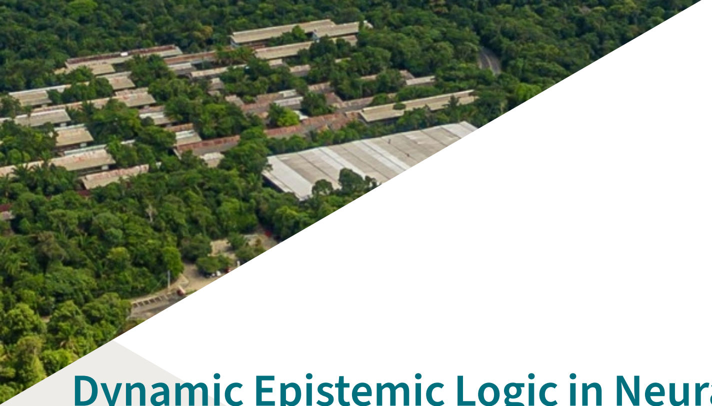

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

SBMF 2023
26th Brazilian Symposium on Formal Methods

Dynamic Epistemic Logic in Neural Layer Transparency

Towards a Formal Understanding of Knowledge Evolution in Neural Networks

Jefferson O. Andrade

2023-12-05

Motivation

- Artificial neural networks (ANN) have revolutionized areas like image recognition and natural language processing.

Motivation

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- Artificial neural networks (ANN) have revolutionized areas like image recognition and natural language processing.
- But their inherent opacity hinders broad trust and acceptance.

Motivation

- Artificial neural networks (ANN) have revolutionized areas like image recognition and natural language processing.
- But their inherent opacity hinders broad trust and acceptance.
- The main hurdle is their intrinsic complexity; understanding their detailed internal processes remains elusive.

Motivation

- Artificial neural networks (ANN) have revolutionized areas like image recognition and natural language processing.
- But their inherent opacity hinders broad trust and acceptance.
- The main hurdle is their intrinsic complexity; understanding their detailed internal processes remains elusive.
- Many attemps have been made to bring interpretability and transparency to ANN. For example:
  - XAI
  - Saliency maps
  - Attention mechanisms
  - Influence functions

  But they are yet to gain widespread acceptance.

Problem Statement

- Artificial neural networks, especially deep learning models, are **inherently complex** and often lack transparency.

Problem Statement

- Artificial neural networks, especially deep learning models, are **inherently complex** and often lack transparency.
- There is a significant challenge in interpreting how these models process and derive conclusions from data.

Problem Statement

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- Artificial neural networks, especially deep learning models, are **inherently complex** and often lack transparency.
- There is a significant challenge in interpreting how these models process and derive conclusions from data.
- Existing methods fall short in effectively representing the **evolution of knowledge** within neural layers.

Problem Statement

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- Artificial neural networks, especially deep learning models, are **inherently complex** and often lack transparency.
- There is a significant challenge in interpreting how these models process and derive conclusions from data.
- Existing methods fall short in effectively representing the **evolution of knowledge** within neural layers.
- The ‘black box’ nature of neural networks raises concerns in ethical AI applications and decision-making processes.

Problem Statement

- Artificial neural networks, especially deep learning models, are **inherently complex** and often lack transparency.
- There is a significant challenge in interpreting how these models process and derive conclusions from data.
- Existing methods fall short in effectively representing the **evolution of knowledge** within neural layers.
- The ‘black box’ nature of neural networks raises concerns in ethical AI applications and decision-making processes.
- Addressing these challenges is crucial for advancing AI towards more **transparent**, **interpretable**, and **trustworthy** systems.

Background

- **Symbolic methods** offer clear, explicit representations, but they falter with real-world datas unpredictability.

Background

- Symbolic methods offer clear, explicit representations, but they falter with real-world datas unpredictability.
- Neural networks bridge this gap by detecting intricate patterns but, in doing so, turned their internal workings into hard-to-decipher black boxes.

Background

- **Symbolic methods** offer clear, explicit representations, but they falter with real-world datas unpredictability.
- **Neural networks** bridge this gap by detecting intricate patterns but, in doing so, turned their internal workings into hard-to-decipher black boxes.
- **Neuro-symbolic AI** seeks to fuse neural networks empirical strength with classical AIs symbolic reasoning.

Background

- **Symbolic methods** offer clear, explicit representations, but they falter with real-world datas unpredictability.
- **Neural networks** bridge this gap by detecting intricate patterns but, in doing so, turned their internal workings into hard-to-decipher black boxes.
- **Neuro-symbolic AI** seeks to fuse neural networks empirical strength with classical AIs symbolic reasoning.
- One promising approach for Neuro-symbolic AI is the application of **Dynamic Epistemic Logic (DEL)** to understand neural behaviors.

Dynamic Epistemic Logic (DEL) Overview

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- **What is DEL?**
  - DEL studies the effects of **actions** on knowledge and beliefs, particularly in multi-agent systems.
  - It is a modal logic, extending classical logics with modalities for dynamic actions.

Dynamic Epistemic Logic (DEL) Overview

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- **What is DEL?**
  - DEL studies the effects of **actions** on knowledge and beliefs, particularly in multi-agent systems.
  - It is a modal logic, extending classical logics with modalities for dynamic actions.
- **Key Components:**
  - **Kripke Models**: Used to represent possible worlds and agents' knowledge about these worlds.
  - **Actions**: Represented as model transformations in DEL, changing the state of knowledge.

Dynamic Epistemic Logic (DEL) Overview

- **What is DEL?**
  - DEL studies the effects of **actions** on knowledge and beliefs, particularly in multi-agent systems.
  - It is a modal logic, extending classical logics with modalities for dynamic actions.
- **Key Components:**
  - **Kripke Models**: Used to represent possible worlds and agents' knowledge about these worlds.
  - **Actions**: Represented as model transformations in DEL, changing the state of knowledge.
- **DEL Formulas:**
  - Basic form: $[A]F$, meaning "after action A, formula F is true."
  - Actions update the Kripke model, reflecting the change in knowledge.

DEL and Neural Networks

- **Applying DEL to Neural Networks**:
  - DEL provides a formal framework to reason about the knowledge evolution within neural networks.
  - Neural network operations can be viewed as knowledge-transforming actions in DEL.

DEL and Neural Networks

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- **Applying DEL to Neural Networks**:
  - DEL provides a formal framework to reason about the knowledge evolution within neural networks.
  - Neural network operations can be viewed as knowledge-transforming actions in DEL.
- **Representing Neural Layers with Kripke Models**:
  - Each layer of a neural network can be represented as a state in a Kripke model, with connections symbolizing possible knowledge transitions.
  - The propositions in the layer/state can model the weights that each neuron "knows", or activation of a neuron.

DEL and Neural Networks

- **Applying DEL to Neural Networks**:
  - DEL provides a formal framework to reason about the knowledge evolution within neural networks.
  - Neural network operations can be viewed as knowledge-transforming actions in DEL.
- **Representing Neural Layers with Kripke Models**:
  - Each layer of a neural network can be represented as a state in a Kripke model, with connections symbolizing possible knowledge transitions.
  - The propositions in the layer/state can model the weights that each neuron “knows”, or activation of a neuron.
- **Modeling Knowledge Flow**:
  - Knowledge flow through layers can be represented as transitions in the Kripke model: $[L_i \rightarrow L_{i+1}]F$, where F is a knowledge state.
  - This reflects how information is processed and transformed across the network.

Methodology Overview

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- **Developing the DEL Framework**:
  - Formulating a DEL-based framework to model knowledge states and transitions in neural networks.
  - Example: Defining a DEL operator $[L]$ to represent a transition through a neural layer L.

Methodology Overview

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- **Developing the DEL Framework**:
  - Formulating a DEL-based framework to model knowledge states and transitions in neural networks.
  - Example: Defining a DEL operator $[L]$ to represent a transition through a neural layer L.
- **Knowledge Representation in Neural Networks**:
  - Representing neural activations, weights, and connections using epistemic models in DEL.
  - Example: Activation A in layer L can be modeled as $[L]A$ in DEL.

Methodology Overview

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- **Developing the DEL Framework**:
  - Formulating a DEL-based framework to model knowledge states and transitions in neural networks.
  - Example: Defining a DEL operator $[L]$ to represent a transition through a neural layer L.
- **Knowledge Representation in Neural Networks**:
  - Representing neural activations, weights, and connections using epistemic models in DEL.
  - Example: Activation A in layer L can be modeled as $[L]A$ in DEL.
- **Modeling Knowledge Dynamics**:
  - Using DEL to express and analyze the flow of information between layers.
  - For layers $L_1, L_2, \ldots, L_n$, knowledge transition can be represented as $[L_1 \to L_2 \to \ldots \to L_n]F$.

Methodology Overview

- **Developing the DEL Framework**:
  - Formulating a DEL-based framework to model knowledge states and transitions in neural networks.
  - Example: Defining a DEL operator $[L]$ to represent a transition through a neural layer L.
- **Knowledge Representation in Neural Networks**:
  - Representing neural activations, weights, and connections using epistemic models in DEL.
  - Example: Activation A in layer L can be modeled as $[L]A$ in DEL.
- **Modeling Knowledge Dynamics**:
  - Using DEL to express and analyze the flow of information between layers.
  - For layers $L_1, L_2, \ldots, L_n$, knowledge transition can be represented as $[L_1 \to L_2 \to \ldots \to L_n]F$.
- **Formal Verification**:
  - Ensuring the consistency and validity of the DEL model across various neural architectures.
  - Techniques such as model checking may be employed for verification.

Methodology Overview

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- **Developing the DEL Framework**:
  - Formulating a DEL-based framework to model knowledge states and transitions in neural networks.
  - Example: Defining a DEL operator $[L]$ to represent a transition through a neural layer L.
- **Knowledge Representation in Neural Networks**:
  - Representing neural activations, weights, and connections using epistemic models in DEL.
  - Example: Activation A in layer L can be modeled as $[L]A$ in DEL.
- **Modeling Knowledge Dynamics**:
  - Using DEL to express and analyze the flow of information between layers.
  - For layers $L_1, L_2, \ldots, L_n$, knowledge transition can be represented as $[L_1 \to L_2 \to \ldots \to L_n]F$.
- **Formal Verification**:
  - Ensuring the consistency and validity of the DEL model across various neural architectures.
  - Techniques such as model checking may be employed for verification.
- **Visualization of Knowledge Evolution**:
  - Developing tools to visualize the changes in knowledge as interpreted by the DEL framework.
  - Aimed at making the understanding of neural network processes more accessible.

Framework Development

- **DEL Framework Foundations**:
  - Establishing the core structure of the DEL framework specific to neural networks.
  - Utilizing Kripke models to represent neural network layers and their interactions.

Framework Development

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- **DEL Framework Foundations**:
  - Establishing the core structure of the DEL framework specific to neural networks.
  - Utilizing Kripke models to represent neural network layers and their interactions.
- **Kripke Models in Neural Networks**:
  - A Kripke model $M = (W, R, V)$ where W represents possible worlds (neural states), R the accessibility relation (transitions), and V the valuation function (neural activations).
  - Each neural layer corresponds to a set of possible worlds in the model.

Framework Development

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- **DEL Framework Foundations**:
  - Establishing the core structure of the DEL framework specific to neural networks.
  - Utilizing Kripke models to represent neural network layers and their interactions.
- **Kripke Models in Neural Networks**:
  - A Kripke model $M = (W, R, V)$ where W represents possible worlds (neural states), R the accessibility relation (transitions), and V the valuation function (neural activations).
  - Each neural layer corresponds to a set of possible worlds in the model.
- **Representing Neural Activations**:
  - Modeling activations as propositional variables in Kripke models.
  - For an activation a in layer L, represented as (L, a) in the model.

Framework Development

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- **DEL Framework Foundations**:
  - Establishing the core structure of the DEL framework specific to neural networks.
  - Utilizing Kripke models to represent neural network layers and their interactions.
- **Kripke Models in Neural Networks**:
  - A Kripke model $M = (W, R, V)$ where W represents possible worlds (neural states), R the accessibility relation (transitions), and V the valuation function (neural activations).
  - Each neural layer corresponds to a set of possible worlds in the model.
- **Representing Neural Activations**:
  - Modeling activations as propositional variables in Kripke models.
  - For an activation a in layer L, represented as (L, a) in the model.
- **Initial Logic Operators**:
  - Defining DEL operators to capture neural processing, e.g., $[L]P$ signifies knowledge after processing by layer L.
  - These operators represent the transformation of knowledge states within the network.

Case Study Selection

Criteria for selecting case studies:

1. Neural networks with different complexities.
2. Tasks with different domains.
3. Examples of how our framework reveals knowledge dynamics in neural networks.

Visualization Tool Development

Planed features:

- Show how epistemic models and action models vary across layers and after forward or backward propagation.
- Let users modify inputs or parameters and see how knowledge dynamics are affected.
- Work with common neural network libraries, making it easy for researchers and practitioners to use it for their models.

Expected Contributions

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- DEL Framework for Neural Networks.
  - Provides a logical perspective on knowledge representation and dynamics.
  - Opens doors for incorporating advanced reasoning methods.

Expected Contributions

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- DEL Framework for Neural Networks.
  - Provides a logical perspective on knowledge representation and dynamics.
  - Opens doors for incorporating advanced reasoning methods.
- Insights on Knowledge Evolution.
  - We expect to gain deeper understanding of the knowledge flow and decision-making processes within neural networks.

Expected Contributions

- DEL Framework for Neural Networks.
  - Provides a logical perspective on knowledge representation and dynamics.
  - Opens doors for incorporating advanced reasoning methods.
- Insights on Knowledge Evolution.
  - We expect to gain deeper understanding of the knowledge flow and decision-making processes within neural networks.
- Visualization Tool.
  - Develop an interactive visualization tool.

Expected Contributions

AFRiTS 2023
Workshop on Automated Formal Reasoning for Trustworthy AI Systems

- DEL Framework for Neural Networks.
  - Provides a logical perspective on knowledge representation and dynamics.
  - Opens doors for incorporating advanced reasoning methods.
- Insights on Knowledge Evolution.
  - We expect to gain deeper understanding of the knowledge flow and decision-making processes within neural networks.
- Visualization Tool.
  - Develop an interactive visualization tool.
- Trustworthy AI Systems.
  - Better understand how neural networks acquire, transform, and use knowledge, and how they justify their outputs.

Challenges & Risks

- The complexity of modern neural networks.
- The ambiguity in epistemic mapping.
- The reception and integration of our approach within the wider AI community.
- Ethical risks associated with increased transparency of neural networks.

Thank You/Contact

Thank You for Your Attention!

Any further questions or discussions can be directed to:

Jefferson O. Andrade

Professor

Federal Institute of Espirito Santos (Ifes) – Campus Serra

Email: jefferson.andrade@ifes.edu.br

Curriculo Lattes: 7138275599443632

ResearchGate: researchgate.net/profile/Jefferson-Andrade