Source-channel coding for machine learning

***Laboratories*:**
IETR: Institut d'électronique et des technologies du numérique (https://www.ietr.fr)
Lab-STICC (https://www.labsticc.fr/en/index/)
Start date : October 1st, 2021 **– *duration :*** 36 months

Topic
This PhD position will be conducted in the framework of the project CoLearn of the Labex CominLabs and will be supervised by researchers from INSA Rennes and IMT Atlantique.

Every minute, 500 hours of video are uploaded on Youtube, and 240,000 images are added on Facebook. Since it is physically impossible that this huge mass of data is entirely processed and visualized by humans, there is an absolute need to rely on advanced machine learning methods so as to sort, organize, and recommend the content to users. However, the transmission of the data from the location where they are collected toward the server where they are processed must be done as a preliminary step. The conventional data transmission framework assumes that the data should be completely reconstructed, even with some distortions, by the server. Instead, this thesis aims at developing a novel communication framework in which the server may also apply a learning task over the coded data. We aim at developing an information theoretic analysis so as to understand the fundamental limits of such systems, and develop novel coding techniques allowing for both learning and data reconstruction from the coded data.

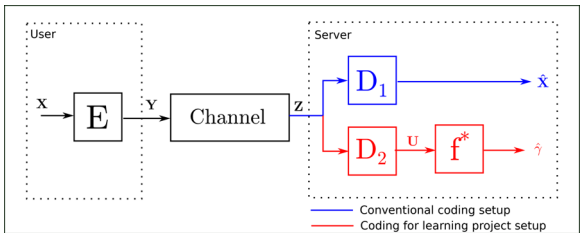

Figure 1: Schematic coding scheme

The problem that is addressed in this proposal is the following and is depicted in Figure 1. A source X is generating data that is encoded and transmitted through a noisy communication channel at a given coding rate R. The server may either reconstruct the data with a certain distortion (decoder D1), or perform learning task with the function f* that follows another decoder D2. As function f*, we may for instance consider a Deep Neural Network already trained for classification, for image segmentation or for content recommendation. For the coding scheme, the design objective is to reduce as much as possible the coding rate R in order to satisfy both the distortion constraint and the learning performance constraint. The main innovation introduced by this project is to consider the part for learning over coded data (in red) together with the conventional setup for data reconstruction (in blue).

To perform learning, one straightforward idea consists in using standard coding techniques for data transmission, and to perform learning after data reconstruction. In our scheme of Figure 1, this would consist of using a standard encoder E designed for data reconstruction, and of building decoder D2 exactly like decoder D1. However, it is questionable whether designing the coding scheme from a distortion point of view may also optimize the learning performance. Hence, the first fundamental question the candidate will address is: "**is there a tradeoff in terms of coding rate between distortion and learning performance?**" Moreover, the source-channel separation theorem states that, under asymptotic conditions, the source coding system and the channel coding system can be designed completely independently from each other, without any loss in performance compared to a joint design of the two systems. Therefore, the second fundamental question which we aim to investigate is: "**is source-channel separation still optimal for learning under both asymptotic or non-asymptotic conditions?**" [1].

The few works in literature that dealt with the tradeoff between reconstruction and learning performance have considered either a particular setup of the general problem depicted here, e.g. [2, 3], or have neglected the channel coding part e.g. [4]. In this PhD, the candidate will consider the general setup depicted above and search for the fundamental information-theoretic limits governing the tradeoff between data reconstruction and

learning performance measure. Moreover, the PhD candidate will investigate the more promising source and channel coding solutions in order to get closer to the bounds that would have been derived in a first step.

One of the envisaged applications is acoustic signal classification from underwater sensors. The data, collected from acoustic sensors, are transmitted via acoustic underwater channel to a gateway in order to be classified, e.g. biological or geological sound. The coding schemes proposed in the PhD may be applied in this context.

Key skills

The candidate should have earned an MSc degree, or equivalent, in one of the following field: information theory, signal processing, applied mathematics. He/She should have a strong background in probabilities and information theory. Some knowledge about the Machine Learning field would also be appreciated. The candidate should be familiar with Matlab and C/C++ language or Python.

Key words:

Asymptotic and non-asymptotic information theory, measure theory, source and channel coding, Machine Learning.

How to apply:

Please send an e-mail to the contacts listed below explaining in a few lines your interest for this subject, and attach:
- Full CV with list project and courses that could be related to the subject
- Complete academic records (from Bachelor to MSc)
- 1 or 2 references
- **Applications will be reviewed when they arrive until one candidate is selected**

Environment and Benefits:

The qualified candidate will be part of the project CoLearn of the *laboratory of excellence* CominLabs, involving researchers from INRIA, Labsticc and IETR. He/She will benefit from the international network of renown experts in information theory and signal processing. In addition:

- 3-year full time employment with fully funded doctoral contract. Remuneration around 1520€/month.
- French national health coverage.
- Partial reimbursement of public transport costs, student housing.
- Approximately 7 weeks of annual leave per year.
- Location: Rennes or Brest. Various stays in both cities, according to the location the candidate is hosted, will be planned all along the thesis.
- Funds to present scientific results at the flagship conferences and workshops of the field.
- Teaching assignment possible but not mandatory.

References

[1] V. Kostina, "Lossy data compression: non-asymptotic fundamental limits", *PhD dissertation*, Princeton University, 2013.
[2] E. Tuncel, D. Gündüz, "Identification and lossy reconstruction in noisy databases", *IEEE Trans. on Inf. Theory*, 2013.
[3] S. Sreekumar, D. Gunduz, "Distributed hypothesis testing over discrete memoryless channels", *IEEE Trans. on Inf. Theory*, 2019
[4] M. Raginski, "Learning from compressed observations", *In Proc. of IEEE ITW*, 2007

Contacts:
Dr. Elsa Dupraz
IMT Atlantique / Lab-STICC
Dr. Philippe Mary
INSA de Rennes / IETR UMR CNRS - 6164

e-mails: elsa.dupraz@imt-atlantique.fr;
philippe.mary@insa-rennes.fr ;
Web sites:
http://elsa-dupraz.fr
http://pmary.perso.insa-rennes.fr