On recognizing emotion of sadness in images of a general nature using CNN

Modestas Motiejauskas, Gintautas Dzemyda

Institute of Data Science and Digital Technologies, Vilnius University

Sadness emotion recognition in images of general nature

Sadness emotion recognition in images of general nature is being constructed as binary classification problem – answering whether image expresses sadness emotion. We chose a convolutional neural network as the mean for such classification.

Convolutional neural networks need for a large sets of images for training. WEBEmo [1] may serve as such a set. WEBEmo dataset contains about 268000 images. It is a large scale weakly-labeled image emotion dataset for possible training of convolutional neural networks. This dataset contains images of general nature, however part of images have some text. Text may carry some emotion. In our case, we should to discard these mentioned images. We have downloaded a part of WEBEmo dataset, 220854 images. We have selected randomly 18520 images with known classes: 8549 images which are labeled as having visual sadness emotion and remaining 9971 images which do not have sadness emotion. From the initial 18520 image dataset we discarded images, having textual information, and obtained 14901 total filtered images: 6697 images which are labeled as having visual sadness emotion and remaining 8204 images which do not have sadness emotion.

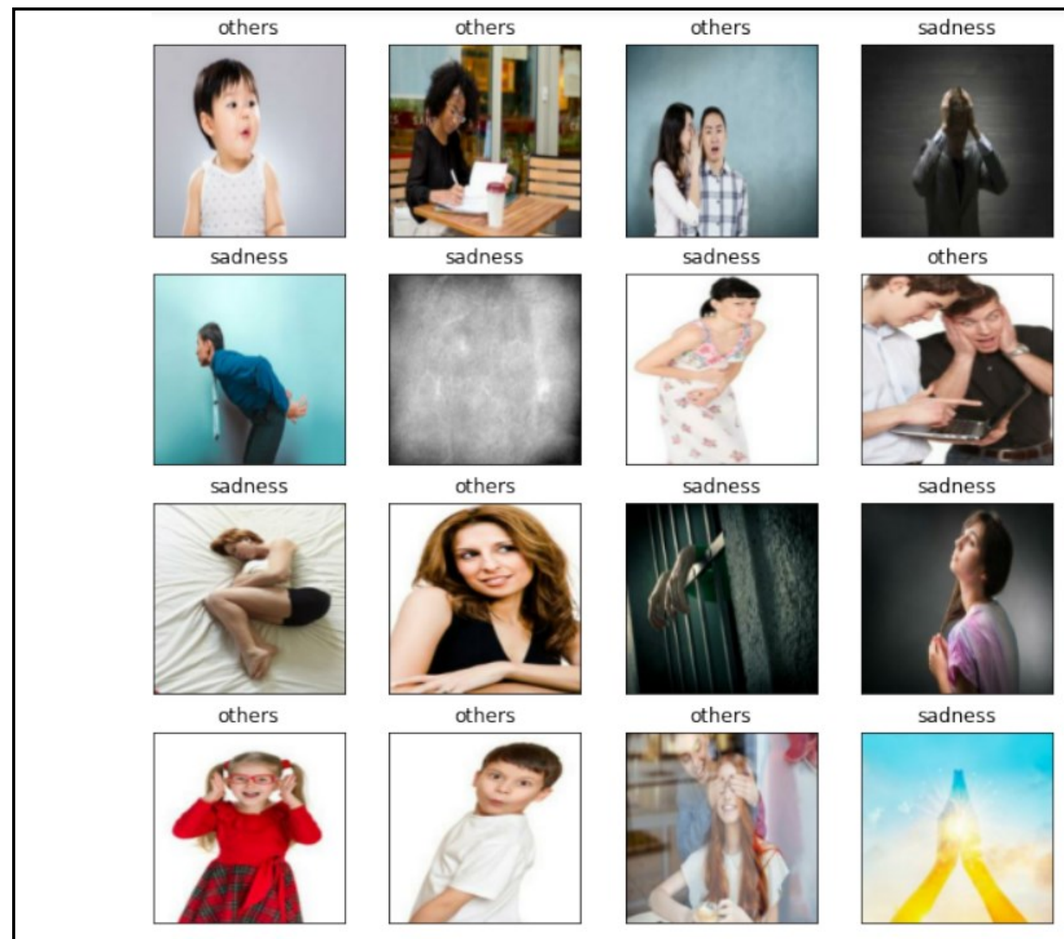
For discarding images containing textual meaning, we have trained a special classifier that may answer whether images have texts inside them. For this filtering out unnecessary content we have trained EfficientNetV2B0 [4] model with images from several datasets. For training of EfficientNetV2B0 we used 6627 images with some text inside from ICDAR2017 robust reading challenge on coco-text [7] and 9720 general purpose images, which do not have textual information from RGP [3]. Trained model achieved 99.31 % overall accuracy. One of the newest convolutional neural network family EfficientNetV2 [4] was published in 2021. According to the authors it provided the best results on the ImageNet [5] dataset classification task. This ImageNet ILSVRC2012 dataset contains 1,281,167 training images, 50,000 validation images and 100,000 test images and classifier was trained to classify 1000 classes from mentioned dataset. This EfficientNetV2 family of models outperforms previous models by introducing more efficient blocks named MBConv and Fused-MBConv. Authors apply search space method based upon their older EfficientNetV1 backbone and they obtain model named EfficientNetV2-S [5]. EfficientNetV2B0 and EfficientNetV2B2 are scaled down versions of the EfficientNetV2S backbone and are usually trained on the smaller image sizes, thus they have less number of parameters. In our case, we took such a pre-trained network that recognizes 1000 objects as an initial state for further additional its training.

Another chosen model for our analysis is named Xception [2]. They presented a step in-between regular convolution and the depthwise separable convolution operation and at that time outperformed InceptionV3 [6] model. Their main introduction is a convolution layer named as depthwise convolution. It is a spatial convolution performed independently over each channel of an input, following by a pointwise convolution - applying 1x1 convolution filter. This whole structure is commonly known as Inception module. Authors improvement comes from the different order of operations of mentioned Inception module and the usage of non-linearity after the first operation. We used the pre-trained network using the ImageNet [5] dataset. Both EfficientNetV2 and Xception were pre-trained on the same data and for recognition of the same objects.

Training process of the chosen networks was carried out as follows: 14901 total filtered images dataset was split into 70 % subset for training, 15 % subset for validation and remaining 15 % for testing. Adagrad optimizer for training was used with 0.001 learning rate parameter. Loss function for evaluating models was sparse categorical cross entropy, input images were provided as 150x150 colored images.

Results

The results of training and classification are presented in Tables [1-3]. We see that EfficientNetV2 [4] networks outperformed Xception [2].



In Table 1 results are presented from the previously mentioned filtered out WEBemo dataset testing subset. We evaluated overall model accuracy and F1 scores, separately we obtained classes named sadness and others F1 scores. We conducted another test without using pre-trained models from ImageNet [5] to determine, whether using pre-trained weights produce better results to our specific task.

Table 1: Trained models comparison results.

Model	Overall accuracy (%)	Sadness F1	Others F1	Overall F1
Xception	72.53	0.70	0.75	0.73
EfficientNetV2B0	74.14	0.72	0.76	0.74
EfficientNetV2B2	74.18	0.71	0.77	0.74
EfficientNetV2S	75.57	0.74	0.77	0.76
Xception no-pretrain	63.53	0.62	0.65	0.64

Table 2: Trained models comparison results. Testing: UnbiasedEmo dataset.

Model	Overall accuracy (%)	Sadness F1	Others F1	Overall F1
Xception	73.29	0.71	0.75	0.73
EfficientNetV2B0	73.54	0.72	0.75	0.74
EfficientNetV2B2	75.00	0.73	0.77	0.75
EfficientNetV2S	74.88	0.75	0.75	0.75

Table 3: Trained models comparison results. Testing: Emotion-6 dataset.

Model	Overall accuracy (%)	Sadness F1	Others F1	Overall F1
Xception	69.10	0.62	0.74	0.70
EfficientNetV2B0	65.52	0.61	0.69	0.66
EfficientNetV2B2	68.88	0.62	0.74	0.70
EfficientNetV2S	68.95	0.65	0.72	0.70

Conclusions

Conducted study in terms of classifying sadness emotion allows us to further understand domain knowledge of the topic better. It is possible expand to more emotion categories conducting same experimental approach. Primary experimental tests also shows challenges common to convolutional neural networks — overfitting, appropriate learning optimizer selection. Tested data on the trained models shows us higher F1-score towards images labelled as others.

References

1. Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K. Roy-Chowdhury. Contemplating visual emotions: Understanding and overcoming dataset bias. 2018.
2. François Chollet. Xception: Deep learning with depthwise separable convolutions. 2016
3. Prasun Roy, Subhankar Ghosh, Saumik Bhattacharya, and Umapada Pal. Effects of degradations on deep neural network architectures. 2018.
4. Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training. 2021.
5. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3): 211–252, 2015.
6. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818-2826, doi: 10.1109/CVPR.2016.308.
7. R. Gomez et al., "ICDAR2017 Robust Reading Challenge on COCO-Text," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, pp. 1435-1443, doi: 10.1109/ICDAR.2017.234.