

Devin Ha

Berlin, Germany · +491607050562 · t.ha@tu-berlin.de · github.com/DevinTDHa

Work Experience

Machine Learning and Software Engineer

April 2021 – Present

John Snow Labs

(Part-time Remote) Berlin, Germany

- Part of Spark NLP Open Source team, porting state-of-the-art language models to Apache Spark using Scala. Contributions include text embedding models such as RoBERTa, BERT, UAE and multi-modal models such as CLIP, Whisper, Swin, Wav2Vec and VisionEncoderDecoder. Users can then infer these models on a Spark cluster.
- Implemented crucial pre-processing steps for these models in Scala, such as BPE tokenization for language models, image normalization for vision models and mel-filter banks for Whisper.
- Implemented auto-regressive language generation for LLMs in Spark NLP.
- Enabled users to import these models into Spark NLP by providing TensorFlow, PyTorch, and ONNX scripts.
- Optimized performance of existing tokenization algorithms of Spark NLP for XLM-based models, resulting in a significant reduction in memory usage and processing time.
- Built TensorFlow binaries for inference on Apple Silicon machines, published as open source.

Graduate Software Engineer

Nov 2019 – Mar 2021

Amadeus IT Group

London, United Kingdom

- Part of Baggage Component Team, processing baggage to international standards of aviation
- Contributed to C++ based high performance back end for processing thousands of agency queries and API calls to pricing providers
- Ensured quality with automated regression testing with pytest-based framework in a BDD style

Student Assistant

Aug 2017 – Sep 2019

DAI-Labor

Berlin, Germany

- Contributed to full-stack development within DAI's Competence Center for Information Retrieval and Machine Learning, building full stack NLP web applications for user interaction.
- Implemented a chatbot to handle support questions for Deutsche Telekom

Education

Technische Universität Berlin

Expected 2024

M.Sc. Computer Science, specialized in Machine Learning and Data Engineering

Berlin, Germany

Technische Universität Berlin

Nov 2019

B.Sc. Computer Science

Berlin, Germany

Skills

Languages: Scala, Python, Java, C++, Shell Script, \LaTeX , Markdown, Lean 4

Frameworks: PyTorch, Apache Spark, Spark NLP, TensorFlow, ONNX, MLlib, transformers, pandas, numpy, matplotlib

Tools and Libraries: Git, Jupyter Notebook, Linux, Docker

Cloud Tools: AWS, Databricks, Google Cloud Platform

Specialization: Deep Learning, Natural Language Processing, Computer Vision, Applied Mathematics

Languages: English (Fluent), German (Native), Vietnamese (Intermediate)

Projects

Predicting Chemical Quantum Interactions: As part of the M.Sc. curriculum, attempted to recreate a model for [chemical quantum interactions](#). Implemented the network from scratch, resolving numerous numerical issues. Additionally, improved interpretability and performance of the proposed model by employing explainability and contemporary regularization techniques respectively.

Classification of dangerous situations based on car sensor data (Bachelor Thesis): Trained 2 different types of time series models (Long short-term memory and temporal convolutional networks) to classify dangerous driving situations based on sensor data provided by a German car manufacturer. Data processing and model training was done in a distributed manner with Spark. Additionally, data stream processing was simulated with Kafka + Spark.