Improving the fit between human and DNN representational spaces using weighted least squares regression

Anna Bavaresco[1], Nhut Truong[1], Homa Priya Tarigopula[1], and Uri Hasson[1]

[1]Center for Mind/Brain Sciences (CIMeC), The University of Trento, Italy

Abstract

Reducing the existing gap between the human visual representational space and that generated by Deep Neural Networks (DNNs) is a core issue for both cognitive and AI research. While initial research has documented that DNNs can predict human visual representations, more recent work has developed approaches to modify DNNs so that their ability to predict human responses is improved. A standard way to quantify these predictions' accuracy is by assessing the correlation between 1) pair-wise distances computed from feature arrays (embeddings) obtained by passing the images through a DNN and 2) human similarity judgments (HSJs) over image-pairs. Previous research seeking to improve the match between DNN data and human data has mainly focused on altering the DNN features (reweighting) or removing some to obtain an optimal feature subset (pruning). Those methods implicitly assume that human similarity judgments across levels of similarity are all equally informative. However, depending on the specific objects involved in the comparison, people's confidence and agreement when expressing similarity judgments may vary. For example, people may be less precise in assigning mid-range similarity values, or low-range ones, which would make those ranges of values more difficult to fit. We, therefore, propose a machine-learning approach where certain ranges of HSJs are up-weighted when learning a fit with DNN data. More specifically, we develop a weighted least squares framework to predict a DNN similarity matrix from a human similarity matrix. The weights applied to the residuals are computed as a function of the HSJ (i.e., the magnitude of the similarity judgment itself). Given that the most informative HSJs are not known *a priori*, several candidate weighting functions are examined, each enhancing a different range of similarity-magnitude expressed in the human judgments. The simplest function was a linear one that assigned greater weight to judgments of higher similarity, and this function already produced better prediction of out-of-sample data to which it was applied. However, a parabolic function that up-weighted both low and high similarity judgments produced the best fit with DNN data. That fit proved to be even higher than the one achievable with the compared state-of-the-art method, i.e. 'pruning'. In order to further increase predictive capacity, we also used weighted judgments to 'prune' a DNN, thus combining pruning and weighting. We found that this weighted pruning altered the DNN representational space so that it increased inter-category distance and decreased intra-category distance between animal categories. To summarize, the contribution of this work is threefold. First, we demonstrate that the high-range and low-range HSJs up-weighted by the parabolic function are the most informative for predicting DNN similarity. Second, we propose a procedure for improving the fit between human and DNN representational spaces which outperforms the compared state-of-the-art method. Third, we show that weighted pruning modifies the network's representational space producing a better alignment with the human organization of animal categories.