Aton Kamanda

https://atonkamanda.github.io/ | atonkamanda@hotmail.com | https://github.com/atonkamanda | 438 543-4133

Education

University of Namur Namur
Bachelor of computer science with distinction (minor in mathematics) *Sept. 2018 – Sept. 2021*

University of Montreal Montreal
Master of artificial intelligence, 3.95/4.3 GPA *Sept. 2021 – Aug. 2023*

Experience

Machine learning engineer March 2024 – September 2024
AwakeAI - Mila incubated startup *Canada, Montreal*

- Developed a video understanding system based on V-JEPA for real-time activity recognition in nursing homes achieving state-of-the-art performance of 81% on the Toyota Smarthome benchmark.
- Demonstrated model adaptation skills by fine-tuning the V-JEPA architecture on proprietary data, resulting in 80% accuracy on complex action recognition tasks in real-world scenarios. Enhanced system capabilities by seamlessly integrating YOLOv8 and StrongSORT for robust multi-person tracking and action recognition.
- Leveraged the self-supervised capabilities of V-JEPA to train the model on hundred hours of unlabeled video data increasing the performances on downstream tasks of up to 5%.
- Implemented a highly scalable Kafka-based real-time video streaming pipeline, showcasing advanced optimization techniques that enabled processing of 100+ simultaneous camera streams on a single GPU at 90 frames per second.

.

Machine Learning Research Engineer May 2023 – February 2024
VMware *Canada, Montreal*

- Implemented a personalized large language model (LLM) trained on private codebase for boosting software engineer productivity in production. Achieved remarkable performance optimizations:
  - ∗ Advanced prompt engineering techniques (Chain of thought, few-shot learning, instruction tuning) resulting in a 30% improvement in code generation accuracy.
  - ∗ Designed and deployed a state-of-the-art retrieval-augmented generation system utilizing CodeLlama, Langchain, and ChromaDB on a 50 GB vector database with 0.1 to 10 milliseconds per query.
  - ∗ Optimized a CodeLLAMA 70B model, achieving a 4x reduction in inference time and 10x memory savings by implementing advanced techniques such as variational dropout, pruning, and sparse matrix representations.

Teacher assistant for a graduate deep RL for robots class January 2023 – May 2023
Mila - Montreal institute for learning algorithms *Canada, Montreal*

- Course focused on deep RL for robotics and composed mainly of PhD students, I have been in charge of creating entirely new assignments with recent research papers, writing automated tests on Gradescope, grading students, and helping students in their research contributions for the final project. (Course website).

Machine learning intern June 2021 – September 2021
SkalUP *Belgium, Namur*

- Developped a full-stack search engine using CodeBERT to retrieve code snippets based on natural language queries, rank them by cosine similarity, and direct the user to the specific GitHub repo related to the snippet.

Publications

CodeUltraFeedback: LLM-as-a-Judge for coding preferences alignment | *ACM 2024* March 2024

- Pioneered CodeUltraFeedback, a 10,000-instruction dataset for LLM alignment, and CODAL-Bench for assessing coding preferences; demonstrated that CodeLlama7B-Instruct, fine-tuned with this data using DPO, outperformed 34B LLMs on CODAL-Bench and improved HumanEval+ functional correctness.The project has reached 60+ stars on github.

Technical strengths

Languages : Python, Julia, C/C++, R, SQL
Data & Developer Tools: Spark, Hadoop, Pandas, AWS, GCP, Azure, Kafka, Docker, Kubernetes
Machine learning: Pytorch, Jax, TensorFlow, MLFlow, Triton, NumPy, Gym, Mujoco, TensorRT