Spontaneous Learning of Face Identity in Expression-Trained Deep Nets

Spontaneous Learning of Face Identity in Expression-Trained Deep Nets

Emily Schwartz (schwarex@bc.edu)
Department of Psychology and Neuroscience, Boston College, Chestnut Hill, MA 02467

Kathryn O'Nell (kathryn.c.o'nell.gr@dartmouth.edu)
Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755. E.S. and K.O. contributed equally

Rebecca Saxe (saxe@mit.edu)
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Stefano Anzellotti (stefano.anzellotti@bc.edu)
Department of Psychology and Neuroscience, Boston College, Chestnut Hill, MA 02467

Abstract

Recent neural evidence challenges the traditional view that face identity and facial expressions are processed by segregated neural pathways, showing that information about identity and expression are encoded within common brain regions. This article tests the hypothesis that integrated representations of identity and expression arise naturally within neural networks. Deep networks trained to recognize expression and deep networks trained to recognize identity spontaneously develop representations of identity and expression, respectively. These findings serve as a "proof-of-concept" that it is not necessary to discard task-irrelevant information for identity and expression recognition.

Keywords: face processing; identity recognition; expression recognition; deep neural networks; transfer learning.

Introduction

The classical view of face processing proposes that identity and expression information are distinct mechanisms (Bruce & Young, 1986; Haxby, Hoffman, & Gobbini, 2000): an identity-specialized pathway (i.e. ventral temporal; Kanwisher, McDermott, and Chun, 1997; Gauthier et al., 2000) discards expression information, and an expression-specialized pathway (i.e. lateral temporal; Haxby et al., 2000; Hoffman and Haxby, 2000) discards identity information. However, recent evidence weighs against this view. Identity can be decoded in lateral temporal regions (Anzellotti & Caramazza, 2017; Dobs, Schultz, Bülthoff, & Gardner, 2018) and facial expression valence can be decoded in ventral temporal regions (Skerry & Saxe, 2014; Kliemann et al., 2018). An alternative hypothesis suggests that identity and expression might not depend on separate neural mechanisms (Duchaine & Yovel, 2015).

Here, we test whether learning to recognize facial expression necessarily requires discarding identity information (and vice versa), or whether recognition of facial expression and face identity might be mutually beneficial. To evaluate this, we train deep convolutional neural networks (DCNNs) to recognize expression and probe whether they spontaneously learn identity information and, likewise, we train DCNNs to recognize identity and probe whether they spontaneously learn expression information.

Methods

To understand if these face tasks must be implemented by separate mechanisms, we test if discarding irrelevant task information is necessary for successful facial expression and face identity recognition. If this is the case, identity information should decline when learning expression information and vice versa. Using Pytorch (Paszke et al., 2017), a deep DenseNet was constructed for each model, consisting of 1 convolutional (CONV) layer, 3 dense blocks, and 1 fully connected (FC) linear layer (Fig. 1). All networks described were trained 10 times with random weight initialization to test the consistency of the results. For convenience, the 10 runs will be referred to as a single DCNN.

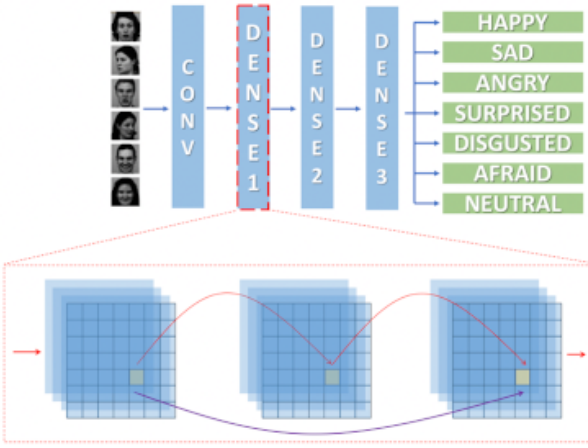

Figure 1: Architecture. *Top:* Each network consist of a convolutional layer, three dense blocks, and a fully-connected layer (face images shown taken from KDEF, Lundqvist et al., 1998). *Bottom:* Structure of a dense block.

Stimuli and Training

Expression and Identity DCNNs The expression DCNN was trained to label facial expressions using images from the Facial Expression Recognition 2013 (fer2013) dataset (Goodfellow et al., 2013), containing 28,709 training images and 3,589 testing images. The identity DCNN was trained to label identities using the Large-Scale CelebFaces Attributes (CelebA) dataset (Liu, Luo, Wang, & Tang, 2015). To match

the dataset sizes for the two networks, a subset of CelebA was used.

Untrained and Scene DCNNs To investigate if findings were due to the architecture and did not depend on training, we tested whether an untrained DCNN also supports identity and expression recognition in later layers. To evaluate if findings were face-specific, we probed a DCNN trained to label scenes using UC Merced Land Use dataset (Yang & Newsam, 2010).

Testing

After training, network weights were fixed to prevent further learning. To test identity and expression labeling, we used an independent dataset of images: Karolinska Directed Emotional Faces (KDEF) (Lundqvist et al., 1998). Accuracy was evaluated for features from the initial CONV layer, and the last layer of each dense block, after being summed with the inputs of the block. To accommodate differing output numbers, layer features were extracted, run through batch normalization, ReLU, and average pooling, followed by an FC linear layer to produce output labels ('readout layer'). A linear layer trained directly on pixel values was used as a control. To control for low-level features, all readout layers were trained using all but one of the viewpoints (frontal, 45 degree left, 45 degree right). Accuracy was tested using the left-out viewpoint (as in Anzellotti, Fairhall, and Caramazza, 2013), averaged across the three conditions.

Results

Expression and Identity Classification

Expression-trained and identity-trained DCNNs performances on an independent dataset Both networks generalized to perform accurately their respective trained tasks on KDEF. The expression DCNN labeled expression with a final accuracy of 53.4% and the identity DCNN labeled identity with a final accuracy of 48.35%.

Expression-trained and identity-trained DCNNs develop identity and expression representations respectively Features extracted from the CONV layer and each dense block of the expression DCNN were used as inputs to a corresponding FC layer for identity readout: accuracies of 9.5%, 6.3%, 14.8% and 20.2% respectively (Fig. 2A, top). In a parallel analysis, identity DCNN features labeled expression with accuracies of 17.6%, 17.1%, 21.5% and 42.1% (Fig. 2A, bottom).

Expression and identity recognition using features from an untrained DCNN and a scene-trained DCNN Features extracted from the CONV layer and each dense block of the untrained DCNN yielded accuracies of 16.5%, 16.2%, 15.5% and 16.5%, respectively, for expression labeling (Fig. 2A, bottom). For identity labeling, untrained features extracted from each layer yielded accuracies of 7.9%, 7.1%, 13.6% and 6.1% (Fig. 2A, top). The untrained DCNN performed similarly within a task for all layers (close to chance level).

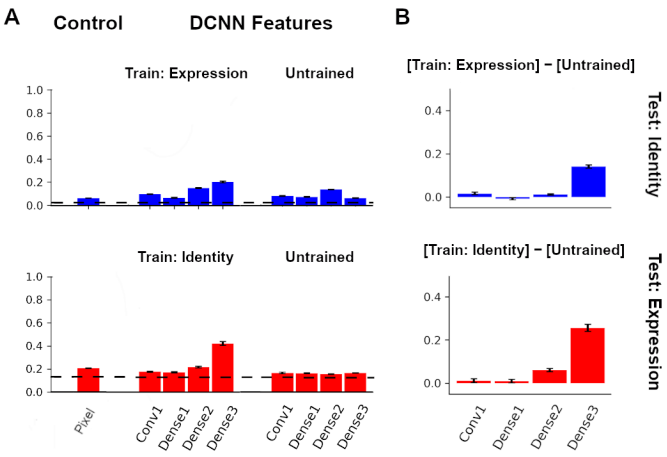

Figure 2: Comparisons with the Untrained Network. A) *Left:* Classification performance using expression features for identity labeling and identity features for expression labeling, *Right:* Classification performance using untrained features. B) Difference between trained and untrained networks. Error bars in plots denote the SEM of the performance of network instances.

When using the scene DCNN to label expression, features from each layer yielded accuracies of 15.9%, 16.0%, 23.5% and 33.0%, respectively. For the scene DCNN when labeling identity, layer features yielded accuracies of 9.5%, 7.8%, 17.3% and 29.6%.

Discussion

We propose that recognition of facial expression and face identity are 'complementary' tasks – that representations optimized to recognize facial expression also contribute to the recognition of face identity, and vice versa. This would account for the observation that identity and expression information coexist within common brain regions (Anzellotti & Caramazza, 2017; Dobs et al., 2018). Features from an expression-trained DCNN can support accurate identity recognition, and reciprocally, features from an identity-trained DCNN can support accurate expression recognition. Our findings serve as an existence proof that in order to perform identity recognition, expression information does not need to be discarded (and vice versa). In fact, within our models, networks trained to perform one task do not just retain information that can help solve the other task: they enhance it. Surprisingly, findings were not category-specific, contrasting with other transfer learning studies (Yosinski, Clune, Bengio, & Lipson, 2014). However, features did not simply arise in an untrained network either. Ongoing work includes the application of these models to neural data.

Acknowledgments

This work is supported by the National Science Foundation under Grant No. 1943862.

References

Anzellotti, S., & Caramazza, A. (2017). Multimodal representations of person identity individuated with fmri. *Cortex*, *89*, 85–97.

Anzellotti, S., Fairhall, S. L., & Caramazza, A. (2013). Decoding representations of face identity that are tolerant to rotation. *Cerebral Cortex*, *24*(8), 1988–1995.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British journal of psychology*, *77*(3), 305–327.

Dobs, K., Schultz, J., Bülthoff, I., & Gardner, J. L. (2018). Task-dependent enhancement of facial expression and identity representations in human cortex. *NeuroImage*, *172*, 689–702.

Duchaine, B., & Yovel, G. (2015). A revised neural framework for face processing. *Annual review of vision science*, *1*, 393–416.

Gauthier, I., Tarr, M. J., Moylan, J., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). The fusiform "face area" is part of a network that processes faces at the individual level. *Journal of cognitive neuroscience*, *12*(3), 495–504.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... others (2013). Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing* (pp. 117–124).

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, *4*(6), 223–233.

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature neuroscience*, *3*(1), 80.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, *17*(11), 4302–4311.

Kliemann, D., Richardson, H., Anzellotti, S., Ayyash, D., Haskins, A. J., Gabrieli, J. D., & Saxe, R. R. (2018). Cortical responses to dynamic emotional facial expressions generalize across stimuli, and are sensitive to task-relevance, in adults with and without autism. *Cortex*, *103*, 24–43.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015, December). Deep learning face attributes in the wild. In *Proceedings of international conference on computer vision (iccv).*

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, *91*, 630.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., ... Lerer, A. (2017). Automatic differentiation in pytorch.

Skerry, A. E., & Saxe, R. (2014). A common neural code for perceived and inferred emotion. *Journal of Neuroscience*, *34*(48), 15997–16008.

Yang, Y., & Newsam, S. (2010). Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th sigspatial international conference on advances in geographic information systems* (pp. 270–279).

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320–3328).