Inhaltsverzeichnis

Sitzung ABC

Recurrent Spatial Attention for Facial Emotion Recognition

Valentin Forch[1], Julien Vitay[1], and Fred H. Hamker[1]

Chemnitz University of Technology, Professorship Artificial Intelligence,
Straße der Nationen 62, 09111 Chemnitz

Abstract Automatic processing of emotion information through deep neural networks (DNN) can have great benefits for human-machine interaction. Vice versa, machine learning can profit from concepts known from human information processing (e.g., visual attention). We employed a recurrent DNN incorporating a spatial attention mechanism for facial emotion recognition (FER) and compared the output of the network with results from human experiments. The attention mechanism enabled the network to select relevant face regions to achieve state-of-the-art performance on a FER database containing images from realistic settings. A visual search strategy showing some similarities with human saccading behavior emerged when the model's perceptive capabilities were restricted. However, the model then failed to form a useful scene representation.

Keywords emotion recognition, attention, LSTM

1 Introduction

As emotions are deeply rooted within psychological and physiological processes, they inform us about personality traits, intentions, physiological states, or important events [1, 2]. Hence, there is a strong interest in the automatic processing of emotion information in the fields of human-machine interaction and machine learning.
For automatic facial emotion recognition (FER), deep neural networks (DNN) have become the preferred approach [3]. As FER requires the detection of subtle changes in facial structure [4, 5] spatial attention might

Figure 1.1: Maps of relevant visual information for facial emotion recognition (adapted from [10]).

improve classification performance, as it allows to select relevant segments of a visual scene [6]. It also has the benefit of reducing the computational cost of image processing, as the whole image does not need to be processed [7].
Eye-tracking studies show that humans preferably attend to the eye and mouth regions when classifying facial expressions [8,9]. Moreover, Blais et al. [10] showed that depending on the displayed emotion different portions of the face are relevant for FER (Fig. 1).
Located in the spectrum of human and artificial information processing, the aim of the present work is twofold: (a) to employ a recurrent DNN incorporating a spatial attention mechanism for FER and (b) to compare the output of the network with results from human experiments on FER (i.e., Blais et al. [10]).

2 Related Work

Generally, attention can be understood as a guiding signal for the processing of relevant information [6]. In humans the main function of visual attention is to guide the gaze towards relevant parts of visual scenes [11]. Mnih et al. [7] took this as inspiration for their recurrent attention model for image processing which perceives only a small part of a visual scene through a so-called glimpse sensor. The sensor's location is controlled by a recurrent network thus extracting information from the image for a fixed number of iterations.
Based on the further developed deep recurrent attention model [12], Ablavatsky et al. [13] proposed their enriched deep recurrent attention model (EDRAM) featuring the spatial transformer [14] as the new glimpse sensor ("Attention Mechanism", Fig. 2). This layer extracts a scaled and rotated patch from an input array by performing an affine

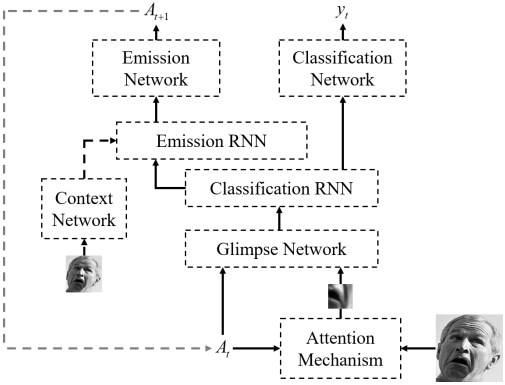

Figure 1.2: EDRAM overview (adapted from [13]).

transformation on a set of grid points defining sampling positions in the input. The number of grid points is equal to the resolution of the resulting output. The transformation is controlled by the fully connected emission network which generates the transformation matrix A.

The glimpse network receives the image patch and extracts a feature vector through a convolutional neural network. This vector is multiplied elementwise with the output of a dense layer of the same dimensionality receiving the transformation matrix as an input thus merging "what" and "where" information [15]. This information is passed to two sequential recurrent networks (RNN) which accumulate the glimpse information thus building up a scene representation. The first RNN's output is used for classification and feeds into the second RNN, which is connected to the emission network. Its hidden state is initialized by the context network – a small CNN using a heavily down-sampled version of the whole input image.

3 Model Training

The model was trained on the AffectNet database [16]. We used examples of the six categories of basic emotions (Happy, Sad, Surprise, Fear, Disgust, and Anger) and the neutral category. As most databases for FER, AffectNet has a strong class imbalance. The predefined validation

set however is balanced with respect to the classes. To counteract the training class imbalance, examples of the i-th of $n_{classes}$ containing N_i examples were weighted by a factor $w_i = \sqrt{\frac{N_{total}}{N_i \times n_{classes}}}$ when computing the classification loss.
In addition to categorial information, AffectNet includes continuous ratings of valence and arousal of each facial expression. To use this information a second fully connected "classification" network was added to the model. Instead of softmax classification the final layer of this network had two units with hyperbolic tangent activation functions.
For preprocessing, images were cropped based on bounding boxes generated by a DNN face detector from the open CV package[1] or the frontal face detector from the Dlib library [17] if the first detector failed, downscaled to 100x100 grayscale, and contrast limited histogram equalization was applied [18]. Data augmentation was applied online.

4 Experiments

We implemented EDRAM within the Tensorflow/Keras framework.[2] The parameters for the first glimpse of an input were based on the initial state of the second recurrent network as in Ba et al. [12]. While Ablavatski et al. [13] used batch normalization layers [19] with shared weights between model iterations our model used unique batch normalization layers for each timestep, as the pattern of layer activations was expected to vary with each timestep. Furthermore, we computed the loss for the emitted transformation matrix only for the zoom parameter thus enabling our model to freely choose the location of the glimpse sensor. The target zoom factor was set to .35. All other model specifications were initially the same as in [13].

The model achieved a mean classification accuracy of 60.4% (recent work of Li et al. [20] achieved 58.8% on the same data). The happy category had an accuracy of 86.7% while all other class accuracies were 50–60%. The visual search strategy of the model was characterized by a uniform application of the glimpse sensor. The model first used a whole-face glimpse and then zoomed in on the left eye region (Fig. 3).

[1] `github.com/opencv/opencv/tree/master/samples/dnn/face_detector`
[2] The spatial transformer implementation was taken from `github.com/oarriaga/STN.keras`

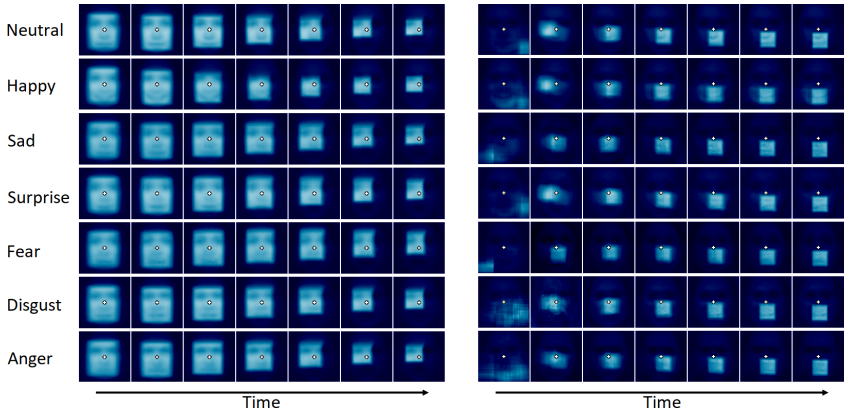

Figure 1.3: Heat maps of glimpse areas. Unrestricted model (left) and model with its zoom factor restricted to .30 (right).

Furthermore, we limited the zoom factor so that the network could only generate glimpses which covered 50% respectively 30% of the image. On average both models then only produced glimpses with their maximum zoom factor. The 50% model performed relatively good (57.8% accuracy). This model shifted its glimpse sensor from the lower left of the face to the upper middle (not shown). Interestingly, the positions of the first glimpses produced by the 30% model showed a high inter and intra-category variance (Fig. 3). Inspection of sample classifications showed that this model produced saccade-like jumps of its glimpse sensor instead of gradually zooming in on a face area or gradually shifting the glimpse position (Fig. 4). The performance of this model however was considerably worse than of other models (41.8% accuracy).

5 Conclusion

We showed that a DNN with a recurrent attention mechanism can achieve state-of-the-art performance in FER. The experiments with restricted zoom factors showed that a good classification performance can be achieved by shifting glimpses with a medium size across the image. The model which was restricted to relatively small glimpses

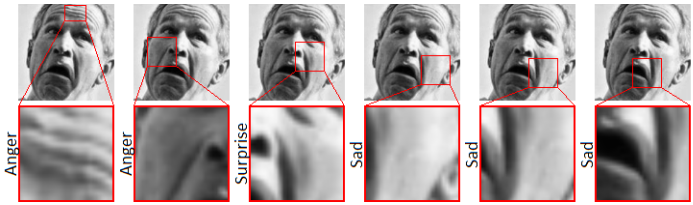

Figure 1.4: Emotion classification sample. Labels represent model predictions.

showed a saccade-like search pattern. However, this model performed considerably worse, indicating that it failed to form a scene representation from the separate glimpses.
Second, it was of interest whether the model would make use of emotion-specific search strategies. The better-performing models showed a relatively uniform search strategy. However, they also processed relatively large regions of the face at once. The model whose glimpse sensor was restricted showed emotion-specific search patterns with a preference for the mouth region as can be found in the data of Blais et al. [10].

This work was partly supported by the European Social Fund at the Free State of Saxony (Grant ESF-100269974).

References

1. P. Ekman, "Are there basic emotions?" 1992.
2. D. Keltner, J. Tracy, D. A. Sauter, D. C. Cordaro, and G. McNeil, "Expression of emotion," in *Handbook of Emotions*. New York: Guilford Publications, 2016, pp. 467–482.
3. S. Li and W. Deng, "Deep facial expression recognition: A survey," *arXiv preprint arXiv:1804.08348*, 2018.
4. Z. Ambadar, J. W. Schooler, and J. F. Cohn, "Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions," *Psychological science*, vol. 16, no. 5, pp. 403–410, 2005.
5. P. Ekman and W. V. Friesen, "Felt, false, and miserable smiles," *Journal of nonverbal behavior*, vol. 6, no. 4, pp. 238–252, 1982.

6. M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature reviews neuroscience*, vol. 3, no. 3, p. 201, 2002.

7. V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

8. O. Beaudry, A. Roy-Charland, M. Perron, I. Cormier, and R. Tapp, "Featural processing in recognition of emotional facial expressions," *Cognition & emotion*, vol. 28, no. 3, pp. 416–432, 2014.

9. H. Eisenbarth and G. W. Alpers, "Happy mouth and sad eyes: scanning emotional facial expressions." *Emotion*, vol. 11, no. 4, p. 860, 2011.

10. C. Blais, D. Fiset, C. Roy, C. Saumure Régimbald, and F. Gosselin, "Eye fixation patterns for categorizing static and dynamic facial expressions." *Emotion*, vol. 17, no. 7, p. 1107, 2017.

11. L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, p. 194, 2001.

12. J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.

13. A. Ablavatski, S. Lu, and J. Cai, "Enriched deep recurrent visual attention model for multiple object recognition," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2017, pp. 971–978.

14. M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

15. H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order boltzmann machine," in *Advances in neural information processing systems*, 2010, pp. 1243–1251.

16. A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *arXiv preprint arXiv:1708.03985*, 2017.

17. D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

18. S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987.

19. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

20. Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-gated cnn for occlusion-aware facial expression recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2209–2214.