Are baboons learning "orthographic" representations? Probably not.

Harald Baayen, Maja Linke, Franziska Broeker and Michael Ramscar

The ability of Baboons (papio papio) to distinguish between English words and nonwords [1] has been modeled using a deep learning convolutional network model that simulates a ventral pathway in which lexical representations of different granularity develop [2]. However, given that pigeons (columba livia), whose brain morphology is drastically different, can also be trained to distinguish between English words and nonwords [3], it appears that a less species-specific learning algorithm may be required to explain this behavior. Accordingly, we examined whether the learning model of [4],which has proved to be amazingly fruitful in understanding animal and human learning [5-7] could account for these data. We show that a discrimination learning network using gradient orientation features as input units and word and nonword units as outputs succeeds in predicting baboon lexical decision behavior - including key lexical similarity effects and the ups and downs in accuracy as learning unfolds - with surprising accuracy. The performance of this model, in which words are not explicitly represented, is remarkable because it is generally assumed that lexicality decisions, including the decisions made by baboons and pigeons [2,3] are mediated by explicit lexical representations. Our results suggest that in learning to perform lexical decision tasks, baboons and pigeons do not construct a hierarchy of lexical units, but rather they make optimal use of low-level information obtained through the massively parallel processing of gradient orientation features. Accordingly, we suggest that skilled fluent reading of both simple and morphologically complex words in humans may involve a transition from a high-level system building on letter representations acquired during explicit instruction in literacy to the use of a similar strategy of exploiting massively parallel processing from low-level visual features to semantics.

These results have several methodological implications for theories of lexical access.

1. Correlation is not causation. If measures tied to words, such as word frequency or by-word random intercepts, have a significant effect in a statistical model, this does not imply that the effect originates from word-specific lexical representations. The lexicality decisions predicted by our model show, just as is the case for the lexicality decisions made by the baboons, a strong and robust random effect for word. Yet, there are no word representations in our model. The only representations in our model are low-level histogram of gradient orientation features and representations for left and right button presses. Therefore, we have to take seriously the possibility that baboon lexical decision behavior simply reflects the changes in the environment in which the baboon is placed, much as a mirror will reflect images of its beholders without having internal representations of those beholders. Likewise, frequency effects in human lexical decision are consistent with resting activation levels for corresponding lexical representations, but are also fully compatible with functional models of lexical access in which no such form representations exist [7].

2. Design decisions determine model behavior. Hannagan et al. (2014) used a deep convolution network to model baboon lexical decision making. The model had a limited number of layers, enabling it to develop units in upper layers that were sensitive to letter

pairs and letter trigrams, but not to words. However, if a network with more layers and sufficient numbers of units had been used, units sensitive to individual words could have developed. Deep learning offers researchers so many degrees of freedom that models can be easily tailored to produce the desired outcome. Wide learning, by contrast, is severely constrained, and is driven almost entirely by the distributional characteristics of the input/output mapping. The only design decision with respect to parameters concerns the learning rate, which we have found to be optimal when fixed at 0.001. [8]

3. Wide learning can achieve nonlinear separation. The potential of two-layer networks (i.e., networks with only input and output units, and no hidden layer) has been severely underestimated ever since Minsky & Papert's (1972) critique of the perceptron. The advances in artificial neural networks, first in the eighties with backpropagation networks, and its recent renaissance with deep learning, have shown that Minsky and Papert were much too pessimistic about the computational possibilities of artificial neural networks. However, two-layer networks are still regarded as inferior and not up to realistic tasks, especially those that require non-linear separation. This negative appraisal of two-layer networks rests on a fundamental misunderstanding of their capacities. Classification problems that do not allow for linear separation can be solved by proper re-representation. For instance, support vector machines use kernels to rerepresent data points in a higher dimensional space such that linear separation becomes possible. Two-layer networks, when given smart input representations, can likewise solve non-linearly separable classification. [9] Therefore, wide learning networks are likely to be sufficiently powerful to provide formalizations for the role of proportional analogy in Word and Paradigm Morphology [10].

References

1. Grainger J, Dufau S, Montant M, Ziegler JC, Fagot J. (2012). Orthographic processing in baboons (papio papio), Science 336(6078):245–248.

2. Hannagan T, Ziegler JC, Dufau S, Fagot J, Grainger J. (2014). Deep learning of orthographic representations in baboons, PLOS-one. 9:e84843.

3. Scarf D, Boy K, Reinert AU, Devine J, Gunturkun O, Colombo M. (2016). Orthographicprocessing in pigeons (Columba livia), Proceedings of the National Academy of Sciences 113(40):11272–11276.

4. Rescorla RA, Wagner AR. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokasy WF, editors. Classical conditioning II: Current research and theory, New York: Appleton Century Crofts, p. 64–99.

5. Miller RR, Barnet RC, Grahame NJ. (1995). Assessment of the Rescorla-Wagner Model. Psychological Bulletin, 117(3):363–386.

6. Ramscar M, Yarlett D, Dye M, Denny K, Thorpe K. (2010). The Effects of Feature-Label-Order and their implications for symbolic learning, Cognitive Science 34(6):909–957.

7. Baayen RH, Milin P, Filipovic Durdevic D, Hendrix P, Marelli M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning, Psychological Review 118:438–482.

8. Bitschnau, S (2015). An exploration of computational learning algorithms: A closer look at the Rescorla-Wagner model and the Danks equilibrium equation in language processing, BA thesis Cognitive Science, University of Tuebingen.

9. Baayen, R. H., and Hendrix, P. (2017). Two-layer networks, non-linear separation, and human learning. In Wieling, M., Kroon, M., Noord, G. van, and Bouma, G. (Eds.) From Semantics to Dialectometry. Festschrift in honor of John Nerbonne. London, College Publications, 13-22.

10. Blevins, James P. (2016) Word and paradigm morphology. Oxford University Press.