# EXTRAPOLATABLE RELATIONAL REASONING WITH COMPARATORS IN LOW-DIMENSIONAL MANIFOLDS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

While modern deep neural architectures generalise well when test data is sampled from the same distribution as training data, they fail badly for cases when the test data distribution differs from the training distribution even along a few dimensions. This lack of out-of-distribution generalisation is increasingly manifested when the tasks become more abstract and complex, such as in relational reasoning. In contrast, human brain is observed to generalise better to unseen inputs (Geirhos et al., 2018), and typically requires only a small number of training samples. Researchers (Spelke & Kinzler, 2007; Chollet, 2019; Battaglia et al., 2018; Xu et al., 2020) argue that the human brain developed special inductive biases that adapt to the form of information processing needed for humans, thereby improving generalisation.

In this paper we propose a neuroscience-inspired inductively biased module that can be readily amalgamated with current neural network architectures to improve out-of-distribution (o.o.d) generalisation performance on relational reasoning tasks. This module learns to project high-dimensional object representations to low-dimensional manifolds for more efficient and generalisable relational comparisons. We show that neural nets with this inductive bias achieve considerably better o.o.d generalisation performance for a range of relational reasoning tasks including visual object comparisons and Raven Progressive Matrices Reasoning test, and thus more closely models human ability to generalise even when no previous examples from that domain exist. Finally, we analyse the proposed inductive bias module to understand the importance of lower dimensional projection, and propose an augmentation to the algorithmic alignment theory to better measure algorithmic alignment with generalisation.

## REFERENCES

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

François Chollet. On the measure of intelligence, 2019.

Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 7538–7550, 2018.

Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

Keyulu Xu, Jingling Li, Mozhi Zhang, Simon S. Du, Ken ichi Kawarabayashi, and Stefanie Jegelka. What can neural networks reason about? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJxbJeHFPS.