sherpa.ai

Neural Language Models Explainability & Ethics

AI Research & Development

Master Thesis Project

HAP-LAP 2021 – UPV-EHU

Project Description

Pre-trained neural language models such as BERT or GPT-3 have became the state-of-the-art end-to-end models for natural language understanding. These pre-trained language models allow one to devote the efforts to fine tune these models for specific tasks.

As a consequence of smarter information systems becoming embedded in our society, many countries are developing AI ethics frameworks to address issues about fairness, transparency and accountability in technology. Accordingly, the ability to interpret and explain machine learning models, and particularly deep learning ones, is becoming a hot research topic. **Compared to other trends, the ability to explain predictions in NLP is still limited and researchers advocate for further work in this area.**

The project consists of exploring explainability and other ethical aspects of deep learning end-to-end language models, such as BERT or GPT-3.

Goals

1. Understand the ethical problems raised when using deep learning techniques.
2. Study and reproduce state-of-the-art approaches in explainable AI for NLP.
3. Identify current difficulties and communicate conclusions.

Materials

- Computer

Working Plan & Expected Results

1. Study and reproduce experiments based on https://arxiv.org/pdf/2010.00711.pdf
2. Agree with Sherpa in a particular experiment of interest for Sherpa and the student.
3. Report.

Academic and Industrial Mentoring

- Dr. Miguel A. Veganzones (Sherpa AI Director)
- Prof. Eneko Agirre (IXA Team, EHU-UPV)

Candidate Profile

Basic knowledge of:

- Python
- Natural Language Processing

Interest on:

- End-to-end neural language models
- Ethics and Explainable AI

Benefits and Practical Information

- Funding: 2600€
- Duration: 3-6 Months
- Location: Aula SHERPA, Fac. Informática San Sebastián