due to the scale of this dataset, we only used this for evaluation purposes (see Section 4.3).

3.2. Experiments on FI Dataset

Table 1 shows the comparison between different variants of the proposed multi-GAP (MG) network, with the different streams (Object, Places, ANP) depicted along the middle three columns. We observe that feature fusion is able to outperform single stream network for most of the time. By fusing the features from ANP, Object and Places networks early, we achieve the best accuracy of 64.30%. We also note that the addition of Bi-GRU was not effective in most cases.

To better depict the per-class performance of our best method, we show its confusion matrix in Figure 3. The performance on 'Anger' and 'Fear' are worst among all as they are inherently very similar kinds of emotion (which are better separated with valence-arousal values), but they are also among the smaller classes. 'Disgust', 'Awe' and 'Sadness' performed better considering their meagre number of samples as compared to larger classes *i.e.* 'Amusement' and 'Contentment'. Overall, the performance of our early fusion is slightly bias towards the larger classes and this underlines much room for improvement in future. Against other methods in literature, our method is able to surpass the deep learning baseline of [5] (58.30%), and is comparable to other recent works: EOC-CNN [13] (64.87%), MldrNet [15] (65.23%).

Network Type	O	P	ANP	Acc.(%)
MG	✔			63.70
MG		✔		59.21
MLP			✔	47.37
MG + Bi-GRU	✔			63.40
MG + Bi-GRU + Att	✔			63.46
MG + Bi-GRU + Att		✔		58.40
MG (EF)	✔	✔	✔	64.30
MG (LF)	✔	✔	✔	63.48
MG (LF)	✔	✔		64.00
MG + Bi-GRU + Att (LF)	✔	✔		63.76

Table 1: Performance comparison between variants of Multi-GAP (MG) network. EF: Early Fusion; LF: Late Fusion

4. DISCUSSION

In this section, we provide several perspectives of the proposed network for analysis and discussion.

4.1. Grad-CAM Visualization

Besides reducing complexity, the insertion of the GAP layer into our networks provides us with a means to look into what our network "sees". We utilize Grad-CAM [22] to visualize the area of activations that contributes to the predicted class. Grad-CAM computes an activation heatmap which is an upsampled projection of the weighted activations of the last convolutional layer.

Particularly, Grad-CAM visualizations from different information streams can facilitate explainability in networks. In Figure 5, the top image shows a test image that has been correctly predicted on both Places and Object networks. Note that the 'Amusement' emotion is strongly associated with

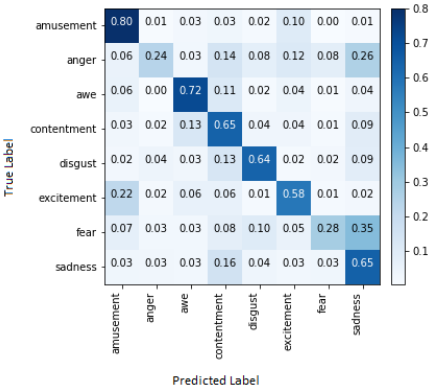

Fig. 3: Confusion matrix of the MG (Early Fusion) approach.

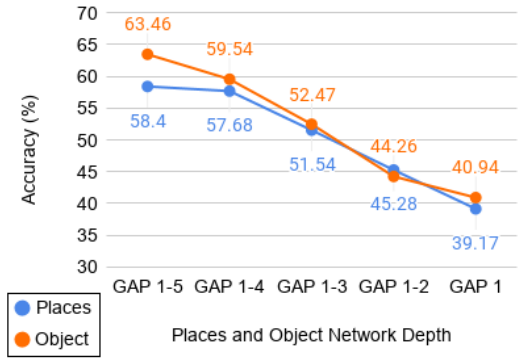

Fig. 4: GAP 1-4 refers to model without GAP 5 and its respective convolutional block.

concepts around the ferris wheel (for Places) and also the sea area (for ImageNet). Meanwhile, the emotion of the bottom image is incorrectly predicted on both networks ('Amusement' instead of the ground-truth 'Disgust') as it fails to activate important visual clues due to lack of scenic information. The middle image shows an image that benefited from the fusion of confidence scores from both networks to correctly predict the 'Awe' emotion. Even though the Places network returned a weak incorrect prediction due to the light streaks, the Object network showed strong confidence in the vast night sky area, linking it to 'Awe' category. Therefore, it is crucial to consider complementary information from multiple streams to boost classification performance. More visualizations and explanations are provided in the Supplementary Materials.

4.2. Ablation Study

To identify the influence of each GAP, a truncation strategy is applied where the convolutional blocks are taken off one-by-one from the top of the network (which is then trained again). Figure 4 shows the result of this ablation study based on MG+Bi-GRU+Att method. Expectedly, for both networks, the accuracy slips upon removal of each block. For both networks, the accuracy drops the most from GAP 1-3 TO GAP 1-2, (Object:-8.21%, Places:-6.26%). This shows that GAP 3

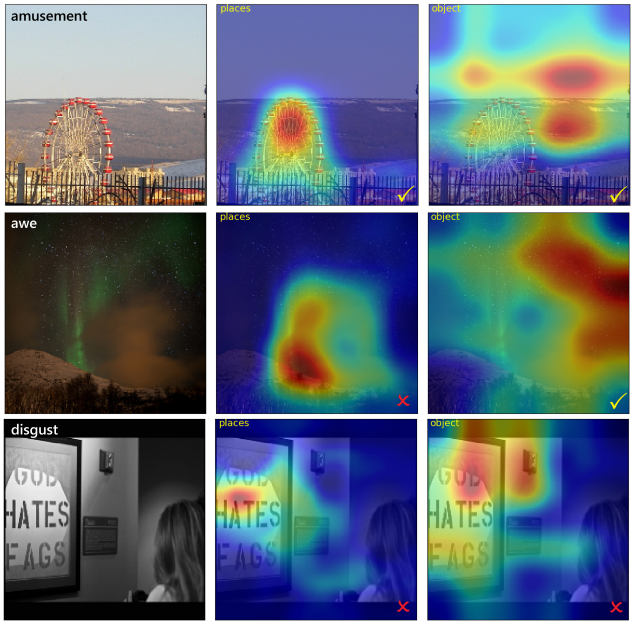

Fig. 5: Grad-CAM visualization of sample test images on the Places and Object networks which resulted in predictions that are both correct (first row), one correct, one wrong (second row) and both incorrect (third row).

is likely the most salient, followed by GAP 4. Meanwhile, the performance drop in Places network from GAP 1-5 to GAP 1-4 is quite minimal, which signifies that the last convolutional block of Places network is less distinctive among all and could be truncated to minimize computations.

4.3. WebEMO

For further evaluation, we also performed evaluation experiments on the WebEMO dataset ($>$ 180K images) using the FI-trained models. Due to computational constraints, we only ran evaluation experiments for the single stream MG networks for Places and Object. The Object Network (46.45%) faired better than the Places Network (42.25%), which indicates that object-based features are likely more meaningful. This also shows that the WebEMO dataset is a much more challenging dataset as compared to the FI dataset and there is room for further research using this dataset.

5. CONCLUSION

In this paper, we propose a Multi-GAP (MG) network coupled with fusion strategies that are able to exploit a wide variety of descriptive features with a reduced number of parameters due to multiple levels of GAP layers. The paper also provides several perspectives: activated regions, depth contributions, that provide further explainability to how deep neural networks perceive emotion in images.

6. REFERENCES

[1] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, 2011.

[2] P. J. Lang, "A bio-informational theory of emotional imagery," *Psychophysiology*, vol. 16, no. 6, pp. 495–512, 1979.

[3] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[4] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behavior research methods*, vol. 37, no. 4, pp. 626–630, 2005.

[5] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark." in *AAAI*, 2016, pp. 308–314.

[6] R. Panda, J. Zhang, H. Li, J.-Y. Lee, X. Lu, and A. K. Roy-Chowdhury, "Contemplating visual emotions: Understanding and overcoming dataset bias," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 579–595.

[7] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 83–92.

[8] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 47–56.

[9] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 223–232.

[10] L. Chang, Y. Chen, F. Li, M. Sun, and C. Yang, "Affective image classification using multi-scale emotion factorization features," in *2016 International Conference on Virtual Reality and Visualization (ICVRV)*. IEEE, 2016, pp. 170–174.

[11] S. Zhao, H. Yao, Y. Yang, and Y. Zhang, "Affective image retrieval via multi-graph learning," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1025–1028.

[12] M. Chen, L. Zhang, and J. P. Allebach, "Learning deep features for image emotion classification," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 4491–4495.

[13] Y.-H. Chew, L.-K. Wong, J. See, H.-Q. Khor, and B. Abivishaq, "LiteEmo: Lightweight deep neural networks for image emotion recognition," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–6.

[14] X. He and W. Zhang, "Emotion recognition by assisted learning with convolutional neural networks," *Neurocomputing*, vol. 291, pp. 187–194, 2018.

[15] T. Rao, X. Li, and M. Xu, "Learning multi-level deep representations for image emotion classification," *Neural Processing Letters*, pp. 1–19, 2016.

[16] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, and D. Xu, "Dependency exploitation: A unified cnn-rnn approach for visual emotion recognition." in *IJCAI*, 2017, pp. 3595–3601.

[17] Y.-L. Hii, J. See, M. Kairanbay, and L.-K. Wong, "Multigap: Multi-pooled inception network with text augmentation for aesthetic prediction of photographs," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 1722–1726.

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE CVPR*, 2018, pp. 4510–4520.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009, pp. 248–255.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.