# Grounding the information flow for frugal AI

## Overview

The problem of understanding the information flow in Deep Neural Networks (DNNs) is still an object of dispute among the scientific community. The extremely large dimensionality of the latent spaces – requiring an exponentially growing number of samples for the current information entropy estimation – makes the problem difficult to address using conventional information theory tools. Therefore, multiple proposals have been proposed in the last decade, with encouraging results that augmented our understanding related to DNNs. Although these discoveries should have led us to the design of more efficient architectures, we are living in an opposite trend – larger models are being constructed, and therefore paying also a toll in energetic terms. Some applications like surveillance and anomaly detection are reaching performance that was unimaginable just a few years ago, thanks to leveraging general knowledge fit by foundation models, but their implementation directly at the edge, close to the user, right now remains a challenge for computation limits. The solution would be to split the DNN in two and transmit the quantized bitstream; however, finding the best way to split the model is not obvious.

Within this PhD, the candidate will attempt to bridge information flow understanding in DNNs and efficiency and derive properties that will favor no losses while at the same time optimizing bandwidth. The work plan can be decomposed into three major phases:

- development of a theoretical framework that analyzes the information flow in a quantized DNN;
- design of an algorithm able to maximally impose computation bottlenecks to the model while preserving relevant information flow;
- benchmark on real devices and validate the approach.

For the latter, the application of coding for machines to the task of airport surveillance for anomaly detection, provided by a partner startup, is the perfect benchmark to verify the applicability of the developed solution.

This PhD is part of a project funded by the Hi!Paris Society, in the context of a Fellowship program. The candidate will be hosted at Télécom Paris, within the doctoral school of the Institut Polytechnique de Paris.

## Profile

Given the nature of the project, the ideal candidate has an excellent academic record and familiarity with the paper writing and submission process. Besides holding a M.Sc. in Computer Science, Physics, Mathematics or related areas, the ideal candidate should have:

- Strong competencies and knowledge of both Deep Learning and Information theory.
- Proficiency in python coding, and advanced coding skills in any deep learning framework (like pytorch or tensorflow).
- Great passion and commitment to research in AI.
- (optional) Publications in A/A* deep learning venues

## Tutor and material conditions

The main responsible for the project is **Enzo Tartaglione**, and the candidate will be hosted by the Multimedia equipe.
All the material allowing the student to progress in his work (laptop, computing facilities, publication fee coverage, and for missions to be agreed with the supervisor) will be provided.
The total duration of the PhD is normally 3 years.
The PhD starting date will be as soon as possible (depending on administrative times - normally, from the moment of acceptance is no less than 2 months).

## How to apply?

Send CV, exam transcripts and motivation letter to enzo.tartaglione@telecom-paris.fr. The dossiers are screened on a first-come first-served basis.
Curious about meeting the student's equipe and learning about my student's research? Visit the website https://enzotarta.github.io/