PHYSICS-INSPIRED INTERPRETABILITY OF MACHINE LEARNING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The ability to explain decisions made by machine learning models remains one of the most significant hurdles towards widespread adoption of AI in highly sensitive areas such as medicine, cybersecurity or autonomous driving. Great interest exists in understanding which features of the input data prompt model decision making. In this contribution, we propose a novel approach to identify relevant features of the input data, inspired by methods from the energy landscapes field, developed in the physical sciences. By identifying conserved weights within groups of minima of the loss function landscapes, we can identify the drivers of model decision making. Analogues to this idea exist in the molecular sciences, where coordinate invariants or order parameters are employed to identify critical features of a molecule. However, no such approach exists for machine learning loss function landscapes. We will demonstrate the applicability of energy landscape methods to machine learning models and give examples, both synthetic and from the real world, for how these methods can help to make models more interpretable.

1 INTRODUCTION

Machine learning methods have achieved impressive results in recent years. Besides famous applications in areas like chess (Silver et al., 2017a) and Go (Silver et al., 2017b), AI plays a critical role in advances to autonomous driving (Grigorescu et al., 2020), protein structure prediction (Jumper et al., 2021), cancer identification (Sammut et al., 2022) and in cybersecurity (Dasgupta et al., 2022). However, in order for AI methods to take the next step and be commonly employed for critical applications without any humans in the loop, we want to be able to understand the decision making process. A critical component towards explainable AI is understanding which parts of the input data are utilised by the model in its decision making. In neural networks, the most popular approach is to study the outgoing weights and gradients from an individual input node. Larger weights are reasonably assumed to indicate a greater significance of the particular input, and indeed, an entire class of interpretability metrics, namely gradient-based methods, are founded on this idea (Simonyan et al., 2013; Linardatos et al., 2020). Yet, given the immense complexity of overparameterised, deep neural networks, current methods are in practice often insufficient to appropriately explain a model. Using methods from the physical sciences, we propose a novel approach as a next step towards interpretable neural networks.

1.1 ENERGY LANDSCAPES

In the physical sciences, energy landscapes (ELs) are employed to explore molecular configuration space (Wales et al., 1998; 2003). Each molecular configuration is associated with an energy value, and local minima of the energy landscape represent stable isomers. The analogy to machine learning loss function landscapes (ML LFLs) is straightforward, the main difference perhaps being that non-minima are valid configurations for sets of weights. Due to this similarity between ELs and ML LFLs, various, well-established methods from the field of energy landscapes can be employed to study ML LFLs. One key area of interest here is interpretability. Employing well-understood methods from a mature field, with a solid mathematical basis in the physical world, to move away from black-box machine learning models may be a helpful step towards interpretable machine learning models.

1.2 Related work

Various approaches to interpretability in deep learning for neural networks exist. Below, we are mostly interested in gradient-based methods due to their applicability to non-image data. Various other methods to interpret the output of CNNs on images exist, as for example summarised in Linardatos et al. (2020), but will not be reviewed below.

Gradient-based methods: All gradient-based methods are concerned with changes in the prediction as the input data is slightly perturbed. For a vector-valued input $\mathbf{x} \subset \mathrm{X} \in \mathbb{R}^d$ and some loss function, $\mathcal{L}$, a gradient-based method computes some expression of the form $\partial\mathcal{L}/\partial\mathbf{x}$, usually for each input node individually. Gradient-based methods were first introduced for images by Simonyan et al. (2013), who used them to compute how changes in the input affect predictions in the neighbourhood of the input, allowing the computation of a salience map (Kümmerer et al., 2014; Zhao et al., 2015). More recently, integrated gradient methods (Sundararajan et al., 2017) consider the derivative of the output (loss) with respect to individual input nodes. If the change in loss is large with respect to some input feature, that feature is more likely to be relevant to the decision making. Various other gradient and perturbation based methods exist (Alvarez-Melis & Jaakkola, 2018), yet their usefulness and accuracy is debated, and is generally agreed to be insufficient (Srinivas & Fleuret, 2020).

Energy landscapes in machine learning: Energy landscapes methods have been employed to study machine learning in previous contributions (Ballard et al., 2017; Chitturi et al., 2020). Niroomand et al. (2022) used energy landscapes to characterise new loss functions, and the landscapes view has been used more broadly to gain insights into machine learning models (Segura et al., 2022; Verpoort et al., 2020). Lastly, other applications of energy landscape methods have employed various concepts from physical sciences in machine learning, including the heat capacity (Bradley et al., 2022; Niroomand et al., 2022), both for characterisation and model improvement.

Interpreting energy landscapes: Due to the associated physical meaning, energy landscapes are usually more easily interpretable. Only minima represent equilibrium configurations, and each minimum is associated with a unique structure. However, for larger, complex molecules, many minima may exist, and enumerating them may be infeasible. Instead, common features between sets of minima, grouped by their energetic properties, may be identified. For example, in (Röder et al., 2020) and (Röder & Wales, 2022) a multi-funnelled landscape is analysed to understand which structural differences of a molecule characterise solutions in a specific funnel.

2 Energy landscapes methods

The study of energy landscapes is a well-established field (Wales et al., 1998). Various approaches exist for constructing a faithful representation of the landscape by optimising the non-convex energy function, and visualising this landscape. Visualisation is commonly performed using disconnectivity graphs (Becker & Karplus, 1997; Wales et al., 1998) as described below.

2.1 Landscape visualisation

A disconnectivity graph is a low-dimensional representation of a complex function landscape, which reduces the function to key characteristic stationary points, namely minima and transition states. A transition state is an index-1 saddle point of the funciton. The vertical axis of a disconnectivity graph represents the energy or loss value, and ordering along the horizontal axis is arbitrary. To identify distinct groups of minima, usually called funnels, we introduce the notion of levels and nodes. Levels are cross-sections of the energy at some evenly spaced, discrete heights in the disconnectivity graph. The highest energy level in the disconnectivity graph is level 1, and the lowest corresponds to the global minimum. Thus, each minimum belongs to one of evenly spaced intervals. Within each level, minima are grouped by a shared parent node, located higher up. In the disconnectivity graphs below, levels and nodes are represented as level_node.

In the molecular sciences, a transition state between two minima describes the energy barrier to be overcome for a molecule to change configuration from one state to the other. This particular notion does not have a direct meaning in machine learning. However, given the optimisation procedure required for model training, the concept of a transition state is highly relevant, since it may determine which minimum basin the optimiser will fall into. Thus, disconnectivity graphs can be employed as a faithful coarse-grained representation of the loss function landscape. In particular, it will be

relevant below to understand that any group of minima close together, perhaps separated from other groups of minima via high-lying transition states, may share commonalities. This effect has been observed in (Röder et al., 2020; Röder & Wales, 2022) for molecular systems, and we find that the same argument holds for ML-LFLs.

3 EXPERIMENTS

We report results for two separate experiments on two datasets. We believe that the underlying idea applies without loss of generality to any neural network architecture. However, further work will be required to validate this suggestion. Figures 1and 2 show disconnectivity graphs for (1) a 2-dimensional synthetic checkerboard dataset (Kluger et al., 2003) and (2) an anonymised, 29-dimensional credit card fraud detection dataset (Dal Pozzolo et al., 2015), which are binary classification problems. The lowest lying node in each graph is the global minimum. To identify groups of minima with conserved weights, we follow a two-step procedure. Firstly, we identify groups of minima, that are separated from other groups by a higher-lying transition state. This segregation leads to the notion of nodes and levels described above. Secondly, we identify groups of minima that share a subset of conserved weights by computing the standard deviation of each weight across each node in each level. A subset of weights $\tilde{w} \subset W$ is conserved if $\sigma(w) < n$ for any $w \in W$, where W denotes the weights of all minima in one node of one level.

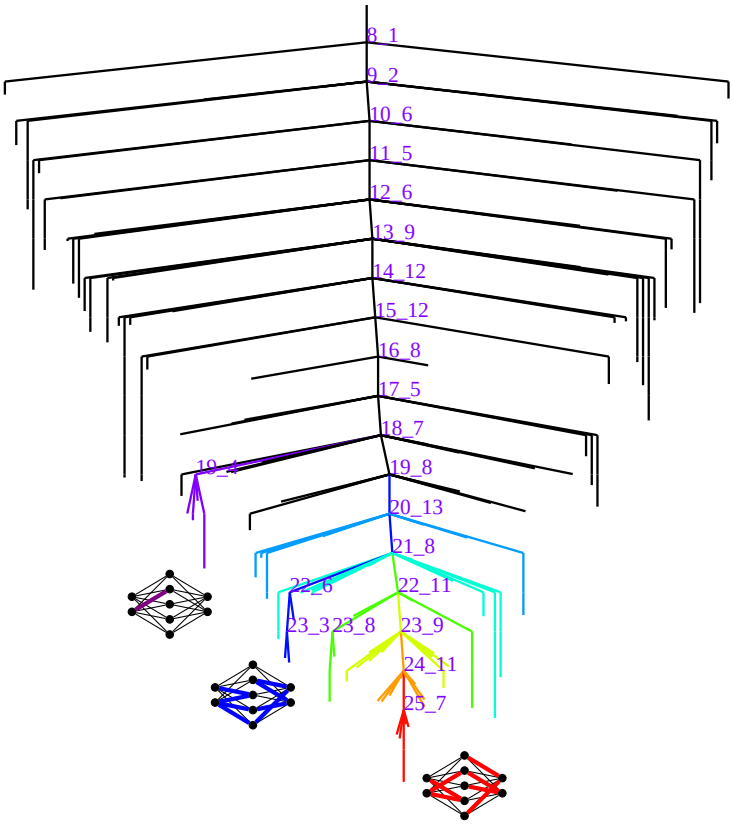

Figure 1: Disconnectivity graph for the checkerboard dataset. The conserved weights for a specific local minimum are highlighted in the respective colour for the chosen examples.

For visualisation purposes, we employ single-layer neural networks, which is sufficient for our analysis, with only a few nodes. The AUC of the best solutions is > 0.95 for both problems. Hence, these networks provide a realistic solutions to the set problems. In both figures, we visualised the conserved weights for a group of minima in the corresponding colour. In Figure 1, various weights across the network are conserved, highlighting how this approach identifies relevant weights for the model. In Figure 2, the funnel containing the global minimum (red) conserves 3 weights, all related

to one specific input node. Randomly permuting the 3 identified weights for the group of minima around the global minimum in figure 2 reduces the best AUC from ≈ 0.95 to 0.76. In contrast, permuting any random set of 3 weights by the same magnitude on average only decreases the best AUC by 0.05 to an average best AUC of ≈ 0.9. In 2, for group 25_7 (red), weights for only a single minimum are conserved, in group 22_6 (blue), weights outgoing from different input nodes are conserved.

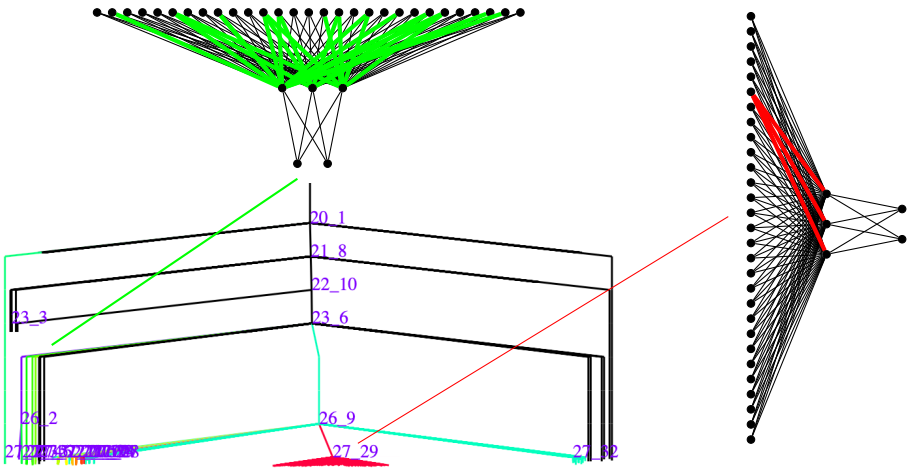

Figure 2: Disconnectivity graph for credit card data. Group 25_7 in red includes the global minimum. Coloured edges indicate that for all minima in the specific group, these particular weights are conserved, i.e. have a standard deviation $< n$ which has been set to $n = 0.01$ here.

3.1 Permutational invariance groups

As discussed in Niroomand et al. (2022), the magnitude of individual weights must always be viewed with caution due to permutational isomers. For a given neural network of H hidden layers, with n_l nodes in hidden layer l, there exist at least $|\mathcal{G}| = \prod_{l=1}^{H} (n_l! \times 2^{n_l})$ sets of weights that are invariant with respect to the model prediction. This effect must be considered when identifying conserved weights; for example, a negative inverse could still be valid and conserved (Niroomand et al., 2022). We account for this effect by identifying permutationally invariant sets of weights and only considering a single minimum $m \in \mathcal{G}$ for each $\mathcal{G}$.

4 Discussion and conclusions

Well-established methods from computational chemical physics can be employed to enhance our understanding of machine learning systems. In this work, we have shown how both concepts and associated tools from the study of energy landscapes can be employed for ML LFLs to guide interpretability. We have shown that groups of minima share conserved weights and importantly, that these weights are critical to model performance. Randomly permuting the conserved weights strongly decreases model performance, much more so than permuting any other random set of weights $\mathcal{S}$ of equivalent cardinality $|\mathcal{S}|$. Figure 2 indicates that all the conserved weights are associated with the particular input node 6. Since the credit card dataset is anonymised and PCA-reduced (Dal Pozzolo et al., 2015), we are unable to say which specific feature it is that helps the model in making a decision, but we can say where it can be found. In Figure 1, we know that both input nodes are relevant, which is confirmed by studying the conserved weights for the three given examples. Importantly, different weights are conserved across different examples, highlighting the importance of studying the loss function landscape. Studying the applicability of our method to larger and more complex architectures, and perhaps also to different types of machine leaning models, will provide valuable insights, and is an interesting direction for future work.

REFERENCES

David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Andrew J Ballard, Ritankar Das, Stefano Martiniani, Dhagash Mehta, Levent Sagun, Jacob D Stevenson, and David J Wales. Energy landscapes for machine learning. *Physical Chemistry Chemical Physics*, 19(20):12585–12603, 2017.

Oren M Becker and Martin Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *The Journal of chemical physics*, 106 (4):1495–1517, 1997.

Arwen V Bradley, Carlos A Gomez-Uribe, and Manish Reddy Vuyyuru. Shift-curvature, sgd, and generalization. *Machine Learning: Science and Technology*, 3(4):045002, 2022.

Sathya R Chitturi, Philipp C Verpoort, David J Wales, et al. Perspective: new insights from loss function landscapes of neural networks. *Machine Learning: Science and Technology*, 1(2):023002, 2020.

Andrea Dal Pozzolo, Olivier Caelen, Reid A Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE symposium series on computational intelligence*, pp. 159–166. IEEE, 2015.

Dipankar Dasgupta, Zahid Akhtar, and Sajib Sen. Machine learning in cybersecurity: a comprehensive survey. *The Journal of Defense Modeling and Simulation*, 19(1):57–106, 2022.

Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Yuval Kluger, Ronen Basri, Joseph T Chang, and Mark Gerstein. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome research*, 13(4):703–716, 2003.

Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

Maximilian P Niroomand, John WR Morgan, Conor T Cafolla, and David J Wales. On the capacity and superposition of minima in neural network loss function landscapes. *Machine Learning: Science and Technology*, 3(2):025004, 2022.

Konstantin Röder and David J Wales. The energy landscape perspective: Encoding structure and function for biomolecules. *Frontiers in Molecular Biosciences*, 9, 2022.

Konstantin Röder, Guillaume Stirnemann, Anne-Catherine Dock-Bregeon, David J Wales, and Samuela Pasquali. Structural transitions in the rna 7sk 5 hairpin and their effect on hexim binding. *Nucleic acids research*, 48(1):373–389, 2020.

Stephen-John Sammut, Mireia Crispin-Ortuzar, Suet-Feung Chin, Elena Provenzano, Helen A Bardwell, Wenxin Ma, Wei Cope, Ali Dariush, Sarah-Jane Dawson, Jean E Abraham, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*, 601(7894):623–629, 2022.

Carolina Herrera Segura, Edison Montoya, and Diego Tapias. Subaging in underparametrized deep neural networks. *Machine Learning: Science and Technology*, 3(3):035013, 2022.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017a.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017b.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Suraj Srinivas and François Fleuret. Rethinking the role of gradient-based attribution methods for model interpretability. *arXiv preprint arXiv:2006.09128*, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Philipp C Verpoort, Alpha A Lee, and David J Wales. Archetypal landscapes for deep neural networks. *Proceedings of the National Academy of Sciences*, 117(36):21857–21864, 2020.

David J Wales, Mark A Miller, and Tiffany R Walsh. Archetypal energy landscapes. *Nature*, 394 (6695):758–760, 1998.

David J Wales et al. *Energy landscapes: Applications to clusters, biomolecules and glasses*. Cambridge University Press, 2003.

Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1265–1274, 2015.