On Learning Bipolar Gradual Argumentation Semantics with Neural Networks

Caren Al Anaissy, Sandeep Suntwal, Mihai Surdeanu, Srdjan Vesic

On Learning Bipolar Gradual Argumentation Semantics with Neural Networks

Caren Al Anaissy[1][a], Sandeep Suntwal[2][b], Mihai Surdeanu[3][c] and Srdjan Vesic[4][d]
[1]*CRIL Université d'Artois & CNRS, Lens, France*
[2]*University of Colorado, Colorado Springs, United States*
[3]*University of Arizona, Tucson, United States*
[4]*CRIL CNRS Univ. Artois, Lens, France*
alanaissy@cril.fr, ssuntwal@uccs.edu, msurdeanu@arizona.edu, vesic@cril.fr

Keywords: Argumentation Semantics, Bipolar Gradual Argumentation Graphs, Neural Networks.

Abstract: Computational argumentation has evolved as a key area in artificial intelligence, used to analyze aspects of thinking, making decisions, and conversing. As a result, it is currently employed in a variety of real-world contexts, from legal reasoning to intelligence analysis. An argumentation framework is modelled as a graph where the nodes represent arguments and the edges of the graph represent relations (i.e., supports, attacks) between nodes. In this work, we investigate the ability of neural network methods to learn a gradual bipolar argumentation semantics, which allows for both supports and attacks. We begin by calculating the acceptability degrees for graph nodes. These scores are generated using Quantitative Argumentation Debate (QuAD) argumentation semantics. We apply this approach to two benchmark datasets: Twelve Angry Men and Debatepedia. Using this data, we train and evaluate the performance of three benchmark architectures: Multilayer Perceptron (MLP), Graph Convolution Network (GCN), and Graph Attention Network (GAT) to learn the acceptability degree scores produced by the QuAD semantics. Our results show that these neural network methods can learn bipolar gradual argumentation semantics. The models trained on GCN architecture perform better than the other two architectures underscoring the importance of modelling argumentation graphs explicitly. Our software is publicly available at: *https://github.com/clulab/icaart24-argumentation*.

1 INTRODUCTION

Computational argumentation theory (CAT) (Dung, 1995; Besnard and Hunter, 2008; Rahwan and Simari, 2009; Baroni et al., 2018; Amgoud and Prade, 2009; Amgoud and Serrurier, 2008; Atkinson et al., 2017) has emerged as a fundamental area of artificial intelligence (AI) and is used in many tasks such as reasoning with inconsistent information, decision making (Amgoud and Prade, 2009), and classification (Amgoud and Serrurier, 2008) across domains such as legal and medical (Atkinson et al., 2017).

[a] https://orcid.org/0000-0002-8750-1849
[b] https://orcid.org/0000-0002-7746-7114
[c] https://orcid.org/0000-0001-6956-8030
[d] https://orcid.org/0000-0002-4382-0928
The first and the second authors contributed equally.

CAT focuses on formalising and automating the process of argumentation. It accomplishes this by constructing argumentation graphs in which arguments and counter-arguments are interconnected through either attack or support edges. Within CAT, *semantics* serves as a formal method for evaluating the strength of each argument in the graph considering its interactions with other arguments. Several *bipolar gradual* semantics have been proposed recently: QuAD (Baroni et al., 2015), DF-QuAD (Rago et al., 2016), Exponent-based (Amgoud and Ben-Naim, 2018), and Quadratic Energy Model (QEM) (Potyka, 2018). This class of semantics have two key properties: (a) they model both attacks and supports edges, and (b) they compute acceptability degrees for the nodes in the graph, which quantify the strength of individual arguments in the argument graph. That is, arguments are not only met with opposition (attack), but also with reinforcement (support). For example, QuAD semantics (Baroni et al., 2015), a state-of-the-art semantics, enables the joint analysis and treatment

of *both* attack and support relations between arguments. That is, it considers the initial weights of the arguments as well as the attack and support relations between them in order to calculate the final acceptability degree for each argument.

To capture the complex landscape of real-world discourse, in this work we focus on bipolar gradual argumentation semantics. While bipolar gradual argumentation semantics provide the final acceptability degree for each node in the graph, there is still no solution to the fact that they are not guaranteed to converge in case of cycles. This issue can be resolved using neural networks. Neural networks are known for converging (Smith and Topin, 2019) however, their ability to learn argumentation semantics is still unknown. In this work, we focus on the latter issue.

In the space of deep learning, recent developments in machine learning and especially in large language models (LLMs) such as GPT-4 (OpenAI, 2023) have generated claims of (sparks of) Artificial General Intelligence (AGI) (Bubeck et al., 2023; Zhang et al., 2023). These claims seem to be supported by the observation that LLMs can solve new tasks that they have not been exposed to during their training such as mathematics, coding, vision, medicine, law, psychology, without needing any domain-specific prompting (Bubeck et al., 2023).

Our paper is motivated by an important observation that connects the two threads above: *a crucial prerequisite for claims of machine reasoning or AGI is that the underlying neural networks understand argumentation theory semantics*. Given that argumentation is an integral part of thinking, making decisions, and conversing, how else would a machine truly reason? To verify if neural networks can model CAT semantics, in this work we train and evaluate multiple neural architectures on their capacity to capture QuAD semantics. In particular, our paper makes three contributions:

1. We construct a dataset of argument graphs that captures QuAD semantics. In particular, we used the argument graphs from two datasets: Twelve Angry Men and Debatepedia provided by the NoDE benchmark (Cabrio and Villata, 2013; Cabrio et al., 2013). We generated multiple versions of these graphs, where the arguments' initial weights are drawn from four different distributions (Beta, Normal, Poisson, Uniform), and the corresponding acceptability degrees are computed using QuAD semantics.
2. We implement three distinct neural architectures that learn to predict QuAD semantics acceptability degrees given a graph with initial weights. Our architectures are based on: Multilayer Perceptrons (MLP), Graph Convolution Networks (GCN) (Kipf and Welling, 2016), and Graph Attention Networks (GAT) (Veličković et al., 2017). To capture the argument graph structure, all these architectures have access to node features that include: (a) node in-degree, (b) total degree, and (c) the initial node weight. In addition, the two graph-based architectures use as features: the initial node weights, edge information, and edge weights (+1 for supports and -1 for attacks edges).
3. We evaluate the capacity of these architectures to predict correct acceptability degrees. We conclude that their predictions indeed come close to the true QuAD semantics acceptability degree (with a mean squared error as low as 0.05). However, this conclusion has an important caveat: this performance is only achieved when the argument graph is explicitly modelled using a graph-based neural architecture.

2 RELATED WORK

Kuhlmann and Thimm (Kuhlmann and Thimm, 2019) and Craandijk and Bex (Craandijk and Bex, 2020) trained neural networks to learn extension-based argumentation semantics. Kuhlmann and Thimm employed a conventional single forward pass classifier to approximate credulous acceptance under the preferred semantics. Craandijk and Bex proposed an argumentation Graph Neural Network (GNN) that learns to predict both credulous and sceptical acceptance of arguments under four well-known extension-based argumentation semantics. Our effort operates under the same goal, i.e., training a neural network to learn an argumentation semantics, but it extends these works considerably. First, these two papers focused on extension-based semantics, whereas ours focuses on gradual semantics. This, in principle, is harder to replicate using neural networks due to the continuous values of acceptability degrees. Second, both these papers take into account only attacks edges; our method addresses both attacks and supports edges. Lastly, we explore three different neural architectures to better understand the best representation for argumentation semantics.

Within CAT, bipolar gradual semantics, which model both attacks and supports edges and com-

It is important to note that these claims are not widely accepted due to suspicions of "contamination," i.e., the data used to train these LLMs often contains the tasks used during testing (Sainz et al., 2023), which invalidates the bold claims of AGI.

pute argument acceptability degrees, are well studied. The QuAD semantics was proposed by Baroni et al. (Baroni et al., 2015) to evaluate the strength of answers in decision-support systems. However, this semantics can sometimes behaviour discontinuously. The DF-QuAD semantics was proposed by Rago et al. (Rago et al., 2016) to fix this discontinuity problem. Amgoud and Ben-Naim (Amgoud and Ben-Naim, 2018) introduce a set of thirteen principles for bipolar weighted argumentation semantics. In their work, the authors explain that both the QuAD and the DF-QuAD semantics do not satisfy some of these principles. This is because, as Potyka explains (Potyka, 2018; Potyka, 2019), the QuAD and the DF-QuAD semantics both have a saturation problem. That means that once an argument has an attacker (supporter) with a degree of 1, it becomes meaningless to take all the other attackers (supporters) into account in the aggregation function. To fix this saturation problem, Amgoud and Ben-Naim (Amgoud and Ben-Naim, 2018) propose the Exponent-based semantics that satisfies the thirteen principles proposed. However, Potyka (Potyka, 2018) explains that the Exponent-based semantics violates the duality principle, meaning that this semantics treats the attack relation and the support relation in an asymmetrical manner. Potyka (Potyka, 2019) also explains that the Exponent-based semantics do not satisfy "open-mindedness," i.e., its ability to change the initial weights is very limited. The Quadratic Energy Model (QEM) proposed by Potyka (Potyka, 2018) satisfies twelve properties among the thirteen properties proposed by Amgoud and Ben-Naim (Amgoud and Ben-Naim, 2018), the duality and the open-mindedness properties. The MLP-Based semantics (Potyka, 2021) consists of viewing a multilayer perceptron (MLP), which is a feed-forward neural network, as a bipolar gradual argumentation framework. This semantics satisfies the same properties as the QEM, except that the MLP-Based semantics satisfies the open-mindedness property excluding the cases where the arguments' intial weight is 0 or 1. However, in this work we focus on the QuAD semantics because it is widely known and easily explainable. We plan to investigate other semantics in future work.

Unfortunately, QuAD and most of the followup semantics are only defined for acyclic graphs. To detail, the DF-QuAD semantics (Rago et al., 2016) and the Exponent-based semantics (Amgoud and Ben-Naim, 2018) are also defined for acyclic graphs. For the semantics proposed by Mossakowski and Neuhaus (Mossakowski and Neuhaus, 2016), the convergence is not guaranteed for all cyclic graphs. The convergence of the Quadratic Energy Model proposed by Potyka (Potyka, 2018) for cyclic graphs is not proven. Mossakowski and Neuhaus (Mossakowski and Neuhaus, 2018) show that twenty-five different semantics can be obtained by combining five aggregation functions with five influence functions. However, only three of these semantics converge for all graphs. Potyka (Potyka, 2019) shows that these three semantics do not satisfy open-mindednes, which makes them unsuitable for any practical application. Potyka shows also that continuizing discrete models can solve divergence problems. However, there is currently no proof of convergence for continuous models in cyclic graphs. The MLP-Based semantics (Potyka, 2021) is fully-defined for all acyclic graphs and for cyclic graphs with convergence conditions.

Since we use QuAD, we also focus only on acyclic graphs in this work. We acknowledge this limitation, and discuss future work on cyclic graphs in Section 5.

3 APPROACH

3.1 Data

We performed experiments using datasets created using argument graphs from two benchmarks: Twelve Angry Men and Debatepedia provided by the NoDE benchmark (Cabrio and Villata, 2013; Cabrio et al., 2013). Note that all the graphs in those datasets are acyclic. The two datasets used in our study are described next.

Twelve Angry Men Dataset The script of "Twelve Angry Men" is the first natural language argument benchmark in our experiments. This dataset contains three acts (Act 1, Act 2 and Act 3). Every argument in each act links to at least another dialogue argument it *supports* or *attacks* within the act. The benchmark contains three graphs (one for each act), where each argument represents a node and is connected to other nodes that it *supports* or *attacks*. The three graphs contain 80 edges and 83 nodes in total. We split the dataset as follows: the graphs from Act 1 and Act 2 (72 nodes) were the training and validation partitions, while the graph in Act 3 (11 nodes) was the test partition.

Debatepedia and ProCon Dataset This dataset consists of two encyclopedia of pro and con arguments. The dataset was manually constructed by selecting a set of topics of Debatepedia/ProCon debates. Here, each debate represents one topic. Within

each topic, an argument is constructed by extracting user opinion. In this dataset, as the attack and support edges are represented as binary relations, the arguments are connected with the starting argument or another argument within the same topic to which the newest argument refers. A chronological order is maintained to ensure a dialogue structure. This dataset contains 20 debates across different topics.The train and validation partition contain 14 graphs; the test partition contains six graphs.

3.2 Method

The QuAD semantics requires each argument to have an initial weight that captures its intrinsic value. However, the datasets do not provide such information. To address this limitation, we generated initial weights for each node in our dataset using four probability distributions: Beta, Normal, Poisson, and Uniform. Next, we computed the acceptability degree score for each node using the QuAD semantics (Baroni et al., 2015). Finally, we used the acceptability degree scores as the gold degrees to train, validate, and test three neural network architectures: a multilayer perceptron (MLP), a graph convolution network (GCN) (Kipf and Welling, 2016), and a graph attention network (GAT) (Veličković et al., 2017). These steps were repeated for 100 different initial weights for each graph and initial weight distribution.

3.2.1 Generating argumentation theory-based baselines

The notation $Sup(a)$ (resp. $Att(a)$) stands for the set of supporters (resp. attackers) of argument a, and the notation $w(a)$ stands for the initial weight of a. $\deg(a)$ stands for the acceptability degree of a. We provide a short explanation of how the QuAD semantics works.

Quantitative Argumentation Debate The QuAD semantics determines the strength, i.e., the acceptability degree of each argument, by considering its initial weight and the aggregated strengths of its attackers and supporters. The functions $f_a(a)$ and $f_s(a)$ recursively aggregate the strengths of an argument a's attackers and supporters respectively with a's initial weight. Figure 1 illustrates the QuAD semantics applied on a bipolar argumentation graph.

Then, the acceptability degree of a is defined as:

$$\deg(a) = \begin{cases} f_a(a) & \text{if } \texttt{Sup}(a) = \emptyset, \texttt{Att}(a) \neq \emptyset \\ f_s(a) & \text{if } \texttt{Sup}(a) \neq \emptyset, \texttt{Att}(a) = \emptyset \\ w(a) & \text{if } \texttt{Sup}(a) = \emptyset, \texttt{Att}(a) = \emptyset \\ \frac{f_a(a)+f_s(a)}{2} & \text{otherwise} \end{cases} \tag{1}$$

where

$$f_a(a) = w(a) \cdot \prod_{b \in \texttt{Att}(a)} (1 - \deg(b))$$

and

$$f_s(a) = 1 - (1 - w(a)) \cdot \prod_{b \in \texttt{Sup}(a)} (1 - \deg(b))$$

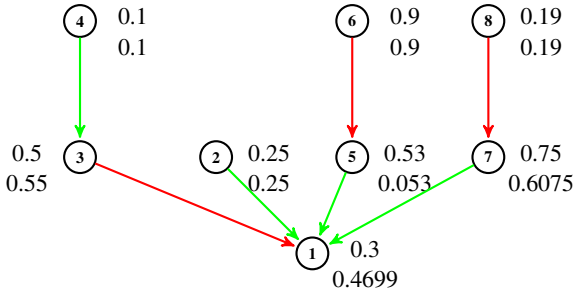

Figure 1: Example of a bipolar argumentation graph extracted from the Debatepedia dataset (Cabrio and Villata, 2014), for the debate called *Sobrietytest*. The red arrows represent attacks while the green arrows represent supports. The first row of numbers next to the arguments represent their initial weights which are assigned randomly (drawn from a given distribution), while the second row of numbers represent their acceptability degrees calculated using the QuAD semantics.

3.2.2 Learning argumentation semantics

To investigate the capacity of neural networks to learn QuAD semantics, we explore three different neural network architectures. Note that the neural network architectures typically employed for learning structured data are not suited for resolving non-euclidean input data such as graphs. Due to the variable size and shape of the input graphs, it is difficult to process them using conventional data structures such as adjacency matrices (Veličković et al., 2017). In addition, adjacency matrices are dependent on the order in which the nodes appear, so they are not node invariant. To avoid the limitations of existing algorithms, we transform the arguments and debates into neural network graph structures, and learn the semantics using Graph Neural Networks (GNNs) (Gori et al., 2005; Scarselli et al., 2008). GNNs aim to utilize conventional deep learning principles for non-euclidean

Table 1: Results for Twelve Angry Men and Debatepedia datasets using QuAD. The top row indicates the neural network architecture. The first column represents the distribution from which the initial random weights were assigned. These results are means across 100 different initial weights for each node. MSE represents the final mean squared error. Since MSE is an error based statistic, a lower value is better. RMSE represents the relative MSE. The MSE results highlighted in bold indicate the best MSE results among the three neural networks.The Kendall Tau correlation metric measures the correlation between the node ranks produced by a given neural network and the ranks produced by QuAD semantics. A statistically significant Kendall's τ represents that rankings produced by NNs and QuAD semantics were similar. *, **, *** indicate statistical significance ($\leq$.05, $\leq$.005, $\leq$.0005).

Model	**MLP**			**GCN**			**GNN-GAT**		
	MSE	RMSE	Kendall Tau	MSE	RMSE	Kendall Tau	MSE	RMSE	Kendall Tau
Twelve Angry Men Dataset									
Beta	0.18	3.33	-0.23	**0.09**	1.67	0.48	0.14	2.55	0.23
Normal	0.15	2.84	-0.18	**0.09**	1.64	0.41	0.11	2.08	0.26
Poisson	0.2	inf	-0.44	**0.14**	inf	0.39	0.18	inf	0.25
Uniform	0.14	2.48	-0.19	**0.07**	1.23	0.47	0.1	1.78	0.22
Debatepedia Dataset									
Beta	0.36	0.67	0.36	**0.07**	0.12	0.72***	0.1	0.19	0.69
Normal	0.34	0.64	0.39	**0.06**	0.11	0.73***	0.08	0.16	0.65
Poisson	0.29	1.06	0.2	**0.08**	0.27	0.62***	0.14	0.5	0.76
Uniform	0.32	0.56	0.33	**0.05**	0.08	0.76***	0.07	0.12	0.64

data. GNNs achieve this by allowing for message passing between nodes (which represent arguments) and edges (which represent relations such as support or attack, and have a weight) through *convolution* operations. The equation for this operation is as follows:

$$\mathbf{n}_i^{(k)} = \gamma^{(k)}\left(\mathbf{n}_i^{(k-1)}, f_{j\in\mathcal{N}(i)}\,\phi^{(k)}\left(\mathbf{n}_i^{(k-1)}, \mathbf{n}_j^{(k-1)}, \mathbf{e}_{j,i}\right)\right) \tag{2}$$

where $\mathbf{n}_i^{(k-1)} \in \mathbb{R}^P$ represents node $i's$ node features in layer $k-1$, $\mathbf{e}_{j,i} \in \mathbb{R}^Q$ represents edge features from node j to node i, informing us about the importance of each neighbor. f represents a permutation invariant function (e.g., sum, avg) to ensure that the methods make no assumptions about the spatial relationships between node features during convolutions, ϕ denotes a neural network that constructs the message to node i for each edge j,i, and γ denotes a neural network that takes the output of this aggregation to update the acceptability degree scores for node i. The final output produces the updated acceptability degree scores for each node in the graph.

We conducted experiments using three architectures: MLP, GCN, and GAT. MLP is a feed-forward neural network architecture suitable for some structure data but not graphs. This architecture serves as our first baseline. GCNs (Kipf and Welling, 2016) extend the concept of convolution, which is widely used in convolutional neural networks (CNNs) (LeCun et al., 1998) for image or text processing, to the graph domain. In CNNs, a convolutional layer applies filters over local patches of an input image (or local textual context) to extract features. Similarly, in GCNs, a convolutional layer processes nodes and their neighboring nodes to aggregate information. GATs, our third architecture, use more complex attention-based architectures to explore the *entire* neighborhood of a node in an order invariant way (Veličković et al., 2017). To capture the graph structure for the MLP architecture, we created node features that contain: (a) node in-degree, (b) total degree, and (c) the initial node weight. These features help overcome some limitations such as uniform structure for each input. For the two graph-based architectures, we used as features: the initial node weights, edge information, and edge weights (+1 for supports and -1 for attacks edges). All neural methods were trained using the mean-squared error (MSE) loss on the corresponding training partitions. We tuned the early stopping hyperparameter on the validation partition.

4 RESULTS AND DISCUSSION

Table 1 presents the results from our study, in which we evaluate the three proposed neural architectures on the graphs introduced in previous section. This table lists results across the four distributions for initial argument weights. To evaluate the performance of the neural architectures we used mean squared error (MSE) between the predicted and the gold acceptability degrees. We also used the Kendall rank correlation coefficient metric (also referred as Kendall's τ) (Kendall, 1938), to measure the correlation between the node ranks produced by a given neural network and the ranks produced by QuAD semantics. Kendall's tau computes the difference between the number of matching and non-matching observation pairs and divides it by the total number of observation pairs. A value of -1 indicates complete negative

association, 1 indicates complete positive association, and 0 indicates no association between the variables. The underlying hypothesis tests if the two lists are identical. A p-value $\leq .05$ and τ value > 0 signifies a statistically significant similarity between the two rankings. Here, a statistically significant Kendall's τ score indicates that the acceptability score rankings produced by QuAD semantics have a high correlation with the acceptability degree scores produced by the neural networks.

Formally, Kendall's τ is calculated as follows:

$$\tau = \frac{2}{k(k-1)} \sum_{i<j} \text{sgn}(x_i - x_j)\,\text{sgn}(y_i - y_j)$$

where k is the number of observations, x_i and y_i are the rankings of the i^{th} observation for the two variables being compared, and *sgn* is the sign function that returns 1 if the argument is positive, -1 if negative, and 0 if zero. We draw the following observations from this experiment:

- First, our experiments indicate that neural networks can indeed learn to predict the acceptability degrees computed by QuAD semantics. To our knowledge, this is the first work to show that neural networks can learn to replicate bipolar gradual CAT semantics. This is an exciting result considering the complexity of the task.
- Second, the best results by far are obtained by neural architectures that model graphs explicitly (GCN and GNN-GAT), which highlights the importance of modelling argument interactions. For example, the MSEs measured for GCN range between 0.05 and 0.09 for the two datasets and the four probability distributions used for initial weights. In contrast, the MSEs measured for the MLP, which uses a limited number of features to summarize the graph structure, range from 0.14 to 0.36 depending on the dataset and probability distribution. The latter results are not only considerably worse, but also show extreme variation between datasets.
- Third, between the two graph-based architectures, GCN performs consistently better than GNN-GAT. For example, for the Debatepedia Dataset, the MSE measured for GCN ranges from 0.05 to 0.08, whereas the MSE for GNN-GAT is approximately twice that ranging from 0.08 to 0.14. We consider this another positive result, which indicates that, as long as the neural architectures capture the argument graph structures, simplicity is better. This is an important observation for the deployment of these neural architectures in real-world software applications.
- Lastly, if we consider the Kendall Tau scores calculated between the ranking produced by GCN and the ranking produced by QuAD semantics, we can observe a moderate positive correlation between the two rankings for the Twelve Angry Men dataset. We can also observe a strong positive correlation between the two rankings for the Debatepedia dataset, which is statistically significant at a very low p-value threshold. These results show us that GCN is not only capturing the general trends in the data (as indicated by low MSE) but is also performing well in maintaining the ranking of the values, therefore producing consistent ranking. These results are particularly valuable since they give us insights about the capability of GCN to effectively predict the ranking among the nodes in a bipolar gradual argumentation graph.

5 CONCLUSION AND FUTURE WORK

Argumentation continues to grow as a key area in AI for several tasks that require decision-making and communication. As such, we argue that understanding argumentation should be a requirement for neural networks that implement machine reasoning. In this study, we investigated the ability of neural networks to learn gradual bipolar argumentation semantics. We conducted several experiments to train and evaluate neural networks' ability to learn the QuAD argumentation semantics. Our findings indicate that argumentation semantics can be learned by neural networks successfully. One important observation is that the best results are consistently produced by GCNs, a graph-based neural architecture, which underlines that the argument graph structure must be explicitly modelled. All in all, this paper is the first to show that neural architectures can learn gradual bipolar argumentation semantics.

One limitation of this work is that all the data used for training and testing the proposed neural architectures consisted only of *acyclic* graphs, where QuAD semantics is guaranteed to converge. In the future, we will expand our work to include cyclic graphs where the QuAD (or other) semantics converge. To the best of our knowledge, there does not exist a dataset or benchmark that contains a significant number of real-world bipolar weighted cyclic graphs. We are currently working on collecting cyclic graphs from real online debates and verify in which situations QuAD semantics converges. Other future work will include training neural networks on other bipolar weighted semantics, based on the behaviour desired in a specified

context.

ACKNOWLEDGEMENTS

This work was supported by the International Emerging Action (IEA) project RHAPSSODY and joint PhD Program SURFING, both funded by the French National Center for Scientific Research (CNRS) and the University of Arizona. Caren Al Anaissy and Srdjan Vesic were also supported by the project AGGREEY ANR22-CE23-0005 from the French National Research Agency (ANR).

REFERENCES

Amgoud, L. and Ben-Naim, J. (2018). Evaluation of arguments in weighted bipolar graphs. *International Journal of Approximate Reasoning*, 99:39–55.

Amgoud, L. and Prade, H. (2009). Using arguments for making and explaining decisions. *Artificial Intelligence*, 173:413–436.

Amgoud, L. and Serrurier, M. (2008). Agents that argue and explain classifications. *Autonomous Agents and Multi-Agent Systems*, 16(2):187–209.

Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., Simari, G. R., Thimm, M., and Villata, S. (2017). Towards artificial argumentation. *AI Magazine*, 38(3):25–36.

Baroni, P., Gabbay, D., Giacomin, M., and van der Torre, L., editors (2018). *Handbook of Formal Argumentation*. College Publications.

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.

Besnard, P. and Hunter, A. (2008). *Elements of Argumentation*. MIT Press.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Cabrio, E. and Villata, S. (2013). A natural language bipolar argumentation approach to support users in online debate interactions. *Argument & Computation*, 4(3):209–230.

Cabrio, E. and Villata, S. (2014). Node: A benchmark of natural language arguments. In *Computational Models of Argument*, pages 449–450. IOS Press.

Cabrio, E., Villata, S., and Gandon, F. (2013). A support framework for argumentative discussions management in the web. In *The Semantic Web: Semantics and Big Data: 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings 10*, pages 412–426. Springer.

Craandijk, D. and Bex, F. (2020). Deep learning for abstract argumentation semantics. In Bessiere, C., editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 1667–1673. ijcai.org.

Dung, P. M. (1995). On the Acceptability of Arguments and its Fundamental Role in Non-Monotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, 77:321–357.

Gori, M., Monfardini, G., and Scarselli, F. (2005). A new model for learning in graph domains. In *Proceedings. 2005 IEEE international joint conference on neural networks*, volume 2, pages 729–734.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kuhlmann, I. and Thimm, M. (2019). Using graph convolutional networks for approximate reasoning with abstract argumentation frameworks: A feasibility study. In *Scalable Uncertainty Management: 13th International Conference, SUM 2019, Compiègne, France, December 16–18, 2019, Proceedings*, pages 24–37. Springer.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Mossakowski, T. and Neuhaus, F. (2016). Bipolar weighted argumentation graphs. *arXiv preprint arXiv:1611.08572*.

Mossakowski, T. and Neuhaus, F. (2018). Modular semantics and characteristics for bipolar weighted argumentation graphs. *arXiv preprint arXiv:1807.06685*.

OpenAI (2023). Gpt-4 technical report.

Potyka, N. (2018). Continuous dynamical systems for weighted bipolar argumentation. In *KR*, pages 148–157.

Potyka, N. (2019). Extending modular semantics for bipolar weighted argumentation. In *KI 2019: Advances in Artificial Intelligence: 42nd German Conference on AI, Kassel, Germany, September 23–26, 2019, Proceedings 42*, pages 273–276. Springer.

Potyka, N. (2021). Interpreting neural networks as quantitative argumentation frameworks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6463–6470.

Rago, A., Toni, F., Aurisicchio, M., Baroni, P., et al. (2016). Discontinuity-free decision support with quantitative argumentation debates. *KR*, 16:63–73.

Rahwan, I. and Simari, G. R., editors (2009). *Argumentation in Artificial Intelligence*. Springer.

Sainz, O., Campos, J. A., García-Ferrero, I., Etxaniz, J., and Agirre, E. (2023). Did chatgpt cheat on your test? Last accessed: 18th July, 2023.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

Smith, L. N. and Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Zhang, C., Zhang, C., Li, C., Qiao, Y., Zheng, S., Dam, S. K., Zhang, M., Kim, J. U., Kim, S. T., Choi, J., et al. (2023). One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era. *arXiv preprint arXiv:2304.06488*.