Computational Models of Argument
F. Toni et al. (Eds.)

© 2022 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA220179

Interpretable Machine Learning with Gradual Argumentation Frameworks

Jonathan SPIELER [a],Nico POTYKA [b], Steffen STAAB [a]
[a] *University of Stuttgart, Germany*
[b] *Imperial College London, United Kingdom*

Keywords. Gradual Argumentation, Explainable AI

As black-box neural networks are increasingly applied in intelligent systems, questions about their fairness, reliability and safety become louder. Recent work tried making them human-understandable by trying to learn parameters that can be well approximated by decision trees [1]. However, the tree remains just an approximation, which leaves the question how faithful it really captures the actual mechanics of the neural network. As it turns out, gradual argumentation frameworks (GAFs) [2] are closely related to multilayer perceptrons (MLPs), one of the main classes of neural networks. More precisely, every MLP corresponds to a GAF under the *MLP-based semantics*, and conversely, every acyclic GAF under this semantics corresponds to an MLP [3].

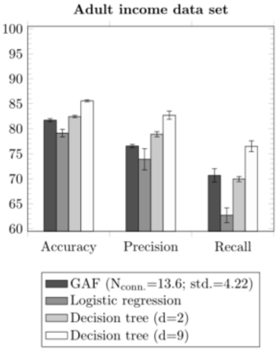
However, since a GAF with millions of attacks and supports between arguments is not easier to interpret than an MLP with millions of connections between neurons, we have to make sure that the original neural network is sparse. While learning sparse neural networks has become a more active area in recent years, current work does not focus on learning an interpretable network, but on decreasing the risk for overfitting, the memory and runtime complexity and the associated power consumption [4]. Even though the learnt networks are significantly sparser than dense networks, they are still too dense to be interpretable. Furthermore, while numerical inputs can be seen arguments with a numerical weight, MLPs result in more intuitive GAFs when all inputs are discrete. This is the opposite of what is usually done in the literature on learning neural networks, where even discrete features are often continuized (e.g., using word embeddings) to improve learning performance (while sacrificing interpretability).

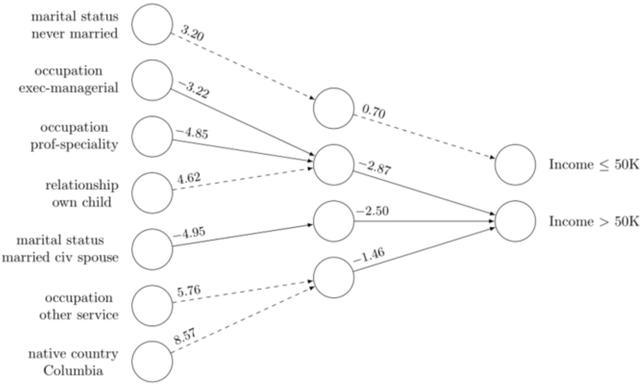
To learn a discrete sparse neural network, we apply structure learning ideas. Our search space consist of the space of all MLP structures that satisfy structural constraints. Examples for such constraints are the maximum depth (number of layers), the maximum width (number of arguments per layer), the maximum outdegree (number of outgoing edges per argument) and possible discretizations of continuous features like bins (the value falls in a particular interval) or fuzzy arguments (e.g., the value is *small*, *average*, *large*). In order to compare candidate structures, we assign a score to every candidate structure C as follows: we train C on the training set using the usual backpropagation procedure for MLPs and compute its accuracy. The score of C is then defined as

$$s_\lambda(C) = (1-\lambda)\cdot \mathrm{Accuracy}(C,\mathcal{D}_{\mathrm{train}}) + \lambda \cdot \frac{n_{max}-n_C}{n_{max}}.$$

The score consists of two terms that are weighted by a hyperparameter $\lambda \in [0,1]$. The first term evaluates the accuracy, the second one the sparsity. In the second term, n_c is the number of edges in C and n_{max} the number of edges in the fully connected GAF corresponding to C .

As the search space is exponentially large, we aim at finding a good structure, rather than the best one. To do so, we implemented a genetic algorithm. Let us emphasize that the genetic algorithm is responsible for finding a good structure, not for learning the parameters of the structure (the latter is done by backpropagation as usual). A detailled description of the algorithm and an evaluation can be found in the technical report [5]. As an example, we show a GAF (solid edges denote attacks, dashed edges supports) found for the Adult income dataset from the UCI machine learning repository and a performance comparison to Logistic Regression and Decision Trees of varying depth.





Overall, the performance of GAFs is usually better than logistic regression (which can only learn linearly separable functions) and comparable to decision trees. However, flat GAFs can sometimes obtain better performance than flat decision trees [5]. They can also be easier to comprehend as they are based on gradual influences rather than on long case differentiations. We are planning to improve the results by adding fuzzy arguments and joint attacks/supports to capture joint effects of inputs without increasing the depth of the network.

References

[1] Wu M, Parbhoo S, Hughes MC, Roth V, Doshi-Velez F. Optimizing for interpretability in deep neural networks with tree regularization. JAIR. 2021;72:1-37.

[2] Baroni P, Rago A, Toni F. How many properties do we need for gradual argumentation? In: AAAI Proceedings. AAAI; 2018. p. 1736-43.

[3] Potyka N. Interpreting Neural Networks as Gradual Argumentation Frameworks. In: AAAI Proceedings; 2021. p. 6463-70.

[4] Ma R, Niu L. A survey of sparse-learning methods for deep neural networks. In: WI Proceedings. IEEE; 2018. p. 647-50.

[5] Spieler J, Potyka N, Staab S. Learning Gradual Argumentation Frameworks using Genetic Algorithms. arXiv preprint arXiv:210613585. 2021.