How do deep neural networks represent faces?

Christian Tsvetkov and Ivan Vankov

Department of Cognitive Science, New Bulgarian University

Deep learning models of vision have recently achieved human-level of performance in tasks such as object recognition and face identification. However, the extent to which these computational models resemble the way biological vision works remains unclear. We address this question by investigating how a deep learning model of face identification represents faces and by comparing the results to data from behavioural experiments. To this end, we use the Bubbles technique (Gosselin & Schyns, 2001) in order to find the critical regions within an image which are needed for correct face identification. The technique is applied to a pretrained deep convolutional neural network model of face recognition (Amos, Ludwiczuk, & Satyanarayanan, 2016), and the results are related to human data from a replication of Gosselin & Schyns (2001). We also use Bubbles to check how faces are represented in a simple connectionist model and in a deep learning model trained on general object recognition.

Previous efforts in studying the details representations in neural networks have, for the most part, focused on visualizing activations for separate units, thus showing locally coded features. However, it can be argued that for a methodology to be comparable to human data, it would need to show a more holistic account of the represented information. To determine region salience in images with Bubbles, multiple transparent windows are generated at random locations on an otherwise opaque mask, overlaid on a learned stimulus. With enough iterations, all locations in an image are sampled. The ratio of bubble masks leading to correct identifications and the set of all presentations, originally called ‘proportion plane’, serves as a measure of diagnosticity. Higher values in this 2-D plane indicate higher degree of diagnosticity. We replicate one of the experiments from Gosselin & Schyns (*2001. Bubbles: a technique to reveal the use of information in recognition tasks. Vision research, 41(17), 2261-2271*) involving identification of individual faces. Using the same set of stimuli, we then use an open source deep learning model of face recognition (*Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). OpenFace: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science.*) to represent the images, and train several classifiers. The same deep model is used to represent all Bubble ‘masks’ generated. Using the same methodology, correct classifications are recorded and used for the production of another ‘proportion plane’. Finally, the planes from the simulation and behavioral experiments are compared using both qualitative and quantitative methods. We repeat this procedure by testing additional models: a more classical early connectionist model with a single hidden layer, and another deep learning network, an instantiation of (*Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-9)*, used for large-scale object categorization (ILSVRC 2014 classification challenge). The former simulation aims to investigate the qualitative differences of the representations owed to the 'deep' architecture, while the latter examines how learned detectors can generalize to unseen categories.