Shubham Krishna

Berlin, Germany
+4915162793372 shubhamkrishna.ism@gmail.com shubhamkrishna shub-kris shub-kris.github.io

Technical Skills

Languages & Frameworks: Python, C++, ShellScripting, PyTorch, TensorFlow, Apache Beam
Python Libraries: Scikit-Learn, NumPy, Pandas, Matplotlib, HuggingFace Transformers, OpenCV, Pillow
Other Tools: GCP, AWS (Beginner), GitHub, Bitbucket, Terraform, Docker

Experience

ML6 | Machine Learning Engineer Dec 2021 - Present
ML Delivery Team Berlin, Germany

- Developed an advanced deep learning-based semantic segmentation model for detecting fungal areas in leaves, utilizing an EfficientNet backbone and Feature Pyramid Network architecture. The model achieved exceptional accuracy, improving the F1 score over the baseline random forest model by 42%. The model's success has been validated by plant experts and has is now being used to save time and detect fungus faster than humans. Utilised Google's Vertex AI for developing an end-to-end pipeline, designed for adaptability, reproducibility and capable of detecting fungi on different types of leaves.
- Successfully finetuned and deployed Text2Img[Link] and Image Variations[Link] Stable Diffusion models as a scalable service on AWS EC2 and models integrated into the company's e-commerce platform, allowing users to generate custom artistic images using text and image prompts. Developed a deep learning model for removing the background from generated images. More than 3 million images were generated in the first month. [Link]
- Implemented highly scalable, automated, and robust ETL pipelines for the ingestion of large volumes of catalog (more than 100k) and user events (more than 1.2 million) data everyday using Apache Beam from client's FTP server to Retail API.
- Utilized established pipelines to build and deploy recommendation models: Similar Items and Frequently Bought Together using Google Cloud Retail API for a retail company's e-commerce platform, resulting in a 300k Euros/week revenue increase due to a 40% increase in conversion rate.
- Developed multiple scalable pipelines using Apache Beam RunInference API and machine learning models in PyTorch, Tensorflow, and Scikit-Learn for efficient and accurate inference, including Per Entity Training, Large Language Model Inference and other applications. The pipelines are featured on the Apache Beam website as examples [Link]
- Developed end-to-end GCP Vertex AI based ML pipelines for various tasks, including text-classification, object detection, and others, easily adaptable to other machine learning tasks. Presented at industry meetups and utilized in multiple projects.
- Technical Skills: Data Engineering, Generative AI, MLOps, Computer Vision, CI/CD, Docker, Vertex AI

Bosch Center for Artificial Intelligence | Master Thesis May 2021 - Nov. 2021
Robust Deep Learning Team Renningen, Germany

- Developed a two stage framework for dealing with label noise for semantic segmentation using semi-supervised learning. The framework reduced the noise to 33.33% and improved the mIoU by 9% on corrupted CityScapes validation dataset.

Max Planck Institute for Intelligent Systems | Research Assistant April 2020 - Oct 2021
Bethge Lab, Deep Learning for Computer Vision Tubingen, Germany

- Worked on multiple projects in the field of computer vision, with a particular focus on topics such as invariant representation learning [NeurIPS Workshop Link], pruning to make neural networks more efficient, and self-supervised learning. Contributed to project ideation and hypothesis development, as well as developing robust codebases to validate research findings.
- Technical Skills: Python, PyTorch, PyTorch Lightning, Matplotlib, Pandas, NumPy, Docker, Slurm

Samsung Research | Applied Research Engineer June 2018 - Sep 2019
On-Device Artificial Intelligence Team Bangalore, India

- Developed and commercialized a deep learning-based keyword extraction model for mobile applications, using app descriptions, resulting in a 25% increase in recall for application search on mobile devices.

- Developed an NER model using OpenNLP to support natural language search in Gallery for smartphones, resulting in improved search. Won the Best Demo award at Samsung's Annual Technical Event, 2018.
- Authored two research papers in the domain of NLP, Deep Learning and Information Retrieval.
- Technical Skills: Machine Learning, Deep Learning, Information Retrieval, NLP, Python,

Education

University of Tübingen — Oct 2019 - Oct 2021
Master of Science in Machine Learning — Tubingen, Germany

Indian Institute of Technology (IIT) — Jul 2013 - May 2018
Integrated Master of Technology in Mathematics & Computing — Dhanbad, India

Publications

- Schneider S*, **Krishna S***, et al. "Generalized Invariant Risk Minimization: relating adaptation and invariant representation learning." NeurIPS pre-registration workshop, 2020
- **Krishna S**, Bajaj A, et al. "Learning Mobile App Embeddings using Multi-task Neural Network." International Conference on Applications of Natural Language to Information Systems, 2019
- **Krishna S**, Bajaj A, et al. "RelEmb: A relevance-based application embedding for Mobile App retrieval and categorization." Computacion y Sistemas 23.3, 2019
- **Krishna S**, Billot R, Jullien N. "A clustering approach to infer Wikipedia contributors' profile." International symposium on open collaboration, 2018

Selected Projects

- Developed an NER model for detecting city and country in a given sentence by fine-tuning the BERT model on a custom curated Ultra Entity dataset. Model downloaded over 200k times on Hugging Face. [Link]
- Authored a Medium blogpost on deploying Transformer models and tokenizers in production using Nvidia Triton's Ensemble Model, which has been read by over 3.5k individuals, highlighting the advantages of server-side tokenization in terms of flexibility, artifact management and ease of use. [Link]