Modeling familiarity through the combination of Deep Learning and Hebbian training

John **Read**[1] & Jacques Sougné[2]

[1] GIGA-CRC In Vivo Imaging, University of Liège, Liège, Belgium; [2] UDI FPLSE, University of Liège, Liège, Belgium

LIÈGE université GIGA CRC In vivo Imaging

LIÈGE université Psychologie, Logopédie & Sciences de l'Éducation

john.read@doct.uliege.be

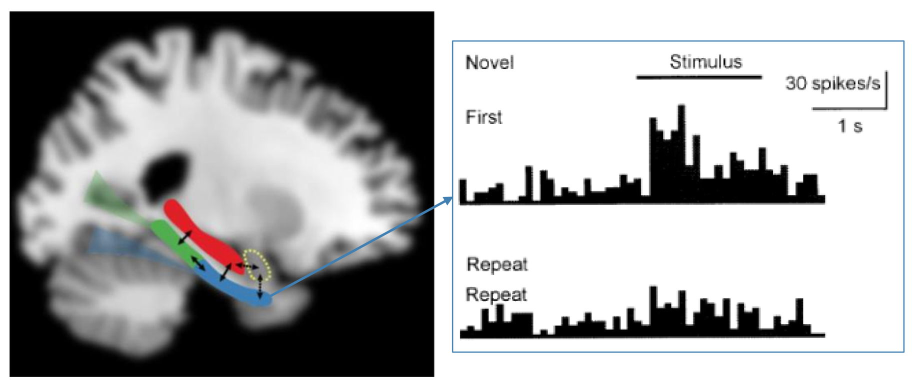
Theoretical Background

Dual-Process Signal Detection theory

Familiarity is a type of recognition that gives a quantitative measure about a previously learned stimulus. That is, familiarity corresponds to the degree of similarity between the characteristics of a perceived stimulus and the characteristics of an old stimulus stored in memory[1]. When familiar, a specific stimulus has a higher level of familiarity than a novel stimulus.

Perirhinal cortex & novelty neurons

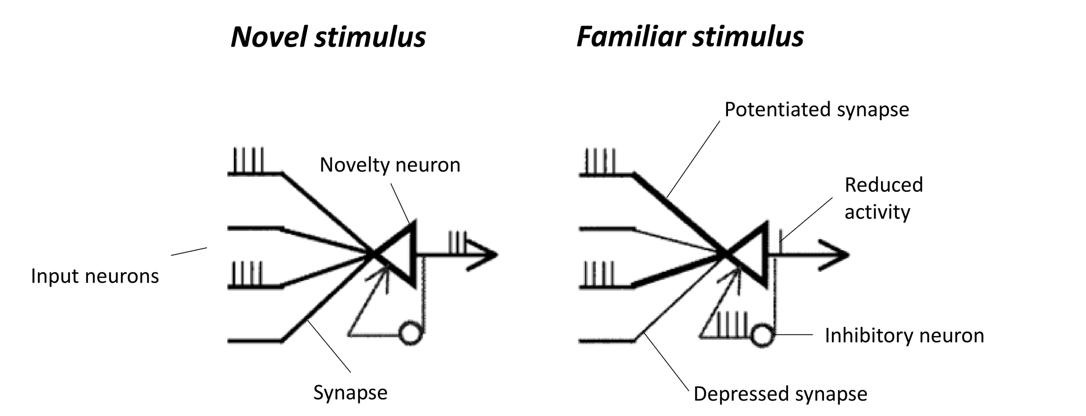
The perirhinal cortex (PRC, in blue) seems to be crucial in the familiarity processes[2]:



Novelty neurons in the PRC respond stronger when presented with a new stimulus. Once familiar, the activity of novelty neurons in the PRC is reduced.

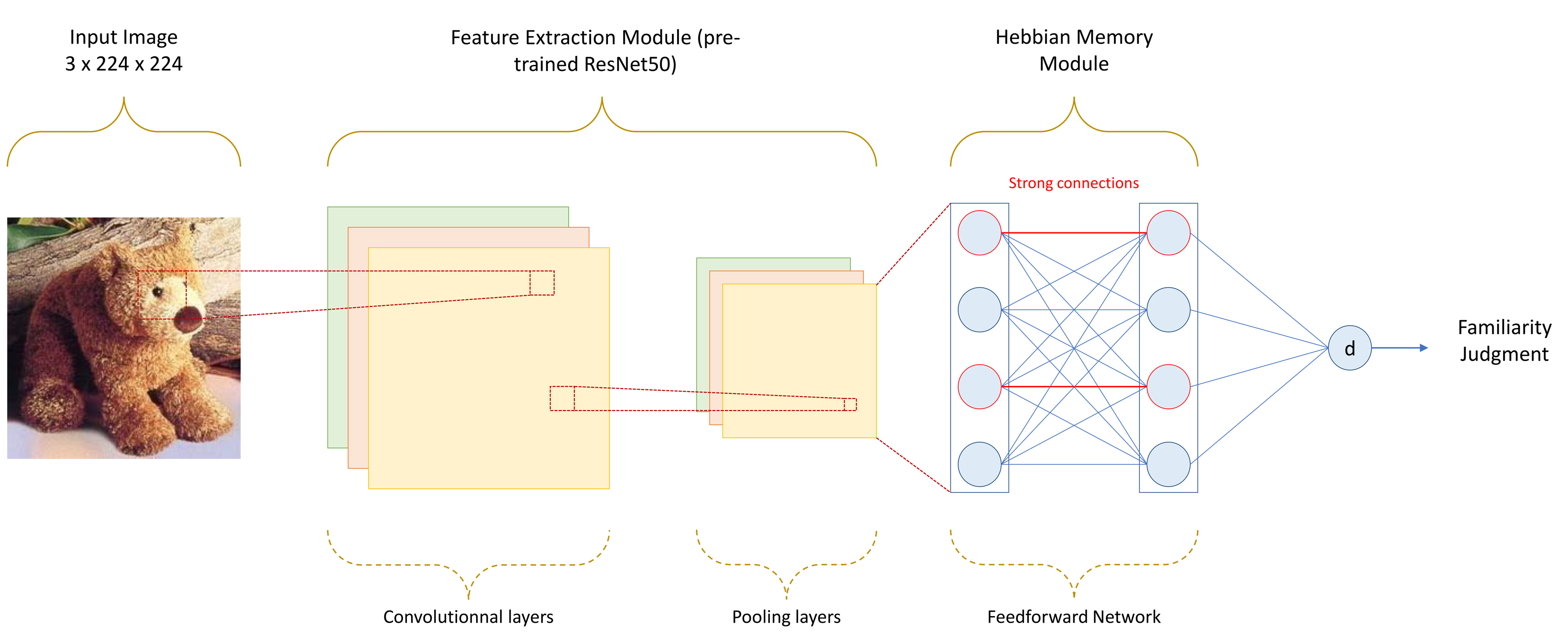
Modeling with Hebbian learning

Artificial neural networks succesfully used Hebbian learning to model familiarity in the PRC on formal binary patterns[3]:



Familiar stimuli have more inhibition than novel ones.

Modeling & Methodology

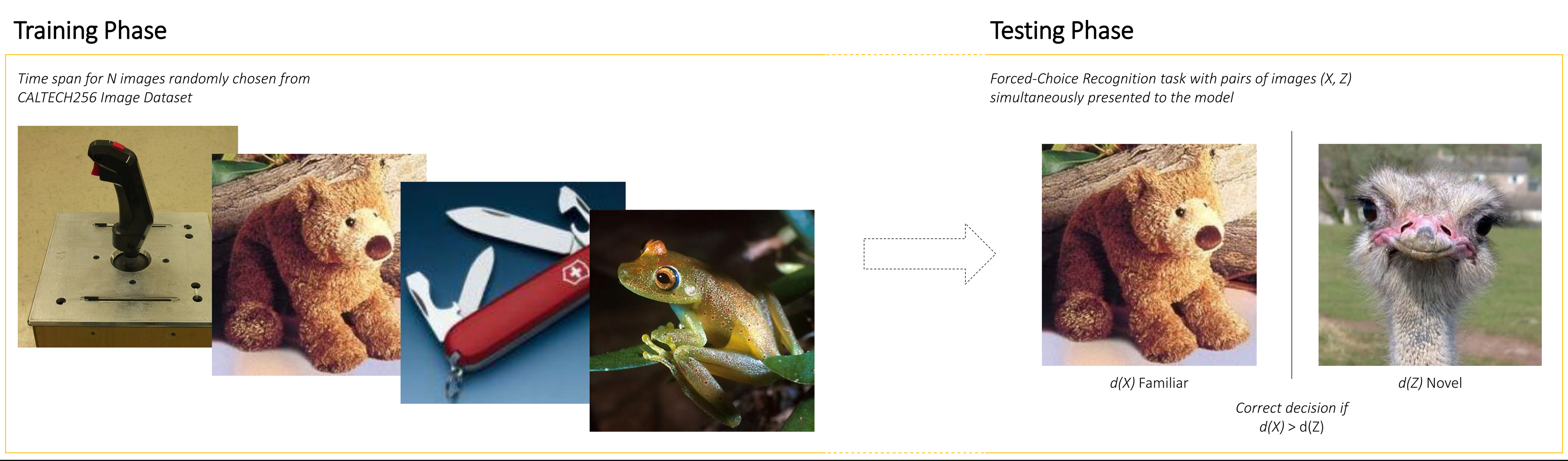


The model was implemented in Python 3.9.11 as three successive modules:

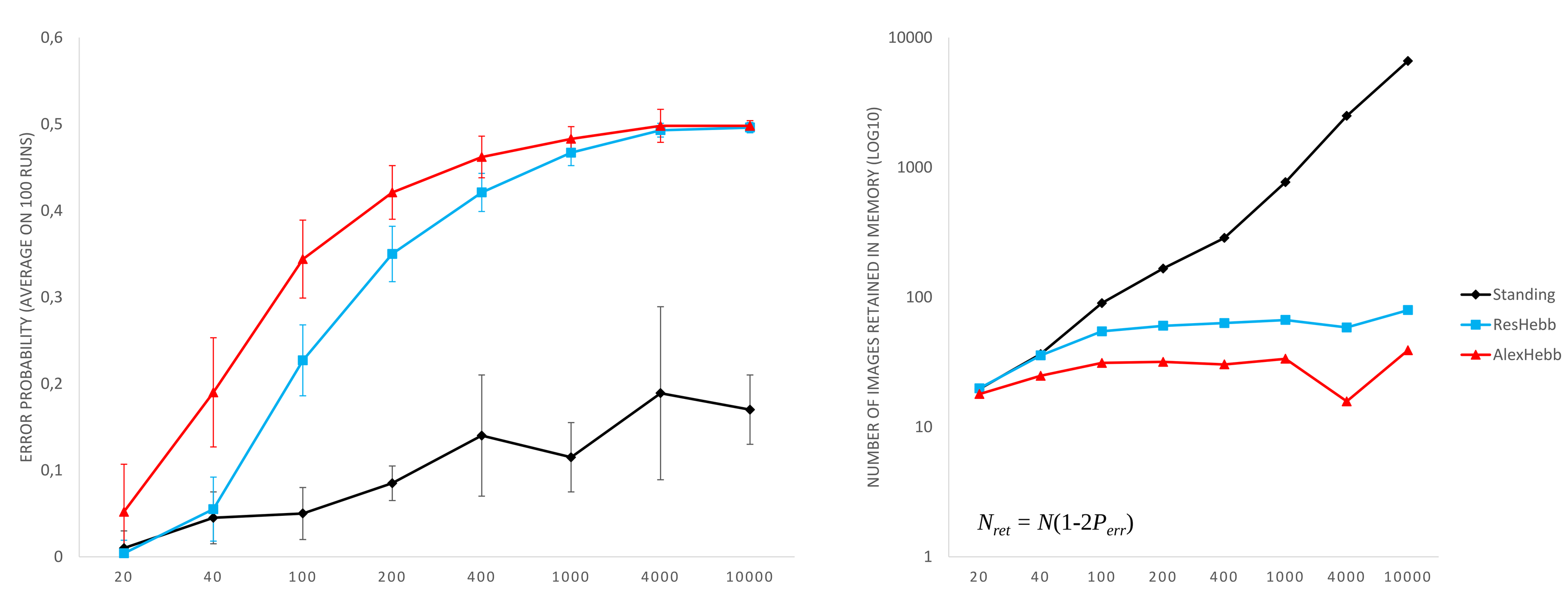
1. A Convolutional Neural Network (CNN) mimics the processing of a stimulus by the visual brain area[4]. We use a pre-trained version of ResNet50 to extract the features of images.
2. A two-layers Feedforward Neural Network learns the features of images with an Hebbian learning rule. The number of active output neurons is limited by the mean of non-modifiable strong connections[3].
3. An Inhibitory Interneuron computes the level of inhibition (*d*) for images.

Simulations took place according to Standing's experiment[5]. Three main simulations were performed to explore:

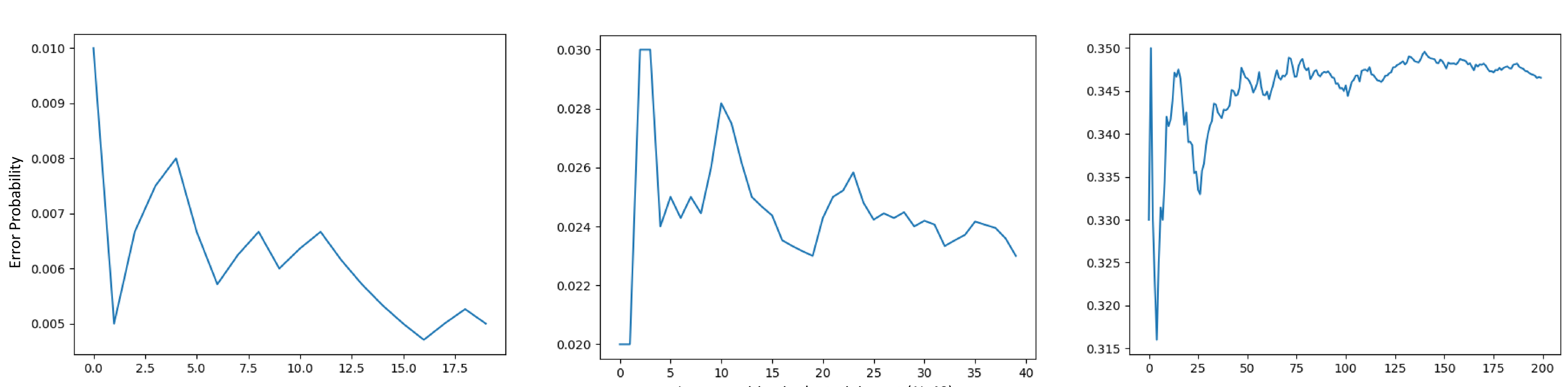
- The memory capacity of the model
- The presence of recency/primacy effects
- The performances on similar images (e.g. only cat images)



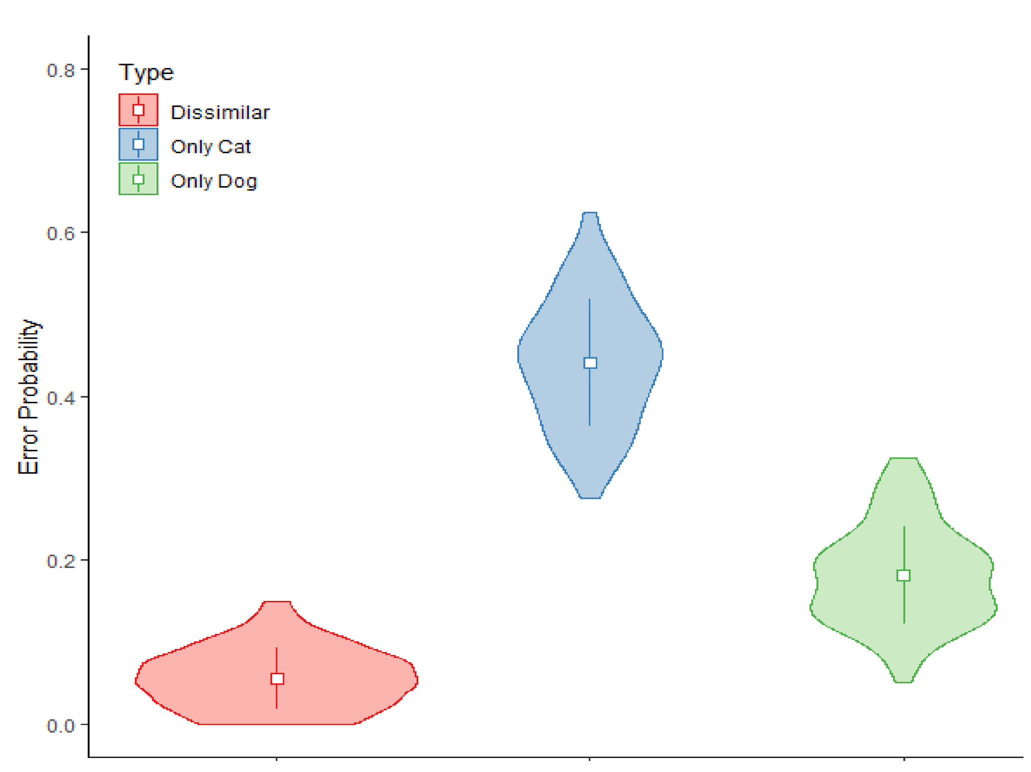
Simulation results



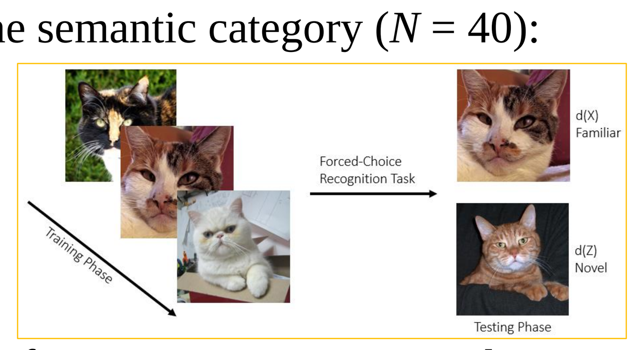
Simulation 1 showed that the Memory Capacity of the model is up to 40 images. Error probability and the number of images retained (N_{ret}) were computed over the entire task. Experimental data are depicted in black[5]. More efficient CNN improves model performances (AlexNet >< ResNet50).



Simulation 2 showed a Recency-like Effect when the number of learned images did not exceed the memory capacity of the model. No Primacy Effect was observed.



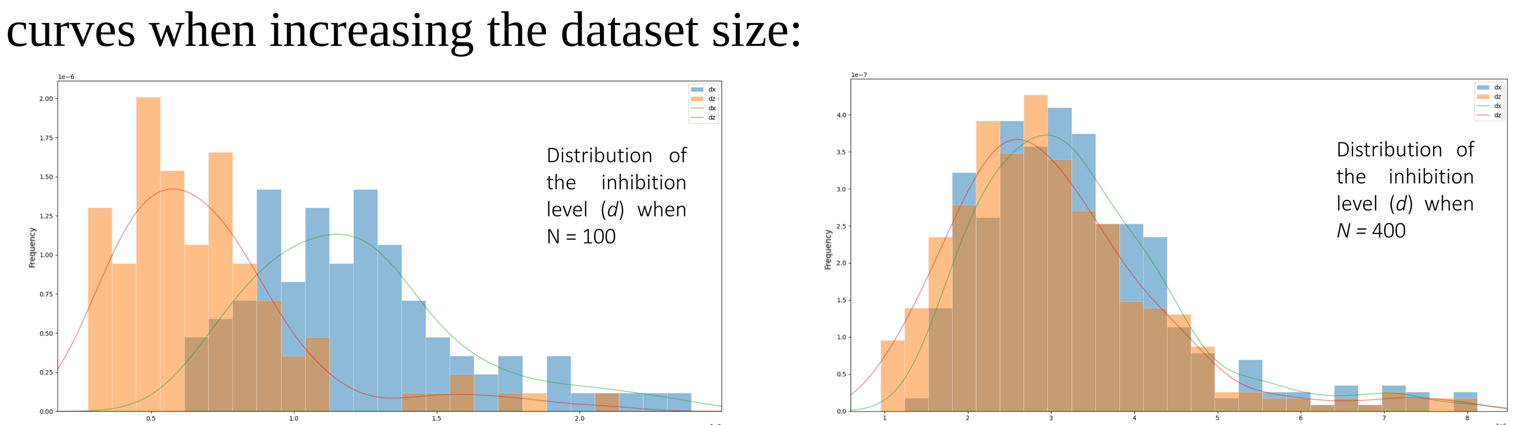
Simulation 3 showed a Similarity Effect. Model performances collapse when trained with images only from one semantic category (N = 40):



Performances are worse when tested with cats than with dogs, thus showing that the model is sensible to homogeneity between inputs during training.

Discussion

Our results suggest the Hebbian model learns the global representation of a stimulus. This could be done by encoding correlations shared by several stimuli[3]. Here, familiarity comes when the representation matches with the stimulus during the testing phase. Interestingly, model performances decrease drastically when too many stimuli are presented as well as when there is high homogeneity between inputs. It is then plausible that the model learns patterns with not enough details to allow discrimination. This would predict the high overlapping observed between *d* curves when increasing the dataset size:



Our computations are consistent with global matching theories of familiarity[6]. However, it is known that familiarity could arise from different coexisting mechanisms (e.g., stimulus-specific reduction of neural activity by repetition suppression)[7]. Nevertheless, it is not clear what are the conditions for a specific mechanism to take precedence over another. Here, we posit that familiarity could be expressed through the overall structure of stimuli (i.e., global matching) if the amount of information to be encoded is limited. Beyond a certain threshold of information to be learned, we hypothesize that familiarity would be expressed through other mechanisms.

References

1. Yonelinas, A. P., Aly, M., Wang, W. C. & Koen, J. D. Recollection and familiarity: Examining controversial assumptions and new directions. *Hippocampus 20*, 1178–1194 (2010).
2. Brown, M. W. & Aggleton, J. P. Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nat Rev Neurosci* **2**, 51–61 (2001).
3. Bogacz, R. & Brown, M. W. Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus* **13**, 494–524 (2003).
4. Kazanovich, Y. & Borisyuk, R. A computational model of familiarity detection for natural pictures, abstract images, and random patterns: Combination of deep learning and anti-Hebbian training. *Neural Networks* **143**, 628–637 (2021).
5. Standing, L. Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology* **25**, 207–222 (1973).
6. Clark, S. E. & Gronlund, S. D. Global matching models of recognition memory: How the models match the data. *Psychon Bull Rev* **3**, 37–60 (1996).
7. Grill-Spector, K., Henson, R. & Martin, A. Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn Sci* **10**, 14–23 (2006).