IITB-Monash
Research Academy

Indian Institute of Technology Bombay

MONASH University

An Indian-Australian research partnership

Project Title:	Explainability, Interpretability of deep learning models via the lens of discreteness of deep neural networks	
Project Number	**IMURA1185**	
Monash Main Supervisor (Name, Email Id, Phone)	Pierre Le Bodic pierre.lebodic@monash.edu	*Full name, Email*
Monash Co-supervisor(s) (Name, Email Id, Phone)	Mario Boley Mario.Boley@monash.edu	
Monash Head of Dept/Centre (Name,Email)	Prof. Jianfai Cai	*Full name, email*
Monash Department:	Department of Data Science and Artificial Intelligence	
Monash ADGR (Name,Email)	**Guido Tack** **guido.tack@monash.edu**	*Full name, email*
IITB Main Supervisor (Name, Email Id, Phone)	Balamurugan Palaniappan balamurugan.palaniappan@iitb.ac.in	*Full name, Email*
IITB Co-supervisor(s) (Name, Email Id, Phone)		*Full name, Email*
IITB Head of Dept (Name, Email, Phone)	Jayendran Venkateswaran jayendran@iitb.ac.in	*Full name, email*
IITB Department:	**IEOR**	

Research Clusters:

Research Themes:

Highlight which of the Academy's CLUSTERS this project will address? *(Please nominate JUST **one.** For more information, see www.iitbmonash.org)*		**Highlight which of the Academy's Theme(s) this project will address?** *(Feel free to nominate more than one. For more information, see www.iitbmonash.org)*	
1	*Material Science/Engineering (including Nano, Metallurgy)*	***1***	***Artificial Intelligence and Advanced Computational Modelling***
2	*Energy, Green Chem, Chemistry, Catalysis, Reaction Eng*	2	*Circular Economy*
3	*Math, CFD, Modelling, Manufacturing*	3	*Clean Energy*
4	*CSE, IT, Optimisation, Data, Sensors, Systems, Signal Processing, Control*	4	*Health Sciences*
5	*Earth Sciences and Civil Engineering (Geo, Water, Climate)*	5	*Smart Materials*
6	*Bio, Stem Cells, Bio Chem, Pharma, Food*	6	*Sustainable Societies*
7	*Semi-Conductors, Optics, Photonics, Networks, Telecomm, Power Eng*	7	*Infrastructure*
8	*HSS, Design, Management*		

The research problem

Explainability and interpretability of AI models obtained by training deep neural networks (DNNs) is an important and challenging problem, given the presence of different DNN layers and structures, huge number of weights (often in billions), task dependent training, supervisory information and data distributions. In this project, we aim to look at explainability and interpretability of deep neural network models via discreteness of neural network weights (e.g. binarization, quantization, discretization). We also aim to investigate the relations between learned representations in such settings and the quality of decisions made by the task dependent layers of the DNN. Development of broad theoretical frameworks and developing theoretical tools for specific situations are envisaged. The project will also involve development of training, inference, interpretability and explainability tools and techniques for special cases (e.g. computer vision, audio/speech data, natural language,etc.).

Project aims

The aim of the project is to improve the efficiency, usability and our understanding of Neural Nets and Machine Learning.

What is expected of the student when at IITB and when at Monash?

At IITB, the student will complete the course work components necessary for PhD and will also clear a qualifier exam. The student will start working on the project while being at IITB. At Monash, the student will continue their work, while being exposed to a brand new academic environment.

Expected outcomes

Development of new techniques, algorithms and models in machine learning (ML) and deep learning (DL) have become the norm of the day. Though the utility of new methods developed in ML and DL has grown, it is still not clear how the decisions and results obtained from these methods can be usefully interpreted by a user. Another aspect is that most of these methods lack the ability to explain the rationale behind the decisions they make. The current project will involve developing tools and techniques which would help improving these aspects of interpretability and explainability of deep neural network models. The developed tools and techniques will also help in improving the interpretability and explainability aspects of deep learning methods for diverse types of data (e.g. computer vision, manufacturing systems, speech and audio, etc.)

How will the project address the Goals of the above Themes?

. The project will involve recent developments in using optimization techniques (especially discrete optimization) and development of novel model architectures, training schemes towards these goals.

Potential RPCs from IITB and Monash

Provide names of the potential research progress committee members (RPCs) and describe why they are most suited for the proposed project

Capabilities and Degrees Required

Expected Skillset:

1. Strong background in linear algebra concepts and probability concepts.
2. Good background in machine learning concepts and deep learning concepts.
3. Exposure to discrete optimization and continuous optimization.
4. Good background in computer programming concepts.
5. Expertise (at least 3 on a scale of 1 to 5, 5 being highest) in Python programming.
6. Good communication skills.
7. We are looking for highly motivated and enthusiastic candidates for this project.

Necessary Courses

Name three tentative courses relevant to the project that the student should complete during his/her coursework at IITB (the student will require to secure 8 point in these courses)

1. *A course on Optimization/Discrete Optimization*
2. *A course on Deep Learning/Machine Learning*
3. *A course on Probability and linear algebra.*

Potential Collaborators

Monash has dozens academics in its DSAI department, so there will be no shortage of potential collaborations.

K**eywords** relating to this project to make it easier for the students to apply.

Explainability, Interpretability, Deep learning, Discrete deep neural networks