



TAUDoS

Theory and Algorithms for the Understanding of Deep learning On Sequential data

PI: **Rémi Eyraud**. Data Intelligence Team, Hubert Curien
Laboratory, Jean Monnet University – Saint-Etienne,
France





Consortium





Context

- Statistical ML, in particular deep learning, allows great practical results
- However: decision process not accessible to human beings (even Machine Learners!)
- A better **understanding** is needed for business development or even legally required (GDPR).





Overview of TAUDoS

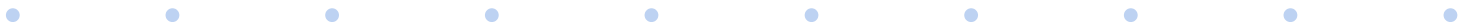
- Focus of this Project : Understanding of neural networks on (discrete) sequential data
- 4 different research paths:
 - Theoretical characterizations
 - Knowledge distillation of grey/white box
 - Learning strategies for interpretability or distillation
 - Definition and learning of metrics for RNN





Theoretical Insights

- **Recent theorems** [Rabousseau et al., 19, Li et al. 20]: linear second order RNN, tensor networks, and Weighted Automata (WA) are **equivalent**
- **Consequences:** Proven learning algorithm for WA extended to RNN
- **In TAUDoS:** extend to other classes (ex: bi-directional RNN and Weighted Context-Free grammars)
- **Possible practical use:** initialization of non-linear RNN





Knowledge Distillation

- **Goal:** Extract simpler, more explicit models from already learn deep networks
- **Recent work:** [Eyraud & Ayache, 20; others] extract Weighted Automata from LSTM/GRU with surprising accuracy
- **In TAUDoS:**
 - direct continuations & improvements of the recent algorithm
 - Subpart detection for subpart distillation





Learning Strategies

- **Goal:** Design new types of layers or of constraints dedicated to understanding of RNN
- **Recent work:** topical subject in the field
- **In TAUDoS:**
 - Discretizing parallel layer to help distillation
 - Compositional constraints (a disentangling approach)
 - Attention with interpretability-based constraints





Metric Learning

- **Goal:** Design and learn metric to compare RNN behavior
- **Recent work:** few...
- **In TAUDoS:**
 - Use the link between WA and 2-RNN (as WA come with a computable distance)
 - Metric learning from Euclidean projection





Valorization

- Open-source toolbox with all the developed approaches
- 2 use cases provided by the firm:
 - Medical (prediction of post-operation complications)
 - Law (Understanding of legal documents)

