Learning Physics with Neural Stethoscopes

F.B. Fuchs O. Groth A.R. Kosiorek A. Bewley M. Wulfmeier A. Vedaldi I. Posner
Department of Engineering Science at the University of Oxford
`{fabian,ogroth,adamk,bewley,markus,vedaldi,ingmar}@robots.ox.ac.uk`

Abstract

Model interpretability and systematic, targeted model adaptation present central challenges in deep learning. In the domain of intuitive physics, we study the task of visually predicting stability of block towers with the goal of understanding and influencing the model's reasoning. Our contributions are two-fold. Firstly, we introduce *neural stethoscopes* as a framework for quantifying the degree of importance of specific factors of influence in deep networks as well as for actively promoting and suppressing information as appropriate. In doing so, we unify concepts from training with auxiliary and adversarial losses. Secondly, we deploy the stethoscope framework to provide an in-depth analysis of a state-of-the-art neural network for stability prediction, specifically examining its physical reasoning.

Previous work has shown that neural networks are highly capable of learning physical tasks such as stability prediction. However, unlike approaches using physics simulators [Furrer et al., 2017, Wu et al., 2017], learning based approaches pose a challenge for model interpretability: Did the model gain a sound understanding of the physical principles or does it take short-cuts following visual cues based on correlations in the data? Occlusion-based attention analyses are a first step in this direction, but insights gained from this are limited [Lerer et al., 2016, Groth et al., 2018]. In this work we introduce stethoscopes to enhance interpretability and influence the learning process on the task of stability prediction, but present it in the following as a general framework which can be applied to any set of tasks.

1 Methodology: Neural Stethoscopes

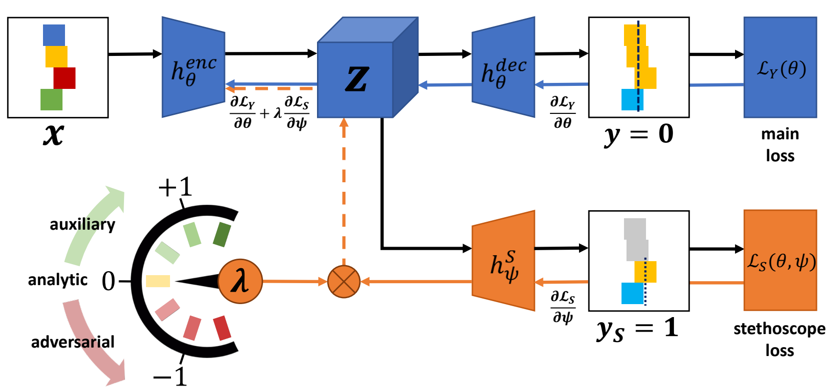


Figure 1: The stethoscope framework. The main network (blue), comprised of encoder and decoder, is trained for global stability prediction of block towers. The stethoscope (orange) is trained to predict a nuisance parameter (local stability) with input is Z, a learned feature from an arbitrary layer of the main network. The stethoscope loss is back-propagated with weighting factor λ to the main network.

Preprint.

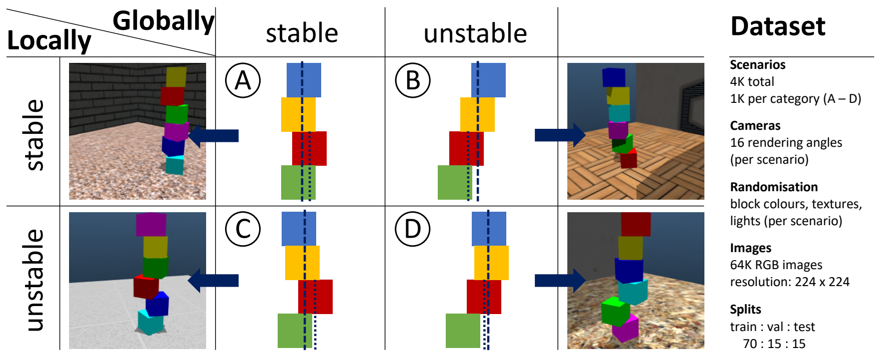


Figure 2: We have four qualitative scenarios (A-D) combining global and local (in)stability. The dashed line shows the projection of the cumulative centre of mass of the upper tower (red, yellow and blue block) whereas the dotted line depicts the projection of the local centre of mass of the red block. A tower is globally stable, if and only if the global centre of mass is always supported whereas the individual local centre of masses are not indicative of global structure stability. Global and local centre of masses for the green, yellow and blue block have been omitted for clarity of presentation.

In supervised deep learning, we typically look for a function $f_\theta : X \to Y$ with parameters θ that maps an input $\boldsymbol{x} \in X$ to its target $\boldsymbol{y} \in Y$. Without loss of generality, we rewrite f_θ as the composition of the encoder $h_\theta^{\text{enc}} : X \to Z$, which maps the input to an intermediate representations $\boldsymbol{z} \in Z$ $\boldsymbol{z} \in Z$, and the decoder $h_\theta^{\text{dec}} : Z \to Y$, which maps features to the output. Let the stethoscope be defined as an arbitrary function $h_\psi^{\text{s}} : Z \to S$ with parameters ψ. We define two loss functions: $\mathcal{L}_y(\theta)$, which measures the discrepancy between predictions f_θ and the true task y and $\mathcal{L}_s(\theta, \psi)$, which measures the performance on the supplemental task (see Figure 1). The weights of the stethoscope are updated as $-\Delta\psi \propto \nabla_\psi \mathcal{L}_s(\theta, \psi)$ to minimise $\mathcal{L}_s(\theta, \psi)$ and the weights of the main network as $-\Delta\theta \propto \nabla_\theta \mathcal{L}_{y,s}(\theta, \psi)$ to minimise the energy

$$\mathcal{L}_{y,s}(\theta, \psi) = \mathcal{L}_y(\theta) + \lambda \cdot \mathcal{L}_s(\theta, \psi). \tag{1}$$

By choosing different values for the constant λ we obtain three very different use cases:

Analytic Stethoscope ($\lambda = 0$) Here, the gradients of the stethoscope, which acts as a passive observer, are not used to alter the main model. This setup can be used to interrogate learned feature representations: if the stethoscope predictions are accurate, the features can be used to solve the task.

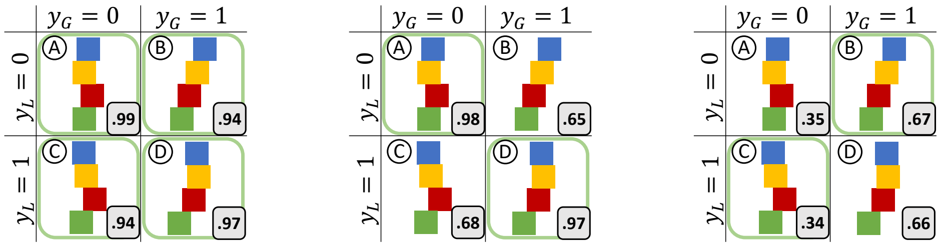
Auxiliary Stethoscope ($\lambda > 0$) The encoder is trained with respect to the stethoscope objective, hence enforcing correlation between main network and supplemental task. This setup is related to learning with auxiliary tasks, and helpful if we expect the two tasks to be beneficially related.

Adversarial Stethoscope ($\lambda < 0$) By setting $\lambda < 0$, we train the encoder to maximise the stethoscope loss (which the stethoscope *still* tries to minimise), thus encouraging independence between main network and supplemental tasks. This is effectively an adversarial training framework and is useful if features required to solve the stethoscope task are a detrimental nuisance factor.

In auxiliary and adversarial mode, we attach the stethoscope to the main network's last layer before the logits in a fully connected manner. This setup proved to have the highest impact on the learning process of the main network. The stethoscope itself is implemented as a two-layer perceptron with ReLU activation and trained with sigmoid or softmax cross-entropy loss on its task $\mathcal{S}$.

2 Vision-Based Stability Prediction of Block Towers

We follow the state-of-art approach on visual stability prediction of block towers and examine as well as influence its learning behaviour. We introduce a variation of the ShapeStacks dataset from Groth et al. [2018] which is particularly suited to study the dependence of network predictions on visual cues. We then examine how suppressing or promoting the extraction of certain features influences the performance of the network using neural stethoscopes. We choose the Inception-v4 network [Szegedy et al., 2017] as it yields state-of-the-art performance on stability prediction [Groth et al., 2018].



(a) Trained on All: $\varnothing_{acc} = 0.96$ (b) Trained on *Easy*: $\varnothing_{acc} = 0.82$ (c) Trained on *Hard*: $\varnothing_{acc} = 0.51$

Figure 3: The influence of local instability on global stability prediction. In setup (a) we train on all 4 tower categories (indicated by green frames). Global stability prediction accuracies on per-category test splits are reported in the bottom right grey boxes. In (b) we train solely on easy scenarios (A & D) where global and local stability are positively correlated. In (c) we only present hard scenarios during training featuring a negative correlation between global and local stability. The performance differences clearly show that local stability influences the network's prediction for global stability.

Dataset As shown in Groth et al. [2018], a single-stranded tower of blocks is stable if, and only if, at every interface between two blocks the centre of mass of the entire tower above is supported by the convex hull of the contact area. If a tower satisfies this criterion, *i.e.*, it does not collapse, we call it *globally stable*. To be able to quantitatively assess how much the algorithm follows visual cues, we introduce a second label: We call a tower *locally stable* if, and only if, at every interface between two blocks, the centre of mass of the block immediately above is supported by the convex hull of the contact area. Intuitively, this measure describes, if taken on its own without any blocks above, each block would be stable. We associate binary prediction tasks y_G and y_L to respective global and local stability where label $\boldsymbol{y} = 0$ indicates *stability* and $\boldsymbol{y} = 1$ *instability*. Global and local instability are neither mutually necessary nor sufficient, but can easily be confused visually which is demonstrated by our experimental results. We create a simulated dataset[1] with 4,000 block tower scenarios divided into four qualitative categories (cf. Figure 2). The dataset is divided into an *easy* subset, where local and global stability are always positively correlated, and a *hard* subset, where this correlation is always negative. The dataset will be made available online.

Local Stability as a Visual Cue Based on the four categories of scenarios described in Figure 2, we conduct an initial set of experiments to gauge the influence of local stability on the network predictions. Figure 3 shows a strong influence of local stability on the prediction performance. When trained on the entire, balanced data set, the error rate is three times higher for *hard* than for *easy* scenarios (6% vs. 2%). When trained on *easy* scenarios only, the error rate even differs by a factor of 13. Trained on *hard* scenarios only, the average performance across all four categories is on the level of random chance (51%), indicating that negatively correlated local and global stability imposes a much harder challenge on the network.

3 Using Neural Stethoscopes to Guide the Learning Process

After demonstrating the influence of local stability on the task of global stability prediction we turn our attention to the use of neural stethoscopes to quantify and actively mitigate this influence.

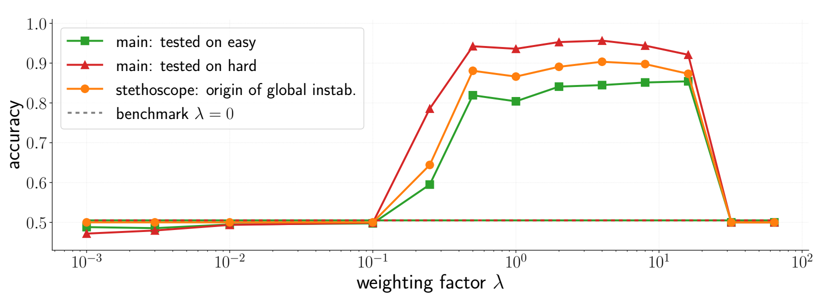
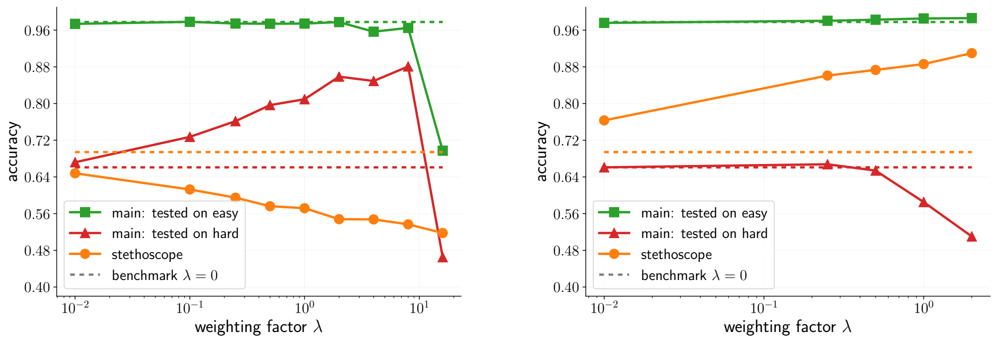


Figure 4: Promoting complementary feature extraction with auxiliary stethoscopes.

Promotion of Complementary Information We test the hypothesis that fine-grained labels of instability locations help the main network to grasp the correct physical concepts. To that end, we

[1]We use the MuJoCo physics engine [Todorov et al., 2012] for rendering and stability checking.



(a) Adversarial Stethoscope (b) Auxiliary Stethoscope

Figure 5: Successful debiasing by suppressing a nuisance factor with adversarial training.

consider the setup from Figure 3c where the training data only consists of *hard* scenarios with a baseline performance of 51%. The main network is trained on global stability while the stethoscope is trained on predicting the origin of global instability, namely the interface at which the instability occurs. Figure 4 shows that auxiliary training substantially improves the performance for weighting parameters $\lambda \in [0.5, 16]$. However, for very small values of λ, the contribution of the additional loss term is too small while for large values, performance deteriorates to the level of random chance as a result of the primary task being far out-weighted by the auxiliary task.

Suppression of Nuisance Information Results from Figure 3 indicate that the network might use local stability as a visual cue to make biased assumptions about global stability. We now investigate whether it is possible to debias the network by forcing it to pay less attention to local stability. To that end, we focus on the scenario shown in Figure 3b, where we only train the network on global stability labels for *easy* scenarios. As shown in Figure 3b, the performance of the network suffers significantly when tested on *hard* scenarios where local and global stability labels are inversely correlated.

The hypothesis is that forcing the network not to focus on local stability weakens this bias. In Figure 5, we use active stethoscopes ($\lambda \neq 0$) to test this hypothesis. We train a stethoscope on local stability on labels of all categories (in a hypothetical scenario where local labels are easier to obtain than global labels) and use both the adversarial and the auxiliary setup in order to test the influence of suppressing and promoting accessibility of information relevant for local stability in the encoded representation, respectively. In Figure 5, the results of both adversarial and auxiliary training are compared to the baseline of $\lambda = 0$, which is equivalent to the analytic stethoscope setup.

Figure 5a shows that adversarial training does indeed partly remove the bias and significantly increases the performance on *hard* scenarios while maintaining its high accuracy on *easy* scenarios. With an increasing magnitude of λ, we observe a monotonic reduction in bias up to a point where further increasing λ jeopardises the performance on the main task as the encoder puts more focus on confusing the stethoscope than on the main task (in our experiments this happens at $\lambda \approx 10^1$).

This scenario could also be seen from the perspective of feeding additional information into the network, which could profit from more diverse training data. However, Figure 5b shows that naively using an auxiliary setup to train the network on local stability worsens the bias. With increasing λ and increasing performance of the stethoscope, performance slightly improves on *easy* scenarios while accuracy deteriorates on *hard* scenarios. Auxiliary training on local stability further shifts the focus to local features. When tested on *hard* scenarios, where local and global stability are inversely correlated, the network will therefore perform worse when it has learned to rely on local features.

References

F. Furrer, M. Wermelinger, H. Yoshida, F. Gramazio, M. Kohler, R. Siegwart, and M. Hutter. Autonomous robotic stone stacking with online next best object target pose planning. *ICRA*, 2017.

O. Groth, F. Fuchs, I. Posner, and A. Vedaldi. Shapestacks: Learning vision-based physical intuition for generalised object stacking. *arXiv*, 1804.08018, 2018.

A. Lerer, S. Gross, and R. Fergus. Learning physical intuition of block towers by example. In *ICML*, 2016.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. *IROS*, 2012.

J. Wu, E. Lu, P. Kohli, B. Freeman, and J. Tenenbaum. Learning to see physics via visual de-animation. 2017.

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.