Vamsikrishna Chemudupati

Experience of 5 years developing Software Systems in the field of Machine learning and Automation. Knowledge of working on NLP, ASR and Time series problem statements. Additionally, having an expertise in deployment strategies on software products.

Contact:
chemudu.vamsi@gmail.com
LinkedIn
Github
Google scholar
+15148159166

WORK EXPERIENCE

Machine learning Engineer
Avid Technology
08/2023 – Present Toronto, Canada

- Productized the Speaker Diarization library in Media Composer application utilized by 5000+ media companies worldwide.
- Developed Speaker Diarization solution using a combination of ASR model (Whisper), speaker embedding model (Resnet34) and multi-stage spectral clustering method to handle different length of audio inputs. Obtained a Diarization error rate of 7.98% when evaluated on multiple conversational datasets.
- Worked with Fine-tuning whisper (Seq-Seq) for low resource languages and accents using LoRA method to improve Word error rate.
- Implemented a video captioning solution using multimodal models. Used GPT-3.5 and open source LLMs for text-based reasoning of video transcript followed by obtaining cross-modal embeddings from BLIP model.
- Deployed the Diarization solution on windows and mac edge devices using Cython and ONNX libraries. Implemented the Elasticsearch Database in C++ to facilitate the metadata storage.

Researcher | NLP | ASR
Huawei Technologies, Noah's Ark Lab
01/2023 – 08/2023 Montreal, Canada

- Evaluated the robustness and cross-task performance of ASR system Whisper by implementing a data pipeline that corrupts input speech using environment noise and room reverberations. Achieved SOTA results in both clean and noisy environments across multiple ASR tasks given in SUPERB benchmark, indicating the potential of deployment in real-time products.
- Working on developing Web-assisted QA systems using RAG methods with MongoDB as the document database. Developed data processing pipelines involving web scrapping and data cleaning from publicly available chatbots, resulting in 500k QA samples. Performed training on GPT based LLM's (LLaMa) using the multidomain data collected.
- Created an automated solution for evaluating the predicted answer using LLM's fine-tuned on Entailment and paraphrase detection modules.

Research Scientist Intern | ASR | Model Compression
Nuance Communications, a Microsoft Company
05/2022 – 12/2022 Montreal, Canada

- Developed a single End-to-End ASR model with configurable accuracy/latency trade-off at runtime using Universally Slimmable techniques and in-place distillation techniques.
- Implemented the channel-wise and depth-wise sliming techniques over Conformer Encoder in RNN-T setup using TensorFlow. Evaluated the WER and Real time factor (RTF) tradeoff for 4 scales by exporting model graph and running it on CPU.
- Obtained a relative WER difference of 1-2% between Jointly trained/single configurable and Baseline dedicated models built with same number of parameters.

SKILLS

Programming: Python, C, C++, SQL, Shell, Matlab, Simulink, Jenkins, Ansible

ML: Tensorflow, Pytorch, Sklearn, Pandas, NumPy, Matplotlib, Flask, Huggingface, LlamaIndex, LangChain, Cython

MLSkills: Natural language processing, Automatic speech recognition, Deep learning, Time series analysis

Tools: Git, Slurm, Azure, Apache Airflow, Docker, REST Api, CI/CD, Elasticsearch

Operating systems: Linux, Windows.

PUBLICATIONS

- Kaushik. M, Ankit R, *Vamsikrishna C*, Xing Han. **TASTY: A Transformer based approach to space and time complexity.** ICLR 2023, Deep learning for code workshop.
- Vamsikrishna Chemudupati *et al.* **On the Transferability of Whisper-based Representations for "In-the-Wild" Cross-Task Downstream Speech Applications.** Arxiv publication

AWARDS & CERTIFICATIONS

- International Student Scholarship (Bourse C) - Offered funding to pursue a Master's degree in Computer Science with University of Montreal.
- Awarded the DIRO x Quebec Ministry of Higher Education international students scholarship worth C$4,000 in 2022 and C$3000 in 2023 respectively.

COURSE PROJECTS

Transformer-based approach to space and time complexity

- Prepared a novel labelled dataset (2800 samples) of time and space complexities for code snippets spanning multiple languages such as Python, C++.
- We study Code-based LM models on novel code comprehension task of cross-language time complexity predictions where we fine-tune on one language and run inference on another language

Machine learning engineer | NLP
Hexaware Technologies
12/2020 – 08/2021 Bangalore, India

- Implemented an automatic IT ticket allocation system using text classification methods involving layers such as LSTM and GRU. Obtained an accuracy of 76% in production on 6 labels using ITSM tool ServiceNow.
- Constructed an ETL pipeline to automate the periodic IT operations on servers at a global level for a reputed banking client using Airflow framework, Rest API, CI/CD tools, Ansible and python.

Machine learning engineer | Time series prediction
Hyundai Mobis
07/2018 - 11/2020 Hyderabad, India

- Implemented a hybrid method combining deep CNN and GRU network for instantaneous vehicle speed estimation in Antilock braking systems.
- Achieved a target Mean Squared Error (MSE) of less than 2% of the maximum vehicle speed considering various conditions of roads with respect to mu. Deployed the Vehicle speed prediction model onto the ECU using Matlab embedded coder toolkit.

Intern
Sony India Software Centre
11/2017 - 06/2018 Bangalore, India

- Worked on the development of a scene recognition deep learning library using Mobilenetv2 network to be deployed on Sony mobile phones.
- Developed an end-to-end ML pipeline for 60000 image dataset preparation, Feature extraction, training the model and evaluating using python and shell scripting. The pipeline processes incoming test data and returns the calculated f-score.

EDUCATION

Master of Science - Computer Science (Machine learning)
University of Montreal (MILA Lab, Montreal)
Sept 2021 – May 2023 Montreal, Canada

- Fundamentals of Machine learning (A)
- Representation learning (A+)
- NLU with Deep learning (A+)
- Data Science
- Applied ML projects
- Neural Scaling laws (A+)

Bachelor of Technology, Electronics and Communication
Vellore Institute of Technology **GPA: 8.75/10.0**
July 2014 – May 2018 Vellore, India

- GraphCodeBERT performs the best in complexity prediction by achieving 78% acc.
- Cross language study reveals BERT and GraphCodeBERT to have the most generalized representations.

Ablation study on Conversational Question Answering system

- Performed a reproductivity challenge for FlowQA and BERT-based Question answering algorithms using the CoQA dataset and Pytorch.
- Performed a comparative study of implemented algorithms based on the criteria of previous history of the conversation included.

Goal predictor system for live NHL games using ML algorithms

- Collected data from NHL website using REST API and developed an interactive widget for analysis. Additionally, performed feature extraction on the dataset to obtain significant features such as Shot angle, Shot Distance, and previous game event details.
- Performed a comparative study of algorithms such as Ensemble models, fully connected networks and Gradient boosting techniques developed for predicting the chances of a Goal.
- Deployed the live predictor system using Flask app and Docker.