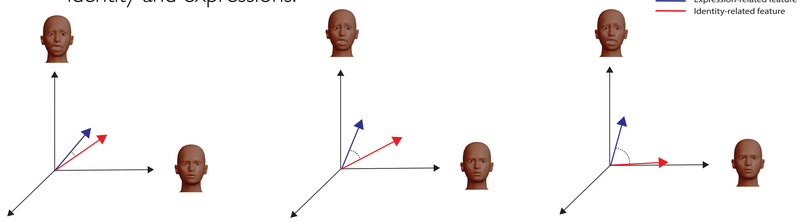
Investigating the emergence of expression representations in a neural network trained to discriminate identities

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 response patterns 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 expressions and identity is driven by computational efficiency. Recognition of identity and expressions might be "complementary" and benefit from each other. For example, if a face image has curved eyebrows, and it is overall consistent with an angry expression, we can conclude that the curved eyebrows are not a lasting property that is diagnostic of that face's identity, improving our accuracy to recognize that same identity with a neutral expression in future images.
- Our lab has recently 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), face processing might disentangle identity and expressions.



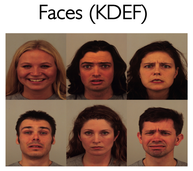
We investigate transfer learning in the reverse direction: Can ANNs trained to distinguish between identities learn features that support recognition of facial expressions?

Methods

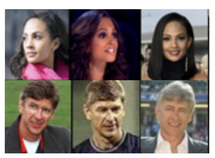
Face stimuli:

Labeled Faces in the Wild (LFW)

Karolinska Directed Emotional Faces (KDEF)



VGGFace2

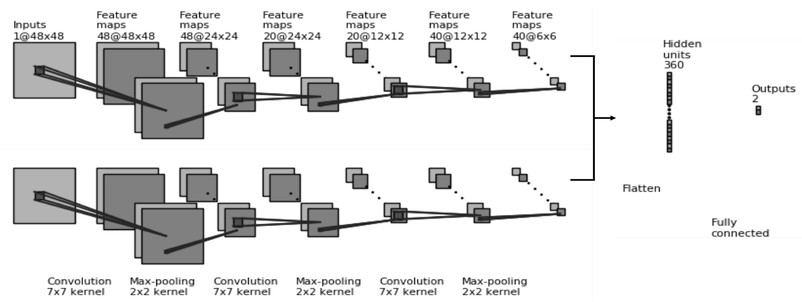


Network architectures:

1. Siamese network
   - Trained without handcrafted features using the LFW[5] database to discriminate between identities.
   - 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.
2. ResNet-50[6]
   - Pre-trained with VGGFace2[7] database to perform identity recognition.
   - The net was trained to minimize the cross-entropy loss.

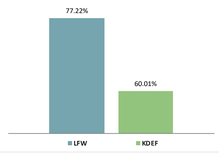
All of the networks were tested using the KDEF[8] dataset.

Face verification network



Siamese network architecture trained to perform a face verification task. The network learns to determine if two identities are the same or different.

- The network is able to discriminate between identities when tested on LFW, but performs poorly on the KDEF dataset.

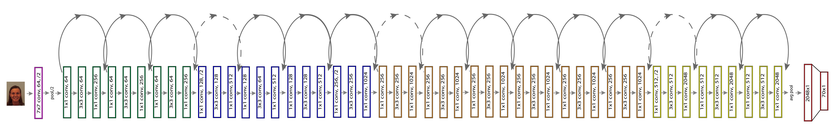


Transfer learning from net to expression recognition

- A FC layer was attached to the pre-trained network and re-trained to generate the correct output.
- The pre-trained network's weights are fixed after the initial training, so that nonlinear features cannot be learned from the loss based on expression.
- The network performed with an accuracy of 15% (at chance) when tested on expression labeling, failing to transfer.
- Since the net did poorly on identity as well, it can be difficult to interpret the results. Therefore, we wanted to use a more accurate network.

Face identification network: ResNet-50

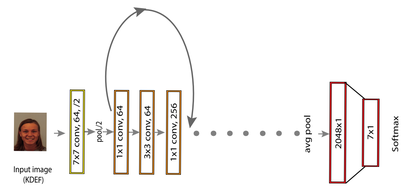
- We tested the net's ability to generalize to the KDEF dataset.
- The pre-trained network was able to perform face identification on the KDEF dataset with an accuracy of **98.6%**.



ResNet-50 architecture pre-trained with VGGFace2 to perform a face identification task. The last layer was removed and a FC linear layer was attached to generate labels for KDEF.

Transfer learning using ResNet-50 architecture to expression recognition

- 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 rely on the identity-based training.
- The net is able to label expression above chance (14.2%) with an accuracy of **62.6%**.



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.

Ongoing and future directions

- Deep networks trained to recognize identity might yield good transfer to expressions either because 1) identity and expressions rely on common features, or because 2) they need to disentangle identity from expression, leading to disentangled expression representations as a byproduct. Analyses in the opposite direction (training on expression and testing on identity) provided support for the second hypothesis. We are currently testing this for the network trained on identity.
- 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.

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