Combining Multiple Deep-learning-based Image Features for Visual Sentiment Analysis

Alexandros Pournaras, Nikolaos Gkalelis, Damianos Galanopoulos, Vasileios Mezaris
CERTH-ITI, Greece
{apournaras,gkalelis,dgalanop,bmezaris}@iti.gr

ABSTRACT

This paper presents our team's (IDT-ITI-CERTH) proposed method for the Visual Sentiment Analysis task of the Mediaeval 2021 benchmarking activity. Visual sentiment analysis is a challenging task as it involves a high level of subjectivity. 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 our approach we examine the potential of transferring knowledge from several pre-trained networks, some of which are out-of-domain. We concatenate these diverse feature vectors and construct an image representation that is used to train a classifier for each of the three subtasks of this Mediaeval task. Due to a bug in the original submission file, the official scores we got are 0.595, 0.479 and 0.380 for subtasks 1,2 and 3 respectively.

1 INTRODUCTION

Visual sentiment analysis is the problem of identifying the sentiment conveyed by an image. The problem has recently attracted significant attention due to the large-scale use of images in social media. This Mediaeval task focuses on images from natural disasters, thus content that can often induce strongly negative sentiments. A human-labeled disaster-related dataset as well as a deep-learning based approach to solve it was proposed in [17]. In [5] a detailed description of the task is presented.

In general, visual sentiment analysis is challenging because it involves a higher level of human subjectivity in the classification process, compared to other image classification tasks. Similarly to such tasks, deep convolutional neural networks are widely used; many literature works, e.g. [2] [7], rely on transfer learning by performing fine-tuning on pre-trained networks that most commonly have been originally trained on ImageNet [12]. However, little emphasis has been given to 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 image representation. Using this, we train a sentiments classifier that takes the form of either a dense 3-layer neural network or a Mixture of Experts (MoE) [8] classifier.

MediaEval'21, December 13-15 2021, Online

2 APPROACH

Our proposed method is closely based on [11], a method that achieves state-of-the-art performance in many benchmark image sentiment analysis datasets. We are transferring knowledge from 5 trained 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 task because they perform very well in their respective domains. A feature vector is extracted from each network. In the following subsections, 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 1.

2.1 In-Domain Feature Vectors

2.1.1 EfficientNet features. EfficientNet [14] is a recently proposed deep convolutional neural network architecture that achieves state-of-the-art performance on image classification tasks. We used a "B2"-variation 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, **E**.

2.1.2 Resnet features. Resnet [6] 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 [12] and extract the 2048-element "pool5" layer as the feature vector, **R**.

2.2 Out-of-Domain Feature Vectors

2.2.1 YT8M features. YouTube-8M [1] is a large annotated video dataset containing approximately 6 million videos of a total duration of more than 500.000 hours and labeled with 3862 classes. For training a classifier on this dataset, we extract features at a 1fps sampling rate using an Inception neural network [13] pre-trained on Imagenet [12]. The ReLU activation of the last hidden layer of this network is given as input to a rather simple CNN classifier, consisting of a 1D convolutional layer with 64 filters, a max-pooling layer, a dropout and a Sigmoid of 3862 outputs. This is the YouTube-8M-trained classifier that we ultimately use as feature extractor for sentiment classification: the classifier's 3862-element output vector for each image is our feature vector, **Y**.

2.2.2 "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 [3] is used. The network is trained to translate a media

Table 1: Summary of employed deep-learning-based image features.

Feature name	Base network architecture	Training datasets	Original task
EfficientNet (**E**)	EfficientNet-B2	Imagenet 1k concepts	image classification
Resnet (**R**)	Resnet-152	Imagenet 11k concepts	image classification
YT8M (**Y**)	Inception	Youtube8M	video classification
Signature (**S**)	Dual encoding network	MSR-VTT, TGIF, Vatex, ActivityNet	ad-hoc video search
GCN (**G**)	Resnet-152, Faster R-CNN, GCN	ImageNet 1k, FCVID, YLI-MED	video event recognition

item (i.e. an entire video) **V** or a textual item (i.e. a natural-language video caption or search query) **T** into a new joint 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. This network is trained using large datasets of video-caption pairs: MSR-VTT [16], TGIF [10], Vatex [15] and ActivityNet [9]. For leveraging this pre-trained network as a feature generator in the image sentiment analysis 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**.

2.2.3 Graph Convolutional Network (GCN) features. To obtain this feature vector, we employ a neural network trained for the task of video event recognition [4]. Following the application of an object detector on the frames of the video, a neural network is used to extract the objects' 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. 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**.

2.3 Sentiment Classifiers

We concatenate the 5 feature vectors described above, resulting in a final 11414-element feature vector, that will be used to train our classifiers for the 3 subtasks.

2.3.1 Subtask 1. For subtask 1 we employ a Mixture of Experts classifier. The first layer of this classifier is a fully connected layer which transforms the input vector to 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, $e_1^c(), e_2^c()$, which are defined for each class c, as well as to the associated gates, $g_1^c(), g_2^c()$. For each class, an extra "dummy" expert is also defined to represent 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.

Table 2: The results of our method (F1 weighted score) for the 3 subtasks.

Subtask	dev set	test set (with bug)	test set (corrected)
Subtask 1	0.757	0.595	0.740
Subtask 2	0.612	0.479	0.604
Subtask 3	0.510	0.380	0.510

$$o_c(I) = \sum_{i=1,2} \sigma(e_i^c(I)) * Softmax(g_i^c(I)) \quad (1)$$

2.3.2 Subtask 2. For subtask 2 we employ a dense 3-layer neural network classifier. This classifier comprises three fully-connected layers with 1000, 200 and, finally, 7 neurons, as there are 7 target classes in this subtask. Between consecutive layers there is a ReLU and a dropout layer with 0.4 probability. Finally the output passes through a sigmoid nonlinearity.

2.3.3 Subtask 3. For subtask 3 we use the same classifier used in subtask 2, with the exception of the output layer, which in this case comprises 10 neurons, as there are 10 target classes.

3 RESULTS

For optimizing the parameters and choosing the classifiers for each subtask, we randomly split the development set to a training and a validation set with 80% and 20% of the images respectively. For subtask 1, we measured the cross-entropy loss. We employed the Adam optimizer and trained with 10^{-5} learning rate. We trained the model for 300 epochs. For subtasks 2 and 3 we additionally performed augmentations to the images of the training set: random crop, blurring, change in brightness and random rotations. We measured the binary cross-entropy loss and optimized with Adam. For subtask 2 we trained for 300 epochs with $3x10^{-6}$ learning rate, while for subtask 3 for 200 epochs with $5x10^{-6}$ learning rate. The learning rate is scheduled to drop by half every 70 epochs in all the subtasks. The batch size for the training was set to 64 for all the subtasks. Following the selection of the above parameters, we used the entire development set provided by the task organizers to train our final models. The experimental results we got on the development set as well as the official (with bug) and unofficial (corrected) test set results are shown in Table 2.

ACKNOWLEDGMENTS

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

REFERENCES

[1] S. Abu-El-Haija, N. Kothari, J. Lee, A. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. In *arXiv:1609.08675*. https://arxiv.org/pdf/1609.08675v1.pdf

[2] V. Campos, B. Jou, and X. Giró-i-Nieto. 2017. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. *Image and Vision Computing* 65 (2017), 15–22.

[3] D. Galanopoulos and V. Mezaris. 2020. 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 ((ICMR '20))*. ACM.

[4] N. Gkalelis, A. Goulas, D. Galanopoulos, and V. Mezaris. 2021. 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*. 3375–3383.

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

[6] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 770–778.

[7] J. Islam and Y. Zhang. 2016. 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, 124–130.

[8] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton. 1991. Adaptive Mixture of Local Expert. *Neural Computation* 3 (02 1991), 78–88. https://doi.org/10.1162/neco.1991.3.1.79

[9] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. 2017. Dense-Captioning Events in Videos. In *Int. Conf. on Computer Vision (ICCV)*.

[10] Y. Li, Y. Song, and others. 2016. TGIF: A new dataset and benchmark on animated GIF description. In *Proc. of IEEE CVPR*. 4641–4650.

[11] A. Pournaras, N. Gkalelis, D. Galanopoulos, and V. Mezaris. 2021. Exploiting Out-of-Domain Datasets and Visual Representations for Image Sentiment Classification. In *2021 16th International Workshop on Semantic and Social Media Adaptation Personalization (SMAP)*. 1–6. https://doi.org/10.1109/SMAP53521.2021.9610801

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, and others. 2015. Imagenet large scale visual recognition challenge. *Int. journal of computer vision* 115, 3 (2015), 211–252.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[14] M. Tan and Q. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Int. Conf. on Machine Learning*. 6105–6114.

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

[16] J. Xu, T. Mei, T. Yao, and Y. Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Proc. of IEEE CVPR*. 5288–5296.

[17] S. Zohaib, K. Ahmad, N. Conci, and A. Al-Fuqaha. 2019. Sentiment Analysis from Images of Natural Disasters. (2019). arXiv:cs.CV/1910.04416