Replicating Posner & Keele (1968) with Standard Convolutional Neural Networks

Milan Van Maldegem (milan.vanmaldegem@esat.kuleuven.be)
Department of Electrical Engineering, KU Leuven, Belgium

Gido van de Ven (gido.vandeven@esat.kuleuven.be)
Department of Electrical Engineering, KU Leuven, Belgium

Hans Op de Beeck (hans.opdebeeck@kuleuven.be)
Laboratory for Biological Psychology, KU Leuven, Belgium
Leuven Brain Institute, KU Leuven, Belgium

Tinne Tuytelaars (Tinne.Tuytelaars@esat.kuleuven.be)
Department of Electrical Engineering, KU Leuven, Belgium

Abstract

This study tests whether standard convolutional neural networks (CNNs) replicate the human prototype effect in category learning. Using a setup based on Experiment III of Posner & Keele (1968), three CNN architectures were trained on abstract dot patterns. While none matched the human results, AlexNet and DenseNet-121 showed better accuracy for unseen prototypes than for new exemplars, suggesting a weak prototype bias. These results provide a foundation for further research on category learning in humans and CNNs.

Keywords: Prototype Effect, Category Learning, Convolutional Neural Networks

Introduction

Humans organise their experiences into categories – groups of similar objects, ideas and events that are treated as functionally equivalent (Minda et al., 2024).

But how do we classify unfamiliar instances? For example, how do we recognise a new animal as a dog or a cat without prior exposure to that specific animal? Cognitive science offers two hypotheses:

1. *Prototype Theories* assume that a category is represented by a single mental abstraction – a prototype (Posner & Keele, 1968). Novel instances are classified according to their similarity only to the central prototype.
2. *Exemplar Theories* assume that people learn categories by storing memory traces for all items in a stimulus set and comparing new samples with these stored examples (Medin & Schaffer, 1978).

Early research showed that humans can often easily abstract and recognise prototypes even when they haven't encountered them before – a phenomenon known as the *Prototype Effect* (Posner & Keele, 1968). Well-documented across various domains (see Cabeza et al., 1999 for an example), this effect supports the view that human behaviour can, at least in part, be explained by prototype theories.

This raises a key question: Do artificial vision models like CNNs behave similarly under the same experimental conditions? While a recent study explored this question (Singh et al., 2021), direct comparisons remain challenging due to key differences in the training and evaluation procedures used for humans and CNNs.

The current investigation aims to address this by closely replicating a pioneering study on the prototype effect in humans (Posner & Keele, 1968), using standard CNNs.

Experimental Setup

The dataset and training procedure closely match the third experiment of Posner & Keele (1968), excluding *day 2 evaluation* due to its limited impact on the results.

In short, the dataset consisted of six classes of abstract dot patterns (9 dots), each anchored by a central prototype (Figure 1). Within each class, stimuli included: one prototype, four training exemplars generated by applying 7.7-bit perturbations to the prototype, two testing exemplars with different 7.7-bit perturbations, and two testing exemplars with milder 5-bit perturbations. Across all classes, the perturbations for the exemplars were the same, as was the case in the original study (Posner & Keele, 1968).

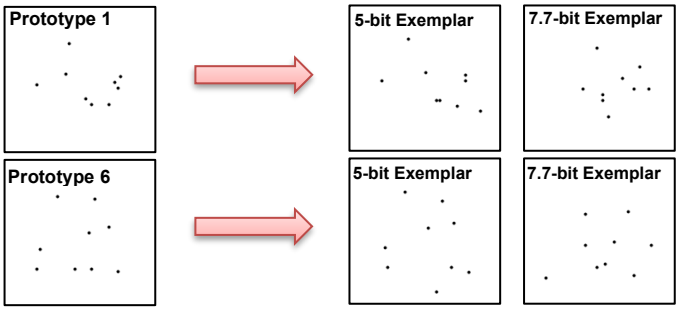

Figure 1. Two example prototypes and exemplars from the same classes (1 & 6).

We used three popular CNN architectures – AlexNet (Krizhevsky et al., 2012), DenseNet-121 (Huang et al., 2017), and ResNet-18 (He et al., 2016) – as learning agents. Models learned to classify 7.7-bit exemplars (4 per class) of three different classes (20 class combinations in total) until achieving two consecutive epochs where all images were correctly classified. The average number of errors to criterion was 37.5 for Humans (Posner & Keele, 1968), 34.72 for AlexNet, 9.59 for DenseNet-121, and 6.52 for ResNet-18. During each epoch, the mini batch size was 1, optimisation was performed using the *Adam Optimiser* (Kingma & Ba, 2014) with a learning rate of 0.0001, and

Cross-Entropy Loss was used as the loss function. Models were initialised with *ImageNet1K-pretrained weights* to mimic prior knowledge in humans, and fine-tuned by replacing the final fully connected layer to match the number of classes.

After training, performance was assessed by evaluating the models on four different conditions: already seen 7.7-bit exemplars (2 per class), the prototypes (1 per class), new 5-bit exemplars (2 per class), and new 7.7-bit exemplars (2 per class). In total, we obtained 20 performance scores – one for each combination of three classes – per CNN architecture per condition.

Results

Figure 2 presents the evaluation accuracies of all learning agents – including the human results from Posner and Keele (1968) – across conditions, revealing similar trends.

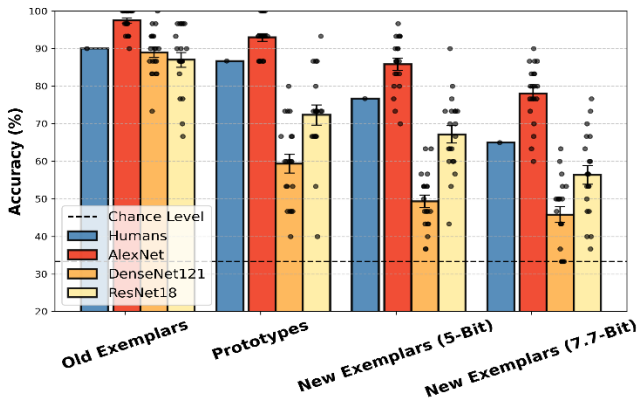

Figure 2. Bar plots show the average accuracies (with standard errors) on test examples after training for all learning agents. Every dot represents a different combination of 3 classes. The original publication provided only one human value (mean).

As in the original study, we used two-sided sign tests to compare conditions and assess whether CNNs replicate human behaviour. After confirming the necessary assumptions (independent and ordered differences), the results varied across models (Table 1). For AlexNet, performance differed significantly across all conditions. In contrast, DenseNet-121 showed no significant difference between new 5-bit and new 7.7-bit exemplars, and ResNet-18 showed similar performance for prototypes and new 5-bit exemplars. None of the CNNs matched the human results that showed similar performances for old exemplars and prototypes – the *Prototype Effect*.

Table 1. Two-Sided Sign Test Results.

Comparison	Humans	AlexNet	DenseNet	ResNet
Old vs Prototype	*ns* t = 12	0.002 t = 6	< 0.001 t = 0	< 0.001 t = 1
Prototype vs New 5-bit	0.01 t = 2	0.001 t = 2	< 0.001 t = 3	*ns* t = 2
New 5-bit vs New 7.7-bit	0.05 t = 5	0.001 t = 1	*ns* t = 1	0.012 t = 0

t = number of ties

Conclusion

Replicating human experiments with artificial models often involves significant deviations in setup (Jacobs & Bates, 2018). In the current study, we partially address this issue by setting up a CNN experiment that preserves almost all aspects of an early human study on category learning (Posner and Keele, 1968).

Unlike the human results, we found no strong evidence of a prototype effect in CNNs. However, in AlexNet and DenseNet-121, unseen prototypes were classified more accurately than new exemplars, possibly suggesting a weak prototype bias. Overall, these preliminary findings hint that CNN behaviour in this specific setup may be best explained by a combination of prototype and exemplar theories.

In conclusion, these results lay the groundwork for further research, including exploring differences in category learning across CNN architectures and evaluating how well CNNs replicate other (more recent) category learning experiments under human-like experimental conditions.

Acknowledgements

This work has been supported by KU Leuven C1-project “Rethinking Replay: understanding and avoiding the stability gap in continual learning” (project code: 3E230465) and by a senior postdoctoral fellowship from the Research Foundation – Flanders (FWO) with grant number 1266823N.

References

Cabeza, R., Bruce, V., Kato, T., & Oda, M. (1999). The prototype effect in face recognition: Extension and limits. https://doi.org/10.3758/BF03201220

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778. https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition,* 4700-4708. https://openaccess.thecvf.com/content_cvpr_2017/papers/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.pdf

Jacobs, R. A., & Bates, C. J. (2018). Comparing the visual representations and performance of humans and deep neural networks. *Current Directions in Psychological Science*, *28*(1), 34–39. https://doi.org/10.1177/0963721418801342

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv*. https://doi.org/10.48550/arXiv.1412.6980

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*(3), 207-238. https://doi.org/10.1037/0033-295X.85.3.207

Minda, J. P., Roark, C. L., Kalra, P., & Cruz, A. (2024). Single and multiple systems in categorization and category learning. *Nature Reviews Psychology, 33*, 536-551. https://doi.org/10.1038/s44159-024-00336-7

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3), 353-363. https://doi.org/10.1037/h0025953

Singh, P., Peterson, J. C., Battleday, R. M., & Griffiths, T. L. (2020). End-to-end deep prototype and exemplar models for predicting human behavior. *arXiv*. https://doi.org/10.48550/arXiv.2007.08723