Scrutinizing Physics with Neural Stethoscopes

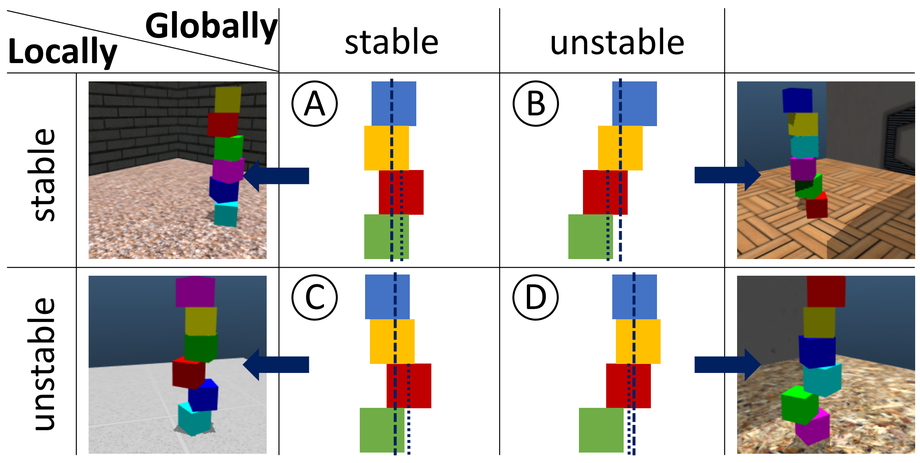
Fabian Fuchs[1], Oliver Groth[1], Adam Kosiorek[1], Alex Bewley[12],
Markus Wulfmeier[13], Andrea Vedaldi[1], Ingmar Posner[1]
[1]Oxford University, [2]Google Research, [3]Google DeepMind

APPLIED ARTIFICIAL INTELLIGENCE LAB
OXFORD ROBOTICS INSTITUTE

Overview

- **Querying** physical understanding of **intuitive physics** models
- **Guiding** the **learning** process in presence of **misleading visual clues**
- Introducing **Neural Stethoscopes** as unifying framework for **querying**, **promoting** and **suppressing** information in latent representations

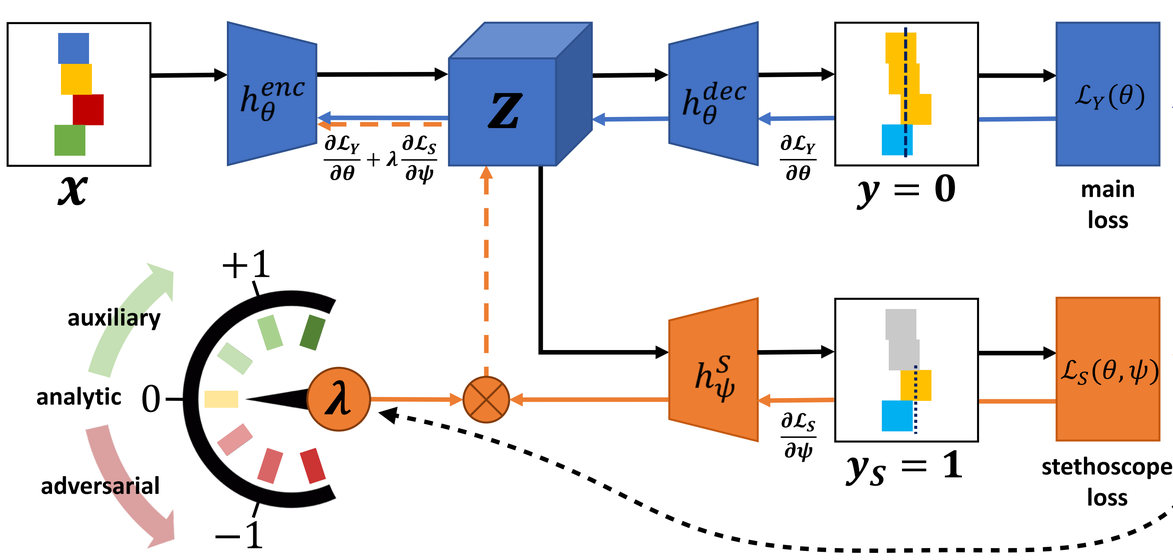
Dataset: Stability Prediction and Visual Cues[1]



Two Labels:

- **global stability**
  Will the tower fall over?
- **local stability**
  Does the tower look crooked?

Stethoscopes: Unifying Adversarial Learning, Auxiliary Learning and Interpretability



The network is trained on two classification tasks:

Main task: in this case (global) stability prediction.

Secondary task: in this case either local stability or origin of global instability.

The value chosen for hyper parameter **λ** determines whether second task is seen as **adversarial**, **auxiliary**, or purely **analytic.**

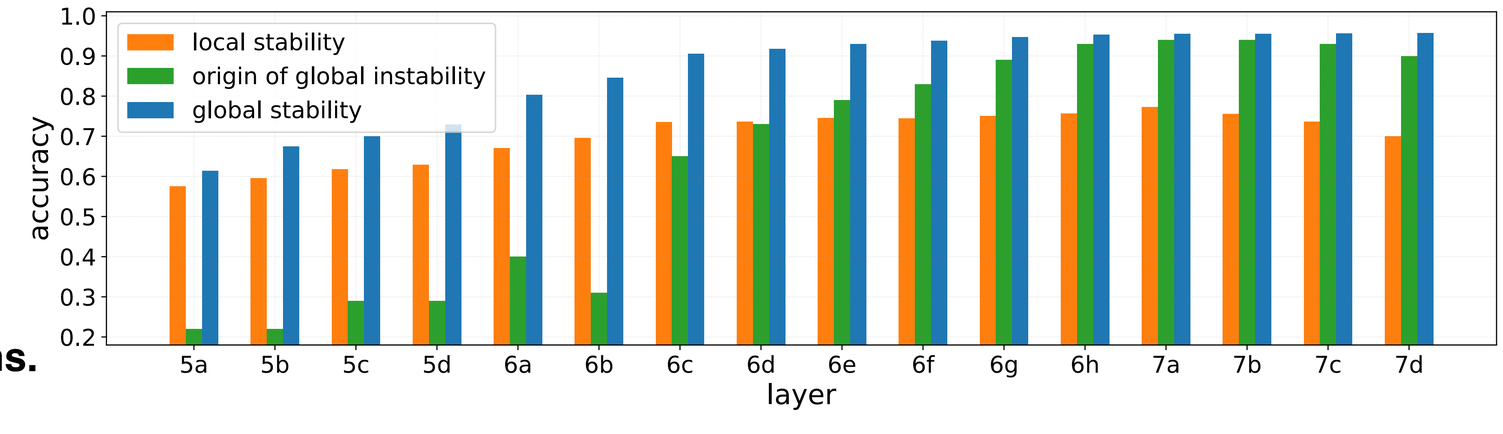
$$\mathscr{L}_{y,s}(\theta, \psi) = \mathscr{L}_y(\theta) + \lambda \cdot \mathscr{L}_s(\theta, \psi)$$

Results: Querying, Suppressing & Promoting Information in Latent Representations

Analytic Mode (Querying, $\lambda = 0$)

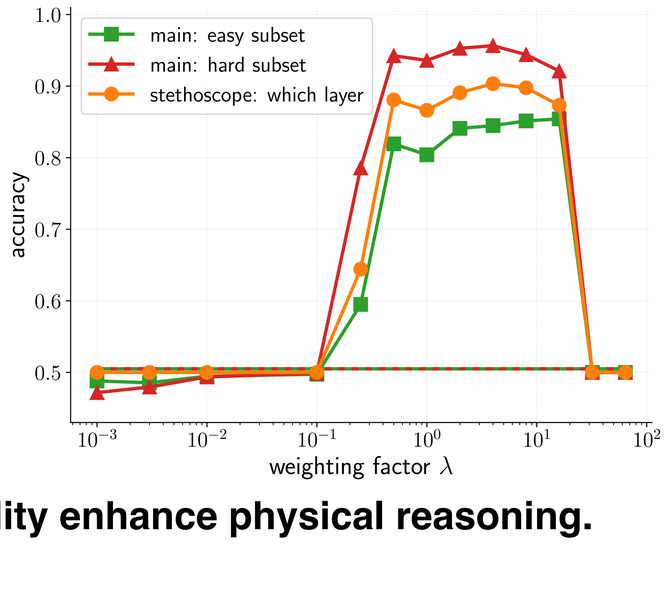
- What information does the network extract?
- One Stethoscope is attached per network layer.
- This helps to understand the learning process.
- Analysis shown is for the Inception v4 network.

⇒ **Visual cues influence learned representations.**



Auxiliary Mode (Promoting Information, $\lambda > 0$)

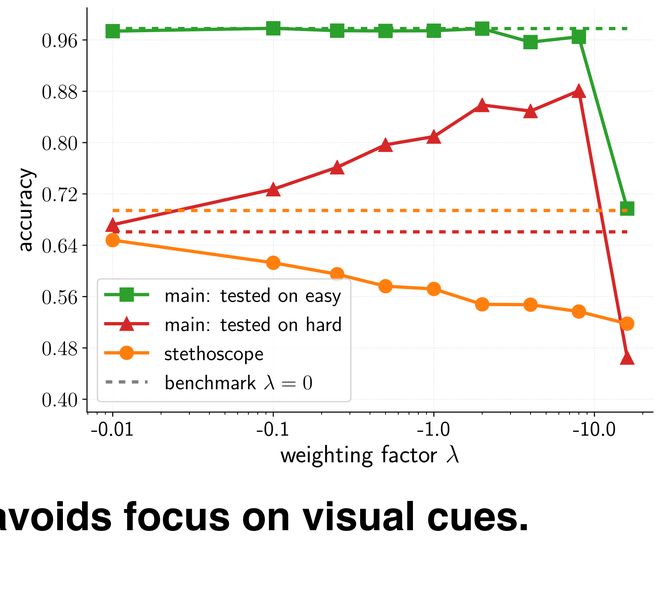
- Training on a particularly hard dataset leads the baseline algorithm to fail.
- Using 'origin of global stability' as a secondary label helps network by promoting helpful features.



⇒ **Labels for origin of stability enhance physical reasoning.**

Adversarial Mode (Suppressing Information, $\lambda < 0$)

- Training on a simple dataset leads to bias for visual cues (local stability).
- Suppressing extraction of respective information de-biases the network and leads to performance gains.



⇒ **Adversarial training avoids focus on visual cues.**

Acknowledgements

This research was funded by the EPSRC AIMS Centre for Doctoral Training at Oxford University, the EPSRC under Programme Grant DFR01420 and the European Research Council under grant ERC 677195-IDIU. We acknowledge use of Hartree Centre resources in this work. The STFC Hartree Centre is a research collaboratory in association with IBM providing High Performance Computing platforms funded by the UK's investment in e-Infrastructure. The Centre aims to develop and demonstrate next generation software, optimised to take advantage of the move towards exa-scale computing.

[1] The Dataset is an extension of: O. Groth, F. Fuchs, A. Vedaldi, I. Posner, *ShapeStacks: Learning Vision-Based Physical Intuition for Generalised Object Stacking*. *ECCV, 2018*.