JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Reinforce Certainty In Artificial Intelligence

Author: Shariqua Razi, Department of Computer Science & Engineering, UIET, MDU Rohtak.

CO- Author: Dr Chhavi Rana, Assistant Professor and Ex Coordinator, Department Of Computer Science & Engineering, University Institute of Engineering & Technology, MDU , Rohtak-124001

E-Mail: shariqua.razi.sr@gmail.com

chhavi.uiet@mdurohtak.ac.in

Abstract: Artificial Intelligence is the simulation of human brain power operates by machines, mainly computer systems. Machine vision, expert systems, speech recognition, and natural language processing are their specified applications. As to try hard to ensure that Artificial Intelligence is successfully used in ways that do not harm humanity. Also Artificial Intelligence (AI) software operates techniques like stochastic optimization and deep look-ahead search of huge neural networks to fit gigantic datasets, it frequently results in complicated behavior that is laborious for people to understand. Concerning trust-positioned AI systems, we must not only enhance their firmness but also progress in ways to make their reasoning intelligible. There is increased implementation mechanization in manufacturing, assistance, and communications or transmissions. Besides organizations are positioning AI algorithms in many mission-critical settings. To confidence their deportment it is essential to construct Artificial Intelligence intelligible, either by evolving fresh methods or by using congenitally interpretable models for describing and administrating otherwise immense composite decisions using local estimation, vocabulary calibration, and communal explanation. This paper discusses the significance of Artificial Intelligence safety in various fields. Also, innovative ideas are required for intra-logistics for progressively automated processes.

Keywords: Machine learning

Artificial intelligence

Interpretability

Manufacturing

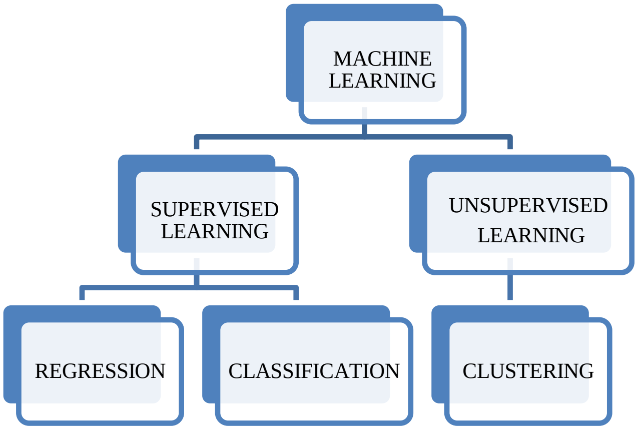
Neural Networks

INTRODUCTION:

Simulating human behavior or intelligence as technology using computers and robots is Artificial Intelligence. A quality of machine doing the tasks that are usually done by humans as they are mind tasks.

Artificial Intelligence is usually categorized into these few categories namely Strong AI, Weak AI, and General AI. Observing, understanding, analyzing, and also self-correction are a few tasks terminologies included in this vast technology. Indulge the working of AI with the structured, unstructured, and semi-structured types of data.

Machine Learning is a component of AI. Forecasting new values by using old data inputs is done by algorithms in Machine Learning. MI is a priority of AI facilitating machinery's ability to understand commands automatically.



Figurre1: Classification in Machine Learning

Both supervised learning and unsupervised learning techniques are included in Machine Learning.

- Models using evidence-making predictions are done in supervised learning.
- Forecasting discrete responses like an email is spam or real is done in the classification technique.
- The regression technique anticipated uninterrupted responses.
- Playing I-spy with structures and patterns is a task of unsupervised learning.
- Clustering catches concealed patterns as well as groupings.

LANGUAGE USED:

For machine learning, web apps, data science, networking apps, AI (artificial intelligence), etc. "Python" is an ideally used language, and has vast standards libraries.

Article 1

Neurons Compositional Explanation

By Jesse Mu in 2021.

METHODOLOGY:

Composition concepts

Future Work:

Model pruning and probe more hidden layers in neurons

Article 2

For Deep RL Counterfactual Survey of Saliency Methods:

By Akansha Atrey in 2020.

Methodology:

Saliency Maps

Future work

Utilize saliency maps as an explanatory tool.

Article 3

Saliency Maps Sanity Updates

By Julius Adebayo in 2020.

Methodology:

Gradient-based approach

Randomization

Visualization

Metrices

Future work:

Try new explanation methods

Article 4

Attention lacking Description

By Sarthak Jain in 2019.

Methodology:

NLP tasks

RNN-based models

BiLSTM architecture

Future work:

Think about seq2seq assignments

Article 5

Language Data Description Understanding

By David Harbecke in 2018.

Methodology:

Data linear model

Linear relation

Future work:

Assess existing signal estimators with the latest criterion.

Article 6

Explicable AI Stakeholders

By Alun Preece in 2018.

Methodology:

Stakeholder communities

Future work:

Finishing disappointed expectations investment

Article 7

Known Consequences RL Contrastive Description

By J. van der Waa in 2018.

Methodology:

MDP (Markov Decision Problem)

Future work:

Focusing multiplex RL benchmark(s)

Article 8

Continuous Explanation for ML Training

By Andrew Slavin Ross in 2018.

Methodology:

Implicit decision methods

Gradient approach

Future work:

Develop greater diagnostic tools

Article 9

Ensemble Tools in Interpretable RL

By Alexander Brown in 2018.

Methodology:

Neural network terminologies

RL algorithm: SARSA

Future work:

Attempt in vast

complex domains

Article 10

DkNN: Regarding Interpretable, Robust and Cofident Deep Learning

By Nicolas Papernot in 2018.

Methodology:

DkNN algorithm

Machine learning models

Future work:

Explore the area of Availability and Integrity benefiting from ML

Article 11

Craft Intelligible Mindset Dare

By Daniel S. Weld in 2018.

Methodology:

Generalised additive models

Future work:

Update an underlying enigmatic model

Article 12

Self-describing Neural Networks regarding Interpretable Robustness

By David Alvarez-Melis in 2018.

Methodology:

Linear regression model

Self-describing models

Black-box methods

Future work:

Assess interpretable models in more knowledgeable areas

CONCLUSION:

With the help of Artificial Intelligence and Machine Learning techniques, on utilizing historical inputs we get understandable or explainable outputs like finding hidden or secret layers, worked with various linear models, gradient-based approach utilization, matrices and more other terminologies are playing major role in the field of machine learning and artificial intelligence. As we conclude from these mentioned previous papers we find work on saliency methodology, deep k-nearest networks, craft intelligible mindset challenge, AI stakeholders and neural networks etc. are the areas have been worked out and requires more enhance safety and security for future purpose to be more efficient and reliable for the users and humans. Implementations done using various programming languages ideally Python. Handling almost every industry with vast knowledge of AI and ML is coming future of living species, as it reduces the hard work and efforts of humans by performing the tasks and not just perform but complete the tasks in given deadline.

REFERENCES:

[1] Jesse Mu, and Jacob Andreas (2021). "Compositional Explanation of Neuron". 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. (2 Feb, 2021) https://proceedings.neurips.cc/paper/2020/file/c74956ffb38ba48ed6ce977af6727275-Paper.pdf

[2] Akanksha Atrey, Kaleigh Clary & David Jensen (2020). "EXPLORATORY NOT EXPLANATORY: COUNTERFACTUAL ANALYSIS OF SALIENCY MAPS FOR DEEP REINFORCEMENT LEARNING."https://scholar.google.co.in/scholar?q=EXPLORATORY+NOT+EXPLANATORY:+COUNTERFACTUAL+ANALYSIS+OF+SALIENCY+MAPS+FOR+DEEP+REINFORCEMENT+LEARNING&hl=en&as_sdt=0&as_vis=1&oi=scholart

[3] Julius Adebayo∗, Justin Gilmer], Michael Muelly], Ian Goodfellow], Moritz Hardt]†,& Been Kim (2020). Sanity Checks for Saliency Maps. https://arxiv.org/abs/1810.03292

[4] Sarthak Jain, & Byron C. Wallace (2019). Attention is not Explanation. https://arxiv.org/abs/1902.10186

[5] David Harbecke, Robert Schwarzenberg, & Christoph Alt (2018). Learning Explanations from Language Data. https://arxiv.org/abs/1808.04127

[6] Alun Preece and Dan, & Dave Braines and Richard (2018). Stakeholders in Explainable AI. https://arxiv.org/abs/1810.00184

[7] J. van der Waa [1], J. van Diggelen[1], K. van den Bosch[1], & M. Neerincx (2018). Contrastive explanations for reinforcement learning in terms of expected consequences. https://arxiv.org/abs/1807.08706

[8] Andrew Slavin Ross, & Finale Doshi-Velez (2018). Training Machine Learning Models by Regularizg their Explanations. https://scholar.google.co.in/scholar?q=Training+Machine+Learning+Models+by+Regularizing+their+Explanations&hl=en&as_sdt=0&as_vis=1&oi=scholart

[9] Alexander Brown and Marek Petrik (2018). Interpretable Reinforcement Learning with Ensemble Methods. https://scholar.google.co.in/scholar?q=Interpretable+Reinforcement+Learning+with+Ensemble+Methods&hl=en&as_sdt=0&as_vis=1&oi=scholart

[10] Nicolas Papernot and Patrick McDaniel (2018). Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. https://arxiv.org/abs/1803.04765

[11] Daniel S. Weld, & Gagan Bansal (2018). The Challenge of Crafting Intelligible Intelligence. https://dl.acm.org/doi/10.1145/3282486

[12] David Alvarez-Melis, & Tommi S. Jaakkola (2018). Towards Robust Interpretability with Self-Explaining Neural Networks. https://arxiv.org/abs/1806.07538