Attentional Bias in Human Category Learning: The Case of Deep Learning

Leyla Roksan Çağılar (l.r.caglar@gmail.com) & Stephen José Hanson (jose@rubic.rutgers.edu)

Department of Psychology, 101 Warren Street
Rutgers Brain Imaging Center (RUBIC), 197 University Street, Newark, NJ 07102 USA

Introduction

Human category learning is a function of both the complexity of the category rule and attentional bias (Shepard et al.; 1961, Shepard, 1987). A classic diagnostic category problem involves learning integral (correlated) stimuli using a condensation rule and separable (uncorrelated) stimuli using a filtration rule (Garner, 1974). Human category learning shows differential learning speed based on these category rules that either require the attentional binding of features or ignoring of features. Computational models from the 90s using standard backpropagation neural networks (BP) with a single sigmoidal hidden layer have failed to capture this human-like attentional bias (Kruschke, 1993). Two hypotheses can be put forth to explain why BP failed at modelling the human-like attentional bias in the filtration/condensation task. One hypothesis is that BP has no way of doing attentional feature binding and thus is unable to distinguish complexity differences across the two tasks and consequently learns them at the same rate. Another hypothesis is that unlike naturalistic stimuli, previously used stimuli by Garner and Kruschke were "well-defined", binary, and lacked similarity gradients that could be used to distinguish the underlying prototypes.

Methods

Two distinct prototypes generated 40 3D-modelled faces based on family resemblance. Guided by 6D non-metric multidimensional scaling based on similarity judgments (N=20), we created separable and integral stimuli sets based on low and high perceptual space distances to then be used in a filtration separable and a condensation integral task. Human behavioral data was collected in addition to modelling the same task with BP and a Deep Learning neural network (DL).

Results

The results show that both BP and DL are able to successfully model human-like attentional bias in category learning when using naturalistic stimuli, but not with "well-defined" stimuli. Further, BP and DL appear to have qualitatively different learning curves. BP shows a classical negative growth (fitting a negative exponential function) over trials, while DL shows the Thorndike (1932) accumulation learning model (fitting a hyperbolic exponential function) which was less favored over the last century. The negative exponential function of BP is in line with the replacement model, in which learning rates depend directly on how much is left to learn (Estes, 1950). The exponential hyperbolic function of DL however is in line with the accumulation model, in which learning rates depend on the existing rate of learning and the amount left to learn (Gulliksen, 1934; Thorndike, 1923). These differential learning models may give further insight into the fundamental learning processes employed by BP and DL.

Discussion

What are the brain supporting mechanisms of information compression? Information compression is the essence of category learning, which is a simplification of category representations while minimizing loss of information. The internal representation of the hidden units shows that successive re-encoding in DL leads to sequential extraction of features and the development of more sensitive feature detectors. This is the same recursive and hierarchical architecture that can be found in the visual system for information compression.

References

Garner, W. R. (1974). *The processing of information and structure* (Vol. 20).
Gulliksen. (1934). A rational equation of the learning curve based on Thorndike’s law of effect. *Journal of General Psychology, 11*, 395–434.
Kruschke, J. K. (1993). *Human category learning: Implications for backpropagation models* (Vol. 5).
Mazur, J., & Hastie, R. (1972). Learning as accumulation: A reexamination of the learning curve. *Psychological Bulletin, 85*, 1256–1274.
Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1*(4820), 54–87.
Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science, 237*(4820), 1317–1323.
Thorndike, E. (1919). The fundamentals of learning. *New York: Teach. Coll., Bur. Publ.*, 638.