Exploiting Out-of-Domain Datasets and Visual Representations for Image Sentiment Classification

Alexandros Pournaras, Nikolaos Gkalelis, Damianos Galanopoulos and Vasileios Mezaris
CERTH-ITI
6th Km Charilaou-Thermi Road, P.O. BOX 60361 Thessaloniki, Greece
Email: {apournaras, gkalelis, dgalanop, bmezaris}@iti.gr

***Abstract*—Visual sentiment analysis has recently gained attention as an important means of opinion mining, with many applications. It involves a high level of abstraction and subjectivity, which makes it a challenging task. The most recent works are based on deep convolutional neural networks, and exploit transfer learning from other image classification tasks. However, transferring knowledge from tasks other than image classification has not been investigated in the literature. Motivated by this, in this work we examine the potential of transferring knowledge from several pre-trained networks, some of which are out-of-domain. We show that by simply concatenating these diverse feature vectors we construct a rich image representation that can be used to train a classifier with state of the art performance on image sentiment analysis. We also evaluate a Mixture of Experts approach, for learning from this combination of representations, and highlight its performance advantages. We compare against the top-performing recently-published methods on four popular benchmark datasets and report new SOTA results on three of the four.**

I. Introduction

Visual sentiment analysis refers to the problem of identifying the sentiment conveyed by an image. The term sentiment may be used to define either the emotion (sad, fear, joy etc.) or the polarity (positive, negative). The problem has recently attracted significant attention due to the large-scale use of images in social media. Images have the power to convey strong emotions, thus sentiment analysis is an important means of opinion mining with lots of applications in education, entertainment and open source intelligence [1]. Visual sentiment analysis is a challenging task because it involves a higher level of human subjectivity in the classification process than other image classification tasks. Many factors, such as the subject's ethnicity, culture and experiences play an important role to the sentiment that an image will convey to her or him.

Several image classification methods have recently been proposed to deal with the sentiment classification problem. These include training deep convolutional neural networks from scratch on sentiment-annotated training corpora, or fine-tuning pre-trained networks, that most commonly have been originally trained on ImageNet [2]. Many of the published works focus on experimenting with advanced neural network architectures, exploiting the principles of visual attention. However, little emphasis has been given on investigating the potential of transferring knowledge learned from neural networks trained on tasks other than image classification. For this reason, we employ several pre-trained networks, some of which are trained on out-of-domain datasets and tasks. We extract their encodings and concatenate them to create a rich representation. Moreover, a well-known issue in image sentiment analysis is the large intra-class variance, which is much larger than in most other image classification tasks. This motivated us to experiment with a Mixture of Experts approach, with the hope that multiple experts implicitly partitioning and representing different subclasses of the original classes would be able to tackle this issue more effectively. To the best of our knowledge, it is the first time a MoE approach is being investigated for this problem.

Our contribution is:

- To show that combining encodings learned from different architectures, datasets and tasks can generate rich image representations that enable even simple classifiers to reach (and even exceed) state of the art performance on image sentiment classification, and
- To demonstrate that the MoE approach, which implicitly derives appropriate partitions of the feature space and reduces the intra-class variance for the classification problem at hand, can lead to further improved results.

II. Related Work

A. Image sentiment analysis

Early work on visual sentiment analysis focused on utilizing hand-crafted features, such as color and texture, to train classifiers. Theoretical and empirical concepts from psychology and art theory were exploited in [3] to extract features and perform emotion classification. The most important early work was [4], where a large-scale visual sentiment ontology of adjective-noun pairs (ANPs) was created to serve as a mid-level representation, with the aim to bridge the "affective gap" between low-level hand-crafted features and sentiments. A large but noisy dataset was created by retrieving images from each ANP from Flickr. A visual concept detector library, named Sentibank, was then created to detect ANPs in images.

The advent of convolutional neural networks and deep learning brought superior performance in almost all visual recognition tasks [5], and shifted the focus to deep-learning-based approaches. The networks are either trained from scratch or fine-tuned on pre-trained models. In [6] the authors of [4] improved on their classifiers previous performance by employing deep neural nets. Some authors later used transfer

978-1-7281-5919-5/20/$31.00 ©2021 European Union

learning to train networks on the Flickr [4] dataset and then fine-tune them on other datasets [7] [8]. Others, [9] and [10], have followed the same approach, but by using networks pre-trained on ILSVRC/ImageNet [2].

Other works exploit more sophisticated network architectures, such as networks with a residual attention unit [11] or other attention mechanisms, with the aim to extract more localized information [12] [13]. The authors of [14] proposed a two-branch convolutional network. The first fully convolutional branch produces a sentiment map; the second branch utilizes both the holistic and the localized information to perform the classification. Others have proposed using off-the-shelf methods to detect objects [15], or saliency regions [16] and then fuse the global and local features for performing sentiment classification. In [17] a semi-supervised approach based on a teacher-student model was proposed to cope with the lack of high-quality annotated images for sentiment analysis.

Most recently, methods that investigate the value of image semantics have been proposed. In [18], typical visual features are combined with features from a Bayesian network trained on object semantics and sentiment relations. In [19], off-the-shelf visual concept detectors and captioning algorithms are employed to extract descriptive text from the images and then fuse visual deep-learning-based features with the text features to train a sentiment classifier.

B. Mixture of Experts

The original formulation of Mixture of Experts (MoE) was introduced in [20] as a learning procedure involving several "expert" networks that implicitly learn different subsets of the training cases. More recently, the concept of MoE has been applied on several image classification tasks. In [21] it was shown that applying the MoE method to deep convolutional neural networks improved the performance on large-scale image classification tasks. In the architecture of [21], the first part of the network produces a shared encoding that is then passed from the experts and the gate. In [22] a sequential MoE architecture was proposed that can be applied on standard convolutional networks. Contrary to the previous work, in this one the individual layers play the role of the "experts" to dynamically increase the capacity of the network without a proportional increase in computational complexity.

III. Proposed Method

Our proposed method is based on transferring knowledge from 5 different neural networks. These networks have different architectures and are trained on different datasets, some for problems other than image classification. They were chosen for use in this study because they perform very well in their respective domains. A feature vector is extracted from each network. In the following section, we briefly describe each network and how each feature vector is extracted. We classify each feature vector in one of the two categories, either in-domain for those coming from networks trained on image classification tasks, or out-of-domain for those coming from networks trained on other tasks. A summary of the employed feature vectors can be seen in Table I.

A. In-Domain Feature Vectors

1) EfficientNet features: EfficientNet [23] is a recently proposed deep convolutional neural network architecture that achieves state-of-the-art performance on image classification tasks. EfficientNet offers many variations, from the lightweight "B0" model with 5.3 million parameters, to the heaviest "B7" with 66 million parameters. For our work we used the relatively lightweight "B2" model with 9.2 million parameters. We used a model pre-trained on the 1000-class ImageNet dataset. We remove the last fully connected layer, so the network outputs a 1408-element feature vector, $\mathbf{E}$.

2) Resnet features: Resnet [24] is a family of convolutional neural networks based on residual blocks, that have shown state-of-the-art performance in image classification tasks. We use the 152-layer deep Resnet architecture trained on the 11k-class ImageNet dataset [2] and extract the 2048-element "pool5" layer as the feature vector, $\mathbf{R}$.

B. Out-of-Domain Feature Vectors

1) YT8M features: The YouTube-8M [25] is the largest video dataset containing approximately 6 million videos with a total duration of more than 500.000 hours and labeled with 3862 classes. We use extracted features for every second of each video. For this initial feature extraction, an Inception neural network [26] pre-trained on Imagenet [2] is used. The ReLU activation of the last hidden layer of the network is given as input to a rather simple CNN. It consists of a 1D convolutional layer with 64 filters, a max-pooling layer, a dropout and a Sigmoid of 3862 outputs. The 3862-element output vector is our feature vector, $\mathbf{Y}$. In contrast to the EfficientNet and Resnet feature vectors discussed in Section III-A, this feature vector contains semantic-level information.

2) "Signature" features: To obtain the "signature" features, we utilize a cross-modal network designed for ad-hoc video search. More specifically, the attention-based dual encoding network presented in [27] is used. The network is trained to translate a media item (i.e. an entire video) $\mathbf{V}$ or a textual item (i.e. a natural-language video caption or search query) $\mathbf{T}$ into a new common feature space $f(\cdot)$, resulting in representations $f(\mathbf{V})$ or $f(\mathbf{T})$, respectively; such representations, despite being derived from different data modalities, are directly comparable. An illustration of this translation is displayed in Fig. 1. The network is trained using large-sets of video-caption pairs in order to find the optimal common feature space.

At the training stage, the network encodes each video into a three-level representation. As the first level, the entire video is sampled by a fixed number of keyframes. Each keyframe is represented as the output of the flattened pool5 layer of the Resnet-152 network trained on the ImageNet 11K dataset [2]. The mean-pooling of these vectors constitutes the first level representation. The keyframe's vectors are then fed into an attention-based bi-GRU [28] [27] and in this way the second-level representation is generated. Finally, the output of the second-level encoding is fed into a biGRU-CNN [29], resulting in the third-level representation. The global video representation is the concatenation of these three representations, which is forwarded into a fully connected layer. Similarly to the visual modality, a textual global representation $f(\mathbf{T})$ consisting of three levels (i.e. mean-pooling, attention-based

TABLE I. BASIC INFORMATION FOR ALL FEATURE VECTORS.

Feature used	Base architecture	Training dataset	Task
EfficientNet	EfficientNet-B2	Imagenet 1k concepts	image classification
Resnet	Resnet-152	Imagenet 11k concepts	image classification
YT8M	Inception	Youtube8M	video classification
Signatures	Attention-based dual encoding network	Imagenet 11k, MSR-VTT, TGIF, Vatex, ActivityNet	ad-hoc video search
GCN	Resnet-152, Faster R-CNN, GCN	ImageNet 1k, FCVID, YLI-MED	video event recognition

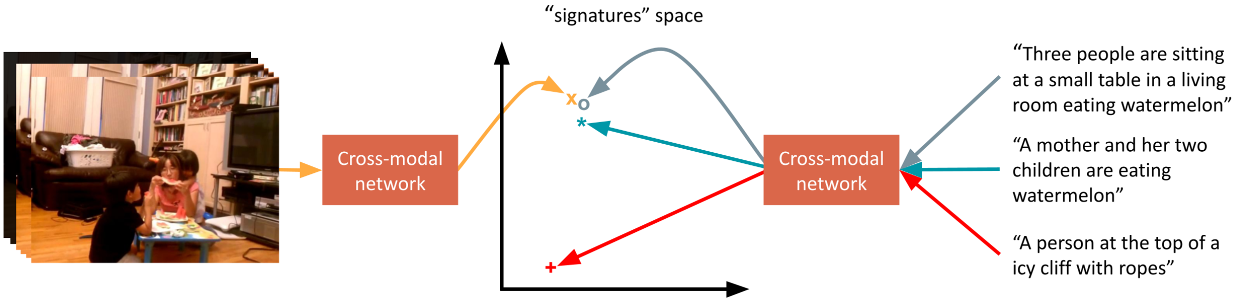
"signatures" space
Cross-modal network
Cross-modal network
x
o
*
+
"Three people are sitting at a small table in a living room eating watermelon"
"A mother and her two children are eating watermelon"
"A person at the top of a icy cliff with ropes"

Fig. 1. An illustration of the transformation that the adopted cross-modal network performs on the videos and captions, representing them into the common "signature" feature space.

bi-GRU sequential model, and biGRU-CNN) is generated for the corresponding captions. The network is trained using the combination of four large-scale video-caption datasets: MSR-VTT [30], TGIF [31], Vatex [32] and ActivityNet [33]. These bring together, in their training portions, about 680.000 video-caption pairs, which were used in this work for training the network architecture of [27].

For leveraging this trained network as a feature generator in the image sentiment classification task, we considered an image as a special type of video comprising only one keyframe. The image is used as input to the visual encoding branch of the network, fed forward through the multi-level encoding layers, and the global image representation $f(\mathbf{V})$, a 2048-element vector, is used as our "signature" feature **S**.

3) Graph Convolutional Network (GCN) features: To obtain this feature vector, we employ a neural network used for the task of video event recognition [34]. Following the application of an object detector on the frames of the video, a neural network is used to extract the object features and graphs are used to model the relations between objects. Then, a graph convolutional network (GCN) is utilized to perform reasoning on the graphs. The resulting object-based frame-level features are then forwarded to a long short-term memory (LSTM) network for video event recognition.

More specifically, uniform sampling is first applied to represent each video with a sequence of 9 frames. Then we employ a Faster R-CNN object detector [35] to derive 50 objects for each video frame, where each object is associated with a bounding box, an object class label and a feature vector of dimensionality F = 2048. Moreover, a feature extractor (the pool5 layer of a pretrained ResNet-152 on ImageNet11K) is applied on the entire frame to derive a 2048-dimensional feature vector, encoding the global appearance information. The extracted feature vectors are then utilized for learning a GCN, a LSTM and two fully-connected layers that comprise our model. We use a two-layer GCN with 2048 units for each layer, an LSTM layer of 4096 units, two FC layers with 2048 and 239 units, respectively, and a sigmoid nonlinearity is utilized on the last FC layer to facilitate multilabel learning. Finally, the model is trained end-to-end on the FCVID [36] dataset.

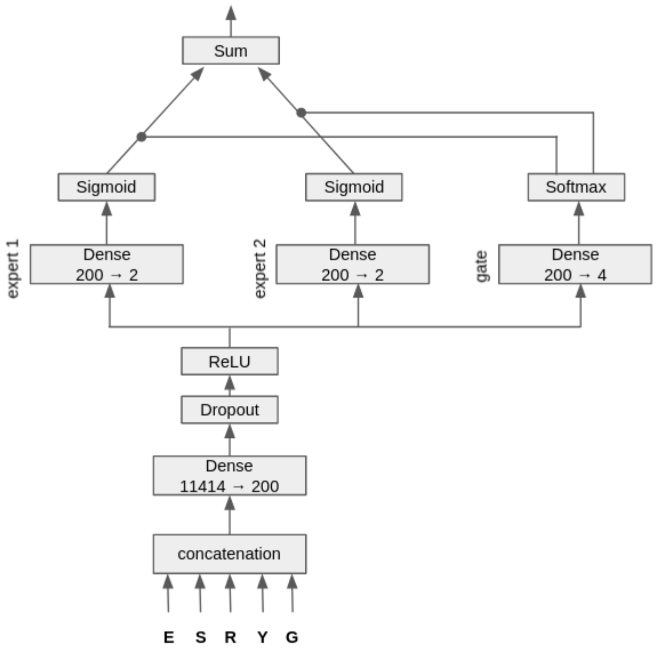

Fig. 2. The overall system architecture. **E**, **S**, **R**, **Y** and **G** are the 5 different feature vectors.

To extract the feature vector that we use for the image sentiment analysis in this work, we fetch the output of the GCN, which is a 2048-element vector **G**.

C. Mixture of Experts

The MoE-based network architecture utilized in our work is illustrated in Figure 2 and is based on the MoE approach described in [25]. More specifically, we concatenate the 5 feature vectors described above, resulting in a final 11414-element feature vector, that will be used to train our Mixture of Experts classifier. This feature vector is fist passed through a fully connected layer, yielding a 200-element vector. After passing through a Dropout and a ReLU block, this 200-element vector is the input I forwarded to the $i = 2$ experts, $e^c_1(), e^c_2()$, which are defined for each class c, as well as to the associated gates, $g^c_1(), g^c_2()$. For each class, there is also defined an extra "dummy" expert (not shown in Figure 2 for simplicity of illustration) that represents the rest-of-the-world class, and only participates in partitioning the feature space through the gate component of the Mixture of Expert classifier. The experts and the gate are implemented as fully connected layers with a sigmoid and a softmax nonlinearity, respectively. A confidence score for the cth class is then computed by merging experts' outputs into a single output $o_c(I)$ according to the gate's decision (Eq. (1)). The whole network is trained end-to-end.

$$o_c(I) = \sum_{i=1,2} \sigma(e^c_i(I)) * Softmax(g^c_i(I)) \tag{1}$$

IV. Experimental Results

A. Datasets

We evaluate our method on four publicly available datasets: TwitterI [7], TwitterII [4], EmotionROI [37] and FI [38]. The TwitterI dataset is the most widely used image sentiment classification dataset and has become the de facto benchmark. It contains 1269 images collected from Twitter and labeled in two categories, as either "positive" or "negative", by Amazon Mechanical Turk (AMT) workers. Each image was annotated by 5 people, creating three distinct subsets "5-agree", "4-agree" and "3-agree", depending on the number of agreements between the annotators. For our experiments we used the more reliable "5-agree" subset, that contains 882 images. TwitterII contains 603 images from Twitter, also annotated as "positive" or "negative" by AMT workers. EmotionROI contains 1980 images from Flickr, belonging to one of six emotion categories (anger, disgust, fear, joy, sadness, surprise), but since we are doing binary classification, images belonging to emotions joy and surprise are labeled as "positive" and images belonging to emotions anger, disgust, fear and sadness as "negative", as it is typically done in the literature works that experiment with this dataset. The FI dataset contains 23308 images from Flickr and Instagram. It is by far the largest of the four datasets. The images are annotated by AMT workers into 8 emotion categories (amusement, anger, awe, contentment, disgust, excitement, fear, sadness). We label as "positive" images that belong to emotions amusement, awe, contentment and excitement, and as "negative" the images that belong to emotions anger, disgust, fear and sadness. It should be noted that for all datasets more than one annotators were used for labeling the images, resulting in strongly labeled datasets.

B. Experimental Setup

In order to compare against the literature, we apply the same training-testing splitting criteria that are typically used for each dataset. For the TwitterI dataset, no pre-specified split to training/testing data exists, however it is common practice to do a 80-20 random split. We perform 5 random splits of 80% training and 20% testing sets and then report the average classification accuracy (ACC) on these splits. On TwitterII we use the dataset's pre-existing 5-fold split to do 5-fold cross validation and report again the average accuracy. On EmotionROI we use the pre-existing 80% training - 20% testing split. On FI, we followed the common practice of splitting randomly 80% for training, 15% for testing and 5% for validation. In our experiments we tested three main configurations:

- using a single feature type to train a neural network classifier. These include EfficientNet+NN, Signatures+NN, Resnet+NN, YT8M+NN and GCN+NN
- using the concatenation of all feature types to train a neural network classifier (combination+NN)
- using the concatenation of all feature types to train a mixture of experts classifier (combination+MoE)

Although more combinations of feature vectors are possible, for our experiments we chose only these three configurations for simplicity. In the first two configurations (single and combination+NN) we used the same parameters and setup after preliminary experimentation. In the second configuration (combination+NN) we applied sample-wise unit normalization for each feature type before the concatenation. The neural network (NN) classifier used was a simple 2-layer fully connected network with a 200-neuron first layer followed by a ReLU, a Dropout of 0.4 and a 2-neuron output layer plus a softmax for the 2 classes (positive, negative). We used binary cross-entropy as our loss function and Adam optimizer with 10^{-4} learning rate; set the batch size to 8, and trained for 20 epochs. For the third configuration (combination+MoE), after preliminary experimentation we ended up with a different set of parameters. Contrary to the first two configurations, we didn't apply any normalization before the concatenation of the features, as it didn't improve the results. We chose 10^{-3} weight decay, 256 batch size, binary cross-entropy loss function and Adam optimizer with 10^{-4} learning rate. For the third configuration we present 2 training schemas: (1) for 120 epochs with a 0.1 learning rate decay every 40 epochs, and (2) for 60 epochs with a 0.1 learning rate decay every 20 epochs. In every configuration we use the same parameters on all datasets; we do not do any dataset-specific parameters optimization.

C. Results

A summary of the scores of our experiments and the best literature methods can be seen in Table II. All reported scores refer to Accuracy (%). The highest score for a dataset is shown in bold, while the second-highest score is underlined. Firstly, we evaluate the models trained with the first configuration i.e. each of the examined feature vectors separately. On TwitterI, Resnet+NN performed the best. On TwitterII, Signatures+NN produced the best results. EmotionROI and FI were topped by GCN+NN and resnet+NN respectively. The models trained with the out-of-domain features, with the exception of YT8M+NN, performed on par with the in-domain ones. YT8M+NN generally underperformed, proving that semantic-level feature vectors are not so suitable for this task. The first configuration experiments were intended as an ablation study to verify the advantage of combining multiple feature types. Still, most of the models exceed the SOTA in TwitterI and one model (GCN+NN) the SOTA in EmotionROI

TABLE II. THE PERFORMANCE OF DIFFERENT METHODS ON THE FOUR DATASETS, IN ACCURACY (%). THE FIRST HALF OF THE TABLE REPORTS ON STATE-OF-THE-ART APPROACHES OF THE LITERATURE, WHILE THE SECOND HALF REPORTS OUR TESTED APPROACHES.

	Method	TwitterI 5-agree (ACC %)	TwitterII (ACC %)	EmotionROI (ACC %)	FI (ACC %)
Literature Methods	Wang et al., 2016 [8]	83.80	-	-	-
	Islam et al., 2016 [10]	86.10	-	-	-
	Campos et al., 2017 [9]	84.40	-	-	-
	Song et al., 2018 [12]	85.10	-	-	-
	Yang et al., 2018 [14]	84.25	81.35	-	-
	Yang et al., 2018 [15]	88.65	80.48	81.26	86.35
	He et al., 2019 [13]	-	82.40	-	-
	Yadav et al., 2019 [11]	86.40	-	-	-
	Wu et al., 2020 [16]	89.50	80.97	83.04	**88.84**
Proposed Approach	EfficientNet+NN	91.64	78.11	81.99	84.14
	Resnet+NN	93.11	80.18	82.66	87.29
	YT8M+NN	81.02	77.34	75.42	81.08
	Signatures+NN	91.30	80.74	81.65	86.65
	GCN+NN	90.62	79.59	84.18	87.06
	combination+NN	92.99	81.81	**84.68**	87.14
	combination+MoE(1)	**93.22**	80.93	84.34	87.81
	combination+MoE(2)	92.77	**82.99**	83.33	87.78

as well. The second configuration (combination+NN) is in general better than any of the first configuration models as it outperforms all of them on 3 out of 4 datasets. It exceeds the SOTA on 2 of the 4 datasets. This was expected, as the model is trained with a vastly richer feature vector. Finally, in the third configuration with the MoE, as it was mentioned earlier, we test 2 training schemas with 120 and 60 epochs respectively. Schema 1 (120 epochs) achieves the best score on TwitterI and FI. The second schema (60 epochs) is the only model that managed to exceed the SOTA on TwitterII and on 3 of the 4 datasets in total. Training for fewer epochs seems to help mitigate the overtraining on the smaller TwitterII dataset. On FI, our models perform slightly worse than the method of [16], which however performs worse than our models in the other 3 datasets. However it should be noted that for TwitterI and FI no pre-specified training-testing split exists, so strictly fair comparison with the literature is not possible. A sample of images from the EmotionROI dataset classified by our method are shown in Figure 3, where we include indicative examples of both correctly classified and misclassified images for both of the two classes.

V. CONCLUSIONS

In this work we focused on the problem of image sentiment analysis. Our motivation was twofold: First, to examine the performance gains of combining multiple feature types to train a sentiment classifier, and second, to evaluate a Mixture of Experts approach for this problem. For evaluating our method, we used 4 benchmark datasets. The results showed that, as was expected, the combination of features resulted in a richer image representation. The models trained on the combination of features were generally able to outperform those trained on a single type of features and outperformed the state-of-the-art on 2 of the 4 datasets. The proposed Mixture of Experts approach further improved the results, exceeding the SOTA performance on 3 of the 4 datasets.

ACKNOWLEDGMENT

This work was supported by the EU Horizon 2020 programme under grant agreement 832921 (MIRROR).

REFERENCES

[1] S. Hassan, K. Ahmad, A. Al-Fuqaha, and N. Conci, "Sentiment analysis from images of natural disasters," in *Image Analysis and Processing – Proc. ICIAP*, Cham, 2019, pp. 104–113, Springer International Publishing.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, et al., "Imagenet large scale visual recognition challenge," *Int. journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[3] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. of the 18th ACM Int. Conf. on Multimedia*, New York, NY, USA, 2010, MM '10, p. 83–92.

[4] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. of the 21st ACM Int. Conf. on Multimedia*, New York, NY, USA, 2013, MM '13, p. 223–232.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. of the 25th Int. Conf. on Neural Information Processing Systems - Volume 1*, Red Hook, NY, USA, 2012, NIPS'12, p. 1097–1105.

[6] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks," *CoRR*, vol. abs/1410.8586, 2014.

[7] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. of the AAAI conf. on Artificial Intelligence*, 2015, vol. 29.

[8] J. Wang, J. Fu, Y. Xu, and T. Mei, "Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks," in *Proc. of the Twenty-Fifth Int. Joint Conf. on Artificial Intelligence*. 2016, IJCAI'16, p. 3484–3490, AAAI Press.

[9] V. Campos, B. Jou, and X. Giró-i-Nieto, "From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction," *Image and Vision Computing*, vol. 65, pp. 15–22, 2017.

[10] J. Islam and Y. Zhang, "Visual sentiment analysis for social images using transfer learning approach," in *IEEE Int. Conf. on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)*. IEEE, 2016, pp. 124–130.

[11] A. Yadav, A. Agarwal, and D. K. Vishwakarma, "Xra-net framework for visual sentiments analysis," in *IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 2019, pp. 219–224.

[12] K. Song, T. Yao, Q. Ling, and T. Mei, "Boosting image sentiment analysis with visual attention," *Neurocomputing*, vol. 312, pp. 218 – 228, 06 2018.

[13] X. He, H. Zhang, N. Li, L. Feng, and F. Zheng, "A multi-attentive pyramidal model for visual sentiment analysis," in *Int. Joint Conf. on Neural Networks (IJCNN)*, 2019, pp. 1–8.

[14] J. Yang, D. She, Y. Lai, P. L. Rosin, and M. Yang, "Weakly supervised

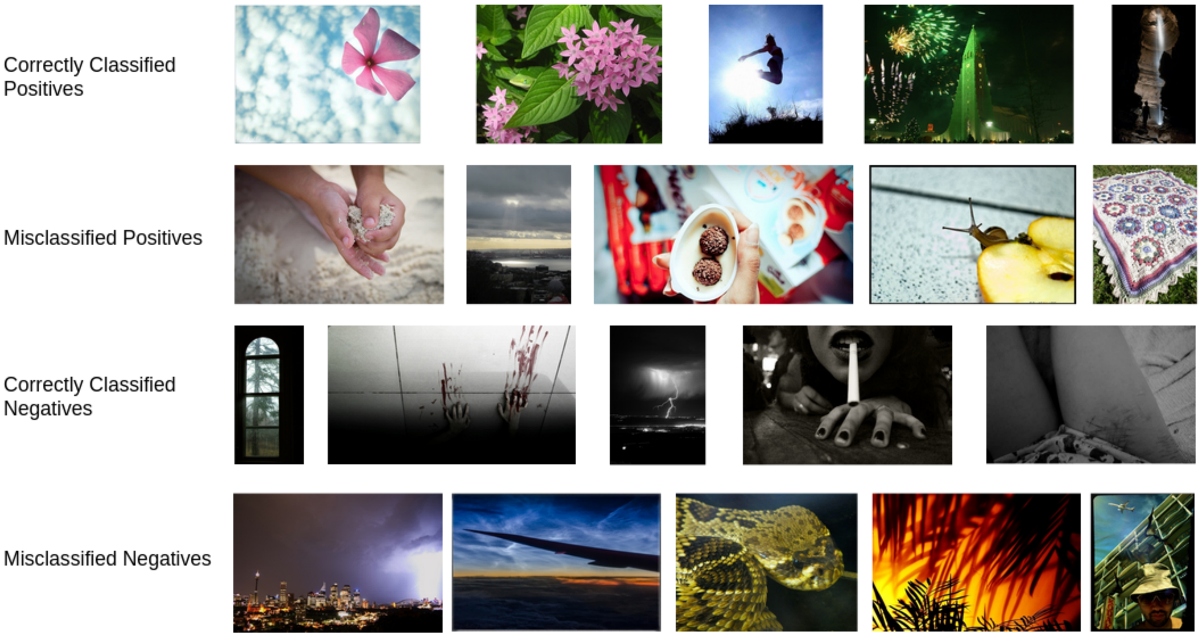
Correctly Classified Positives
Misclassified Positives
Correctly Classified Negatives
Misclassified Negatives

Fig. 3. A sample of images from the EmotionROI dataset, grouped by their classification result.

coupled networks for visual sentiment analysis," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 7584–7592.

[15] J. Yang, D. She, et al., "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Trans. on Multimedia*, vol. 20, no. 9, pp. 2513–2525, Sept. 2018.

[16] L. Wu, M. Qi, M. Jian, and H. Zhang, "Visual sentiment analysis by combining global and local information," *Neural Processing Letters*, vol. 51, pp. 1–13, 06 2020.

[17] Y. Liang, K. Maeda, T. Ogawa, and M. Haseyama, "Cross-domain semi-supervised deep metric learning for image sentiment analysis," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4150–4154.

[18] J. Zhang, M. Chen, H. Sun, D. Li, and Z. Wang, "Object semantics sentiment correlation analysis enhanced image sentiment classification," *Knowledge-Based Systems*, vol. 191, 11 2019.

[19] A. Ortis, G. Farinella, G. Torrisi, and S. Battiato, "Exploiting objective text description of images for visual sentiment analysis," *Multimedia Tools and Applications*, 01 2020.

[20] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton, "Adaptive mixture of local expert," *Neural Computation*, vol. 3, pp. 78–88, 02 1991.

[21] K. Ahmed, M. Baig, and L. Torresani, "Network of experts for large-scale image categorization," in *European Conference on Computer Vision*. 2016, p. 516–532, Springer.

[22] X. Wang, F. Yu, L. Dunlap, Y. Ma, R. Wang, A. Mirhoseini, T. Darrell, and J. Gonzalez, "Deep mixture of experts via shallow embedding," 2019.

[23] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Int. Conf. on Machine Learning*, 2019, pp. 6105–6114.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[25] S. Abu-El-Haija, N. Kothari, J. Lee, A. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," in *arXiv:1609.08675*, 2016.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[27] D. Galanopoulos and V. Mezaris, "Attention mechanisms, signal encodings and fusion strategies for improved ad-hoc video search with dual encoding networks," in *Proc. of the ACM Int. Conf. on Multimedia Retrieval*. ACM, 2020, (ICMR '20).

[28] K. Cho, B. van Merriënboer, C. Gulcehre, et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[29] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751.

[30] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. of IEEE CVPR*, 2016, pp. 5288–5296.

[31] Y. Li, Y. Song, et al., "TGIF: A new dataset and benchmark on animated gif description," in *Proc. of IEEE CVPR*, 2016, pp. 4641–4650.

[32] X. Wang, J. Wu, et al., "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2019, pp. 4581–4591.

[33] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Int. Conf. on Computer Vision (ICCV)*, 2017.

[34] N. Gkalelis, A. Goulas, D. Galanopoulos, and V. Mezaris, "Objectgraphs: Using objects and a graph convolutional network for the bottom-up recognition and explanation of events in video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 3375–3383.

[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. 2015, vol. 28, Curran Associates, Inc.

[36] Y. Jiang, Z. Wu, J. Wang, X. Xue, and S. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 352–364, 2018.

[37] K. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "Where do emotions come from? predicting the emotion stimuli map," in *IEEE Int. Conf. on Image Processing (ICIP)*, 2016, pp. 614–618.

[38] 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 Conf. Artif. Intell.*, 2016.