2021 Machine Learning Internship Proposals

1. Energy-KPI

We are setting up energy performance indicators. We have several energy related metrics: electricity (including solar cells), water, heat consumptions by buildings, meteorological observations and building occupancy estimates. We are looking for performance indicators which relies on missing variables.

Goals are:

- Analyze dataset with Python major Data Science tools (pandas, sklearn, etc.)
- Model missing variables using analytical and statistical tools
- Implement KPIs and perform error estimation (sensitivity check)
- Classify installations based on their consumptions and behaviors
- Detect abnormal behaviors (failure detection)

Challenges are multiple:

- High dimensionality (>3000 features) and cardinality (3 years of 15 min data)
- Real dataset with outliers (<1%) and inconsistencies/idiosyncrasy (~10%)
- Trendy and hot topic on which we must communicate with accurate and consistent results

2. FaceAudit

This project is about Real Time Tiny Face Detection. That is converting a video stream into a time series of people count.

Goals are:

- Design, train and validate AI with a Tiny ML Python major framework (Keras, TF, etc.)
- Generate accurate audience count in auditorium from a basic fisheye camera
- Draft a white paper with state of art and recommendation for industrial implementation

Challenges are multiple:

- we don't have training dataset (no historical data)
- there are privacy issues linked to dataset collation and usage
- targeted hardware is resource limited (Jetson Nano)
- still an open field of research

3. RTM-ML

Real Time Monitoring is a new project owned by the IT Departement and sponsored by the SmartCampus. We are collecting IT system metrics and logs to ease operations and have an objective baseline of our systems. We are starting by monitoring a hundred systems on the Hyperion hyperconverged platform, but the solution should be extended to a broader scope.

Goals are:

- Analyze dataset with Python major Data Science tools (pandas, sklearn, etc.)
- Reduce cardinality and extract relevant metrics from systems
- Detect specific event (exploit, fraud) from metrics and logs
- Predict system loads

Challenges are multiple:

- High dimensionality (>300 features/systems) and cardinality (few months of 10s data)
- Logs are text-based information related to events with specific time distributions
- Databases are still under development; it will offer some flexibility in trade of low documentation