Inversion Effects in Humans and Deep Neural Networks

Samuel Sander[1], Katharina Dobs[2,3,4]

[1] Department Clinical Psychology, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany.

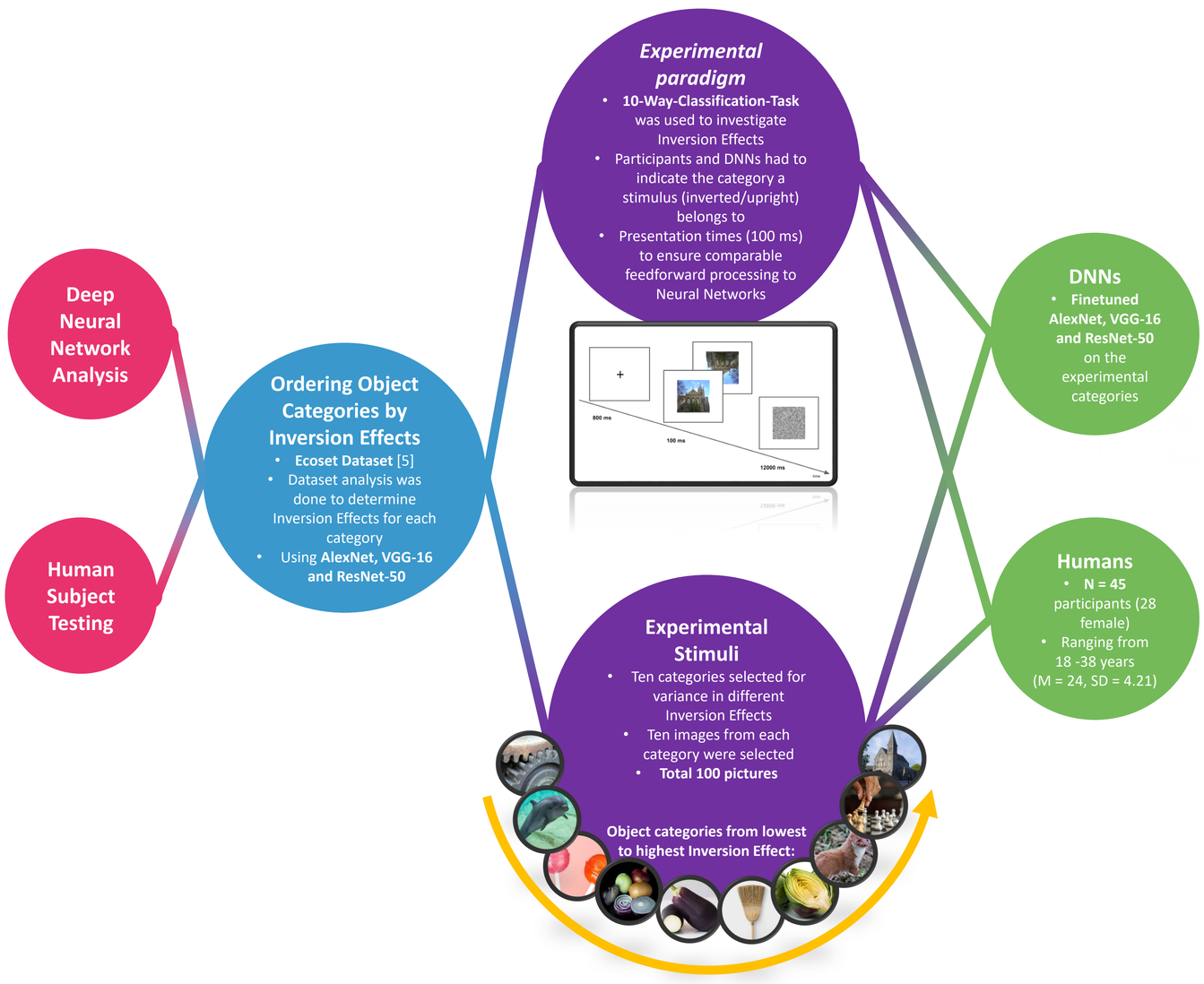
[2] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA.
[3] McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA.
[4] Department of Psychology, Justus Liebig University Giessen, Giessen, Germany.

Zentralinstitut für Seelische Gesundheit

Humans and Deep Neural Networks are prone to Inversion Effects

- Deep Neural Networks (DNNs) have achieved human-like performance in several visual perception tasks, including object classification, face verification, and medical image diagnosis. However, they exhibit sensitivity to variations in input data such as contrast, blur, or orientation.
- The “Inversion Effect" - the decreased recognition performance for objects (and specifically faces) when presented in an inverted orientation – will be compared in DNN and Human responses.
- **Goal:** Determine whether DNNs can predict the human inversion effect, identify the object properties that contribute to this effect, and assess the validity of DNNs as models for human object recognition [1] [2] [3] [4].

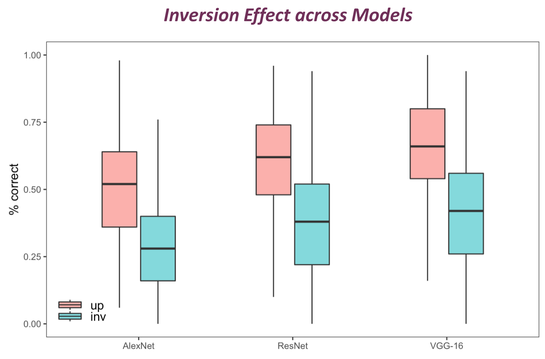
Methods



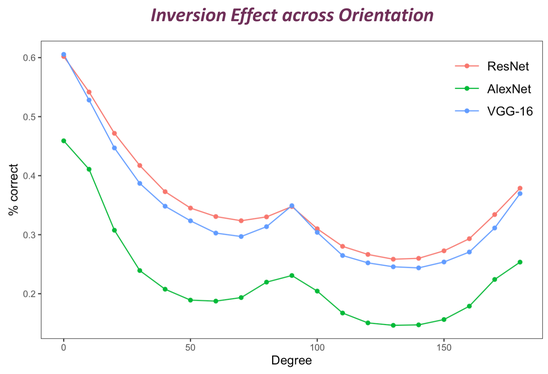
Results for Deep Neural Networks

Behavioral and Representational Inversion Effects in DNNs

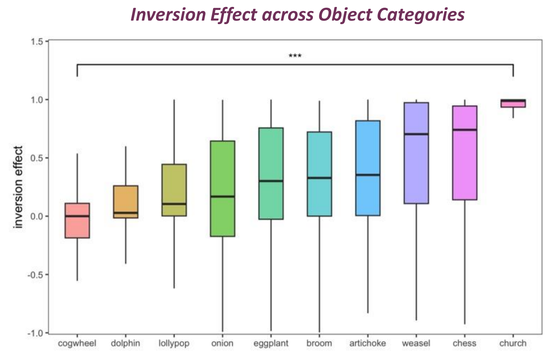
Inversion Effect across Models



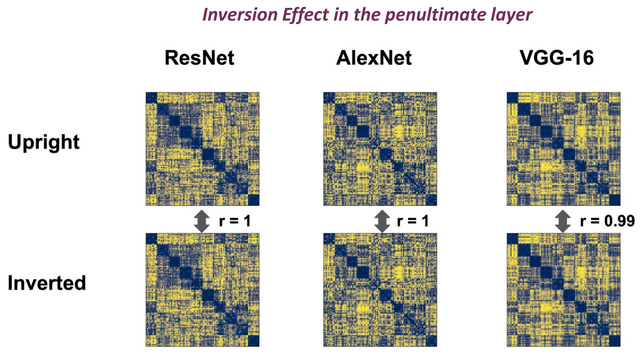
Inversion Effect across Orientation



Inversion Effect across Object Categories



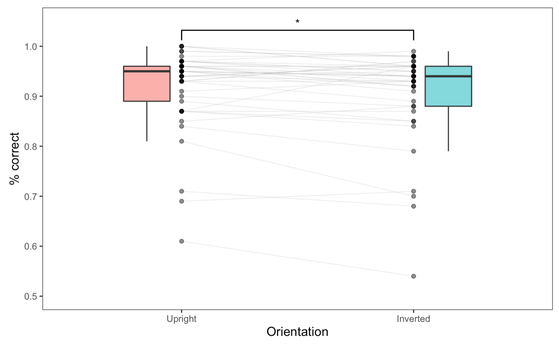
Inversion Effect in the penultimate layer



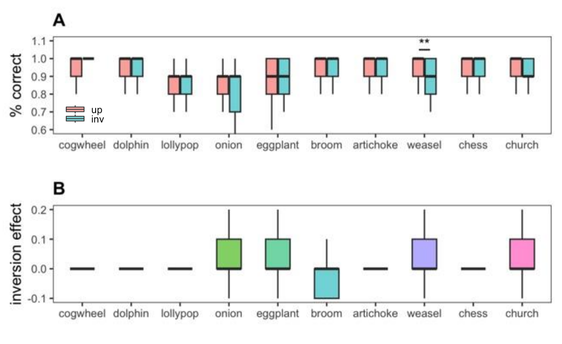
Results for Humans

Behavioral inversion Effects

Inversion Effect across Orientation

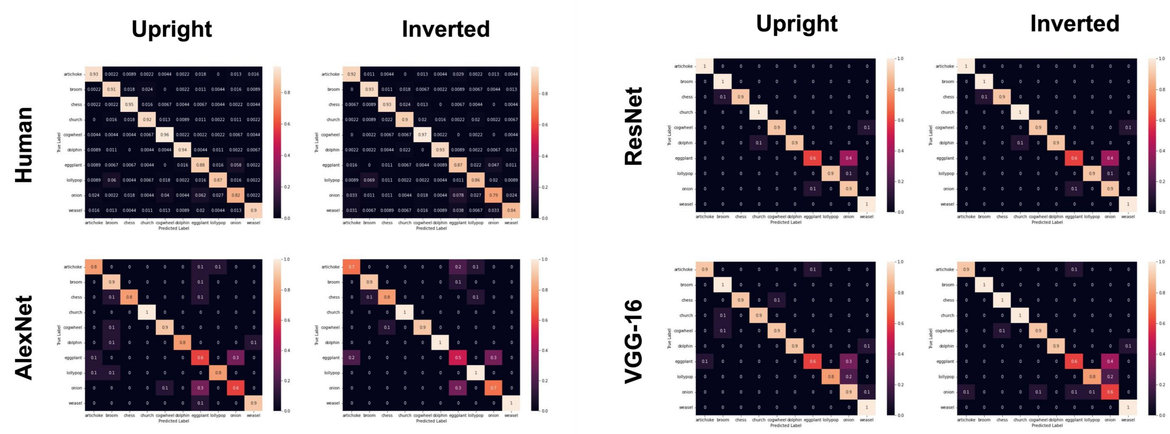


Inversion Effect across Object Categories



Humans vs. Deep Neural Networks

Confusion Matrices



Correlations between Human and DNN response patterns

Correlations for the Classifications

Correlations for the Confusion Matrices of all Responses to Upright Images

Confusion Matrix	1	2	3	4
Human	-			
ResNet	.980***	-		
AlexNet	.974***	.967***	-	
VGG-16	.983***	.981***	.972***	-

Note. *** indicates p < .001.

Correlations for the Confusion Matrices of all Responses to Inverted Images

Confusion Matrix	1	2	3	4
Human	-			
ResNet	.979***	-		
AlexNet	.980***	.971***	-	
VGG-16	.980***	.983***	.978***	-

Note. *** indicates p < .001.

Correlations for the Missclassifications

Correlations for the Confusion Matrices of the Misclassifications to Upright Images

Confusion Matrix	1	2	3	4
Human	-			
ResNet	.462***	-		
AlexNet	.173	.216*	-	
VGG-16	.470***	.666***	.385***	-

Note. * indicates p < .05. *** indicates p < .001.

Correlations for the Confusion Matrices of the Misclassifications to Inverted Images

Confusion Matrix	1	2	3	4
Human	-			
ResNet	.384***	-		
AlexNet	.344***	.216*	-	
VGG-16	.244*	.321**	.435***	-

Note. * indicates p < .05. *** indicates p < .001.

Discussion

- Neural Networks exhibit behavioral Inversion Effect but no differences in activations of the units in the penultimate layer
- The architecture of the model does not significantly influence the size of the Inversion Effect
- Humans do not exhibit significant differences in Inversion Effect across Object Categories
- Behavioral patterns between humans and neural networks regarding object classifications and Inversion Effects align

Literature

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[2] Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1701-1708).
[3] Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. arXiv preprint arXiv:1706.06969
[4] Yin, R. K. (1969). Looking at upside-down faces. Journal of experimental psychology, 81(1), 141.
[5] Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. Proceedings of the National Academy of Sciences, 118(8), e2011417118.

Contact samuel.sander@zi-mannheim.de