Augustas Macijauskas

Email: august.macijauskas@gmail.com • Phone: +37061001317
• LinkedIn: linkedin.com/in/augustas-macijauskas • GitHub: github.com/AugustasMacijauskas

EDUCATION

University of Cambridge (October 2022 – September 2023; Cambridge, United Kingdom)

- Machine Learning and Machine Intelligence (MPhil, 77.66%, **distinction**).
- Notable topics studied: **Deep Learning; Computer Vision; Probabilistic Machine Learning; Neural Machine Translation; Reinforcement Learning; Advanced Machine Learning; Graph Neural Networks**.
- Thesis titled **Eliciting latent knowledge from language reward models** on interpretability and alignment of LLMs. Supervised by Dr Samuel Albanie and Herbie Bradley.

The University of Manchester (September 2019 – June 2022; Manchester, United Kingdom)

- Mathematics (BSc, 91.7%, 1st, 4th rank overall).
- Final project: **Numerical Solutions to the Navier-Stokes Equations**. Supervised by Dr Matthias Heil.

WORK EXPERIENCE and PROJECTS

Eliciting Latent Knowledge from Lange Reward Models (May 2023 – September 2023, Cambridge, UK)

- Created a method that allows using linear classifiers trained on top of a model's activations (referred to as *discovering latent knowledge* (DLK)) **to build reward models that promote truthfulness** (in a narrow sense).
- Utilized the trained reward models to fine-tune pre-trained *large language models* (LLMs) to be more truthful by using the *proximal policy optimization* (PPO) *reinforcement learning* (RL) algorithm.
    - Adopted **efficient fine-tuning strategies**, such as distributed *data-parallel training* (DDP), *low-rank adaption* (LoRA), and *quantization*.
    - Created batch scripts to **automatically launch training jobs** on a computing cluster equipped with **SLURM**.
    - Explored and successfully applied methods to **stabilize and regularize the RL fine-tuning process**.
- Improved the truthfulness of pre-trained LLMs by **up to 1.6%**, as measured by the TruthfulQA benchmark, **without compromising the models' performance on general NLP tasks**.
- Produced a written thesis that was awarded a **distinction-level grade of 78%**.

Baltic Institute of Advanced Technology (BPTI) (*Research Assistant*, July 2020 – September 2022; Vilnius, Lithuania)

- Investigation of object **3D geometry reconstruction** using **neural radiance fields**.
    - Read papers, browsed repositories with implementations and adapted them to our needs.
    - Achieved **satisfactory neural view synthesis and reconstruction quality on a reflective object**.
    - Summarized all the successes and learnings in a scientific report.
- 3D point cloud processing.
    - Replicated the Point Transformer architecture for 3D point cloud classification and segmentation.
    - Tweaked the above model to segment out artificially added noise.
- R&D project in cyber security to research and **improve cyber-attack prediction accuracy**.
    - Compared the ability of various classifiers to detect malicious network packets in manually-generated data.
- Developed a PyTorch model that utilizes **similarity learning using Triplet loss to perform real-world visa stamp recognition** (i.e. classifying the country and direction of travel).
    - Achieved **93% accuracy** on unseen validation data using a ResNet-18 Siamese network architecture.
    - Wrote an API that allowed the team to deploy the trained model for demonstration purposes.
    - Summarized the approach and results in an arXiv preprint: https://arxiv.org/abs/2112.00348.

SKILLS

Programming Languages: Python, MATLAB, JavaScript, C++.

Frameworks and libraries: PyTorch, transformers, datasets, trl, accelerate, PyTorch Lightning, numpy, scikit-learn, fastai.

Soft skills: Leadership, communication, pitching, teaching.

Languages: native in Lithuanian, fluent in English, basic knowledge of Russian.