September 2020

DEEL MAG #06

TOULOUSE & QUEBEC JOIN FORCES TO DEVELOP AI FOR CRITICAL SYSTEMS

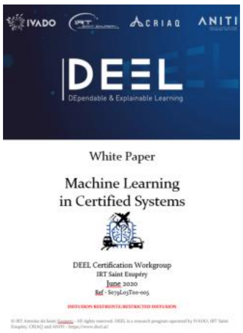
A NEW PARTNERSHIP FOR A MAJOR CATALYST FOR AI PROJECTS

MILA & IRT SAINT EXUPERY joined forces to promote best practices and tackle the scientific challenges of artificial intelligence driven by industry needs and societal impacts. Recognized worldwide for its numerous breakthroughs in artificial intelligence, Mila distinguishes itself both in calibre and in number of its researchers. **The Canadian branch of IRT Saint Exupery was selected to integrate within Mila's ecosystem, considered one of the best in the world for AI research and development,** it is thanks to the partnership established two years ago with the Quebec Institute for data valorization IVADO, which was founded to foster AI partnership projects with industry.

IRT Saint Exupery and its Quebec partners are now developing **new expertise in the field of confidence AI and decision support for critical systems, with applications for land, air and space mobility, but also for health and the environment.**

# Yoshua BENGIO, Geneviève FIORASO, Guillaume GAUDRON

AI CERTIFICATION WORK GROUP: FROM THE BLANK PAGE TO A WHITE PAPER



IVADO ACRIAQ ANITI

DEEL
DEpendable & Explainable Learning

White Paper

Machine Learning
in Certified Systems

DEEL Certification Workgroup
IRT Saint Exupéry
June 2020
Ref - S079L03T00-005

DIFFUSION RESTREINTE/RESTRICTED DIFFUSION

© IRT Antoine de Saint Exupéry - All rights reserved. DEEL is a research program operated by IVADO, IRT Saint Exupéry, CRIAQ and ANITI - https://www.deel.ai/

Using ML techniques in a system submitted to certification constraints raises a lot of fears, questions... and real technical challenges. But what are those challenges exactly? And can they be overcome by selecting appropriate ML techniques, or by adopting new engineering or certification practices?

These are some of the questions addressed by the **ML Certification Workgroup (WG), a team composed of industrial experts and scientists** from various fields (certification, AI, and embedded systems development) , **industrial domains** (aeronautics, railaway, automotive, energy) **and companies** (Airbus, Apsys, Safran, Thales, Scalian, DGA, Onera, SNCF, Continental, Renault, EdF).

In July, after one year of activity, **the WG was proud to deliver a White Paper identifying those challenges**, giving a concrete form to the shared understanding and vision of certification stakes and ML technical issues, and proposing promising research directions. **This White Paper, is organized around seven "main challenges"**: Probabilistic Assesment, Resilience, Specifiability, Data Quality and representativeness, Explainability, Robustness, and Verifiability.

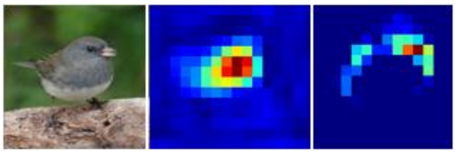
Currently under review by a committee of external experts, it is expected to be made publically available at the end of this year.

# Eric JENN, Grégory FLANDIN, Franck MAMALET

KEY DATES & INFORMATIONS

Project Contract	The DEEL Contract V1 has been signed by all members
Certification Mission	The next workshops : 4th & 5th November * 2sd & 3rd December
Les Carrefours DEEL	1st edition : 1st october, next : 5th November → inscription

ASSESSING THE RELEVANCE OF ATTRIBUTION-BASED EXPLANATION METHODS



In the past few years, many **methods have been proposed to explain the decisions of neural networks.** Among these, **"attribution methods"** are quite popular as they try to assign a score to each input features: the higher the score, the more important the feature in the actual decision. For images, this leads to heatmaps where more meaningful regions are highlighted. Figure 1 shows two different attribution maps from a VGG16 network using the picture of a junco as input.

In spite of their popularity, some attributions methods are not dependable. Indeed, recent works [1, 2, 3, 4] have shown that most of these methods, in particular the ones based on modified back-propagation, are actually not explaining the decision of the network but rather **re-constructing the input image or part of the input image**. In the framework of the DEEL project, we reproduced the experiments in these papers and tried to understand the reasoning behind them. We also stressed several attributions methods to provide recommendations on their reliability. Our experiments confirm the results presented in these works and show that from all the considered attribution methods, only one seems to behave as expected from a dependability point-of-view: **GradCAM**.

1. *Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). "Sanity checks for saliency maps." In Advances in Neural Information Processing Systems (pp. 9505-9515).*
2. *Nie, W., Zhang, Y., & Patel, A. (2018). "A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations." In Proceedings of the 35thInternational Conference on Machine Learning (pp. 3809-3818).*
3. *Gu, J., Yang, Y., & Tresp, V. (2019). "Understanding Individual Decisions of CNNs via Contrastive Backpropagation." In Proceedings of the 14th Asian Conference on Computer Vision (p. 119-134).*
4. *Sixt, L., Granz, M., & Landgraf, T. (2020). "When Explanations Lie : Why Many Modified BP Attributions Fail." In Proceedings of the 37th International Conference on Machine Learning.*

# Mikaël CAPELLE, Florence DE GRANCEY

DEEP REINFORCEMENT LEARNING @MONTREAL

t=0
t=10
t=20
...
t=30
Initial state

Ua
Uc
A
Uo
O
PO-SOKOBAN

Reinforcement Learning is a huge field of AI, where an agent (drone, car, hacker) in its environment (physical world, simulator, rules) tries to learn a "policy" (an intelligent combination of actions) in order to solve one or many problems (usually described by a reward function). **AlphaGo Zero** is maybe the most striking example of RL success, an algorithm having been trained without relying on a data set of examples, just playing against itself, to become the best ever known player of Go, better than AlphaGo that had beaten the world champion of Go. Applications are numerous (robotics, high dimension control, multi-agent collaboration, sim2real…).

As Canada is a major research place on RL, from University of Alberta around Richard Sutton considered as a founder of RL, to Montréal and MILA or McGill where Deep RL – RL enhanced with deep learning – has allowed major breakthrough, IRT Saint Exupery Canada has proposed to explore the DEEL main themes (robustness, explainability) on RL. The idea is to initiate and promote research activities with a potential of local collaborations with academics. Damien GRASSET (IRT Saint Exupery Canada), Patrick SAINT LOUIS (IRT Saint Exupery Canada), Tapopriya MAJUMDAR (Scalian) and Willy LAO (intern from ISAE) explore three main themes, explainable RL, causality in RL, distributional RL, with external collaborators Arthur CHARPENTIER (UQAM), Maxime GASSE (Polytechnique Montréal) and Pierre-Yves OUDEYER (INRIA). IRT Saint Exupery Canada will present each theme in the next newsletters. As there exists a research group on Deep RL at MILA, the more IRT Saint Exupery Canada will work and participate in discussions inside MILA the more interactions with MILA researchers and students will settle.

# Guillaume GAUDRON, Damien GRASSET, Willy LAO, Tapopriya MAJUMDAR, Patrick SAINT LOUIS