Interpreting Neural Networks as Quantitative Argumentation Frameworks (Extended Abstract)

Nico Potyka
Imperial College London
npotyka@ic.ac.uk

Motivated by the flexibility and learning performance of neural networks, there has been increasing interest in using them to approximate argumentation tasks (Garcez, Gabbay, and Lamb 2005; Riveret et al. 2015; Kuhlmann and Thimm 2019; Malmqvist et al. 2020). In the context of quantitative argumentation frameworks, there is a family of argumentation models that is conceptually very close to neural networks. These models are often called bipolar gradual argumentation graphs (BAG for short). Some recent examples can be found in (Baroni et al. 2015; Rago et al. 2016; Amgoud and Ben-Naim 2018; Potyka 2018).

BAGs basically consist of weighted directed graphs with two types of edges. Nodes correspond to abstract arguments that can be accepted or rejected to a certain degree. Every node is associated with a *base score* that reflects its initial weight when ignoring all other arguments. Edges can be attacking or supporting and may have a weight that reflects the strength of the relationship between the arguments. The end goal is to assign strength values to every argument such that the strength of every argument is consistent with the strength of its attackers and supporters. To this end, gradual argumentation models define an update function that iteratively updates the strength values until they converge. The strength of every argument is initialized with its base score and updated based on the strength of its attackers and supporters. Usually, attackers should decrease and supporters should increase the strength of the affected argument based on their own respective strengths. Intuitively, the final strength values correspond to a fixed-point of the update function in which all strength values are in balance.

Conceptually, acyclic BAGs can be seen as neural networks that take some inputs (base score of the arguments without ingoing edges) and compute an output (final strength of arguments without outgoing edges) by performing some transformations on the inputs (intermediate arguments). Conversely, we can view some neural networks as gradual argumentation frameworks, where inhibitory connections between neurons (negative weights) can be seen as attacks and excitatory connections (positive weights) as supports. Given these close relationships, it is then natural to ask, can we transfer results between these two seemingly different fields to their mutual benefit?

As a first step towards linking neural networks and formal argumentation, in (Potyka 2021), we analyzed multilayer perceptrons (MLPs for short) from an argumentation perspective. MLPs process inputs on layered acyclic graphs by successively performing linear and non-linear transformations. We showed that the MLP update function can be generalized to arbitrary graphs and, in this way, can be used to interpret arbitrary BAGs. As it turns out, this new BAG semantics is mechanically close to the Euler-based semantics that has been investigated in (Amgoud and Ben-Naim 2018). Interestingly, the MLP-based semantics satisfies all semantical properties that the Euler-based semantics satisfies, but also solves some symmetry- and bias-problems of the latter. Figure 1 compares semantical properties from the literature satisfied by the Df-QuAD (Rago et al. 2016), Euler-based (Amgoud and Ben-Naim 2018), quadratic energy (Potyka 2018) and MLP-based semantics investigated in (Potyka 2021). It is interesting to note that the MLP-based se-

Property	DfQ	Euler	QEM	MLP
Anonymity	✓	✓	✓	✓
Independence	✓	✓	✓	✓
Directionality	✓	✓	✓	✓
Equivalence	✓	✓	✓	✓
Stability	✓	✓	✓	✓
Neutrality	✓	✓	✓	✓
(Strict) Monotony	(✓)	✓	✓	✓
(Strict) Reinforcement	(✓)	✓	✓	✓
Resilience	(✓)	✓	✓	✓
Franklin	✓	✓	✓	✓
Weakening	(✓)	✓	✓	✓
Strengthening	(✓)	✓	✓	✓
Duality	✓	×	✓	✓
Open-Mindedness	×	×	✓	(✓)

Figure 1: Properties fully satisfied (✓), satisfied when excluding base scores 0 and 1 ((✓)), not satisfied even when excluding base scores 0 or 1 (x).

mantics satisfies almost all properties perfectly even though it has not been designed for this purpose. The only exception is the open-mindedness property that demands that the base score of an argument should not completely determine the final strength of the argument. This property is violated by the MLP-based semantics when the base scores are intialized with 0 (full rejection) or 1 (full acceptance). However, let us note that this case cannot occur in MLPs trained on data because these base scores correspond to a neuron with bias negative infinity or positive infinity.

The computational guarantees for argumentation under the MLP-based semantics are similar to the other semantics in Figure 1. In acyclic graphs (including MLP structures), strength values can be computed in linear time by a simple forward pass through the graph. In acyclic graphs, the strength values typically converge in subquadratic time. However, tools from (Mossakowski and Neuhaus 2018) can be used to construct examples where the strength values start oscillating and fail to converge. We provided sufficient conditions for convergence of the MLP-based semantics in (Potyka 2021) using tools from (Mossakowski and Neuhaus 2018; Potyka 2019). Furthermore, experiments with convergence counterexamples indicate that convergence problems can be avoided my continuizing the semantics as discussed in (Potyka 2018; Potyka 2019). The continuization maintains the original semantics in known convergence cases. Intuitively, this is because the update function has a unique fixed point in these cases. In the difficult cases from (Mossakowski and Neuhaus 2018), the discrete update will fail to reach a fixed point. Intuitively, this is because it takes too large steps. By taking continuous (infinitesimal) steps, the convergence problem can be circumvented.

The connection between neural networks and argumentation frameworks seems interesting to transfer results between the two areas. In (Potyka 2021), we applied previous work on convergence guarantees for BAGs (Mossakowski and Neuhaus 2018; Potyka 2019) to generalize the mechanics of MLPs to arbitrary graphs, which is interesting for the theory of MLPs. The generalization gave rise to a new BAG semantics with strong semantical guarantees, which is interesting for the theory of BAGs. Another interesting direction is using learning ideas for neural networks to advance the state of the art in learning BAGs from data. First ideas in this direction have been sketched in (Spieler, Potyka, and Staab 2021). Ideas from the field of argumentation like sets of attacks or supports may then again lead to novel ideas for additional structure in neural networks that may improve the learning performance in the future. From an explainable AI perspective, the connection also seems useful to create argumentation frameworks from black-box neural networks as suggested in (Albini et al. 2020). Let us also note that (Giordano 2021) recently extended the connection between gradual argumentation frameworks and MLPs to weighted conditionals.

References

Albini, E.; Lertvittayakumjorn, P.; Rago, A.; and Toni, F. 2020. Dax: Deep argumentative explanation for neural networks. *arXiv preprint arXiv:2012.05766*.

Amgoud, L., and Ben-Naim, J. 2018. Weighted bipolar argumentation graphs: Axioms and semantics. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 5194–5198.

Baroni, P.; Romano, M.; Toni, F.; Aurisicchio, M.; and Bertanza, G. 2015. Automatic evaluation of design alternatives with quantitative argumentation. *Argument & Computation* 6(1):24–49.

Garcez, A. S.; Gabbay, D. M.; and Lamb, L. C. 2005. Value-based argumentation frameworks as neural-symbolic learning systems. *Journal of Logic and Computation* 15(6):1041–1058.

Giordano, L. 2021. From weighted conditionals of multilayer perceptrons to a gradual argumentation semantics. *CoRR* abs/2110.03643.

Kuhlmann, I., and Thimm, M. 2019. Using graph convolutional networks for approximate reasoning with abstract argumentation frameworks: A feasibility study. In *International Conference on Scalable Uncertainty Management (SUM)*, 24–37. Springer.

Malmqvist, L.; Yuan, T.; Nightingale, P.; and Manandhar, S. 2020. Determining the acceptability of abstract arguments with graph convolutional networks. In *International Workshop on Systems and Algorithms for Formal Argumentation (SAFA@COMMA)*, 47–56.

Mossakowski, T., and Neuhaus, F. 2018. Modular Semantics and Characteristics for Bipolar Weighted Argumentation Graphs. *arXiv preprint arXiv:1807.06685*.

Potyka, N. 2018. Continuous dynamical systems for weighted bipolar argumentation. In *International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 148–157.

Potyka, N. 2019. Extending Modular Semantics for Bipolar Weighted Argumentation. In *International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 1722–1730.

Potyka, N. 2021. Interpreting neural networks as gradual argumentation frameworks. In *AAAI Conference on Artificial Intelligence (AAAI)*, 6463–6470.

Rago, A.; Toni, F.; Aurisicchio, M.; and Baroni, P. 2016. Discontinuity-free decision support with quantitative argumentation debates. In *International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 63–73.

Riveret, R.; Pitt, J. V.; Korkinof, D.; and Draief, M. 2015. Neuro-symbolic agents: Boltzmann machines and probabilistic abstract argumentation with sub-arguments. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 1481–1489.

Spieler, J.; Potyka, N.; and Staab, S. 2021. Learning gradual argumentation frameworks using genetic algorithms. *arXiv preprint arXiv:2106.13585*.