Investigating the emergence of expression representations in a neural network trained to discriminate identities

Emily Schwartz[1], Kathryn O'Nell[2], and Stefano Anzellotti[1]
[1]Department of Psychology and Neuroscience, Boston College, [2]Department of Experimental Psychology, University of Oxford

Introduction

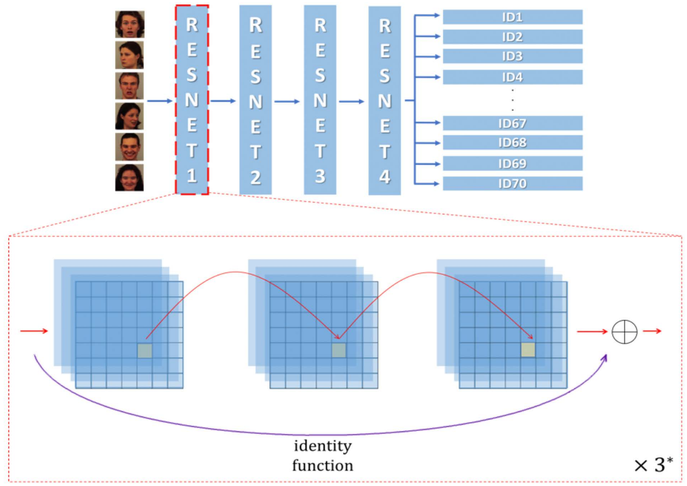
- Face identity and facial expressions are important cues to navigate the social world. According to a traditional account, identity and expressions are processed by separate pathways. Recent evidence has challenged this view: face identity can be decoded from responses in regions previously implicated in expression recognition (pSTS[1,2]); facial expressions can be decoded from ventral temporal regions[3].
- We hypothesize that joint processing of identity and expression is driven by computational efficiency. Recognition of identity and expression might be "complementary" and benefit from each other.
- Our lab has found evidence supporting this: artificial neural networks (ANNs) trained to recognize expressions spontaneously learn features that support identity recognition[4].
- Instead of extracting one property (i.e. identity) and discarding information about other properties (i.e. expression), information about different properties can be extracted by the same mechanism.

We investigate transfer learning in the reverse direction: Can ANNs trained to distinguish between identities learn features that support recognition of facial expressions?

Methods

Network architecture:

- ResNet-50[6] pre-trained with VGGFace2[7] database to perform identity recognition.
- All layers except the fully connected (FC) linear layer used ReLU as the activation function. The net was trained to minimize the cross-entropy loss.
- The network was tested using the KDEF[8] dataset.



Face stimuli:

VGGFace2

KDEF

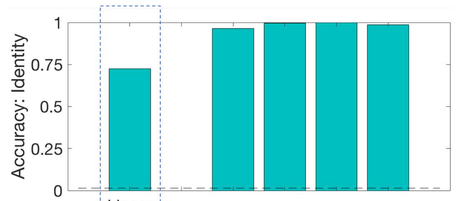
Face recognition accuracy

- We tested the net's ability to generalize to the KDEF dataset.
- The last layer was removed and a FC linear layer was attached to generate labels for KDEF.
- The pre-trained network was able to perform face identification on the KDEF dataset with an accuracy of **98.6%**.

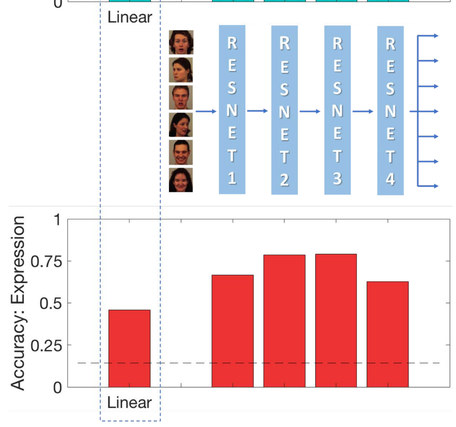
How do the features perform for expression recognition?

Expression recognition accuracy

- Again, an FC layer was attached to the network and the net was re-trained, keeping the weights for the other layers fixed. The nonlinear components of the net fully relied on the identity-based training.



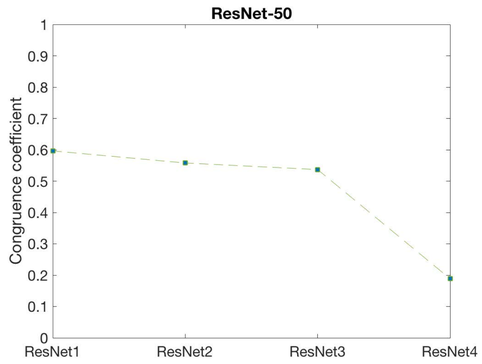
- The net was able to label expression above chance (14.2%) with an accuracy of **62.6%**
- Features from 4 hidden layers were extracted to test for expression (and identity) recognition as the features move through the network.

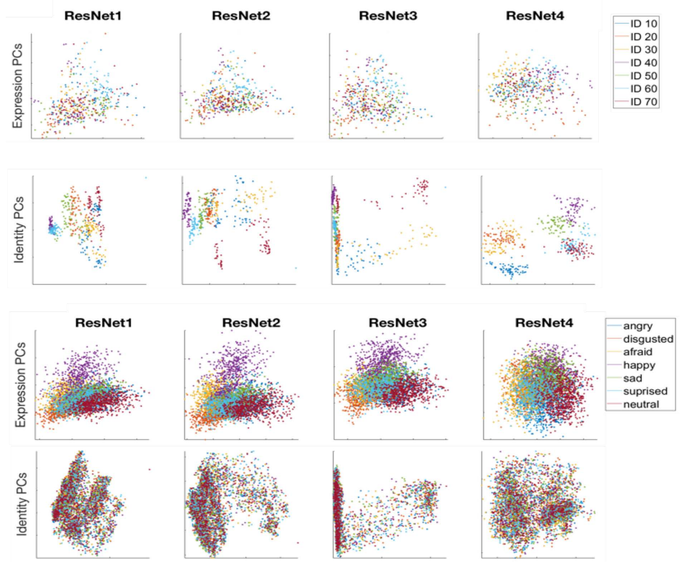


Is the increase in expression accuracy because features are overlapping or disentangled?

Overlap between identity and expression features

- Factor comparison was conducted by performing PCA and calculating the congruence coefficient from the top 5 principal components (PCs) for identity and expression.
- The PCs for identity and expression exhibited higher congruence values in earlier layers, and congruence decreased from layer to layer.





Discussion

- These results show that deep networks trained to recognize identity spontaneously develop representations that support expression recognition. This work has demonstrated transfer learning in the opposite direction[4]. These findings provide a proof of concept of the complementarity between identity and expression.
- We propose that this "complementarity" underlies the empirical observation of identity information in brain regions previously implicated in expression recognition, and vice versa.
- Deep networks trained to recognize identity might yield good transfer to expressions because they need to separate identity from expression, leading to disentangled expression representations as a byproduct. Analyses in the opposite direction (training on expression and testing on identity) has also provided support for this hypothesis.

Ongoing and future directions

- Can these deep network models accurately predict neural responses to face images? We are in the process of investigating this question.

References

1. Anzellotti, S., & Caramazza, A. (2017). Multimodal representations of person identity individuated with fMRI. Cortex, 89, 85-97.
2. Dobs, K., Schultz, J., Bülthoff, I., & Gardner, J. (2018). Task-dependent enhancement of facial expression and identity representations in human cortex. NeuroImage, 172, 689-702.
3. Skerry, A., & Saxe, R. (2014). A common neural code for perceived and inferred emotion. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 34(48), 15997-16008.
4. Kathryn C O'Nell, Rebecca Saxe, Stefano Anzellotti; Deep networks trained to recognize facial expressions spontaneously develop representations of face identity. Journal of Vision 2019;19(10):262.
5. G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07- 49, University of Massachusetts, Amherst, October 2007.
6. Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 770-778.
7. Cao, Q., Shen, L., Xie, W., Parkhi, O., & Zisserman, A. (2017). VGGFace2: A dataset for recognising faces across pose and age.
8. E.Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.
9. WJ Krzanowski. Between-groups comparison of principal components. Journal of the American Statistical Association, 74(367):703–707, 1979.

Acknowledgements

We would like to thank the researchers who created these different databases (Huang et al. 2007, Lundavist et al. 1998, and Cao et al. 2017), as well as the researchers who developed the ResNet-50 architecture (Kaiming et al. 2016).

schwarex@bc.edu