




SHUBHAM KRISHNA

Berlin, Germany

☎ +4915162793372 ✉ shubhamkrishna.ism@gmail.com  [shubhamkrishna](https://www.linkedin.com/in/shubhamkrishna)  [shub-kris](https://github.com/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

- Finetuned Text2Img [\[Link\]](#) and Image Variations [\[Link\]](#) Stable Diffusion models for generating stickers, print designs and artistic inspirations. Deployed it as a scalable service on AWS EC2 instance and models are integrated into the company's e-commerce platform, allowing users to generate custom artistic images using text and image prompts. Finetuned a deep learning model (ISNet) for removing solid background from images generated using Stable Diffusion models. Used as a postprocessing step and helps in creation of images with transparent background. More than 3 million images were generated in the first month. [\[TechCrunch Link\]](#)
- 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 Beam from client's FTP server to Google Cloud Retail API storage. Utilized Terraform and GitHub Actions to design and deploy infrastructure for the utilization of GCP services. 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 machine learning pipelines for efficient and accurate inference using Apache Beam RunInference API. Utilized different machine/deep learning models available in PyTorch, Tensorflow, and Scikit-Learn to illustrate how one can use Apache Beam for building machine learning pipelines. The pipelines were contributed to Apache Beam and are available as examples in the Apache Beam documentation. [\[Link\]](#)
- Developed multiple 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

- Tackled the problem of Label Noise in Semantic Segmentation. Developed a two-stage generic framework for reducing amount of noise using semi-supervised learning. The proposed framework can be easily extended to deal with label noise for other computer vision tasks such as object detection and instance segmentation.
- The framework reduced the noise from 100% to 33.33% and therefore helped in improving the mean Intersection over Union (mIoU) by 9% on the corrupted CityScapes validation dataset.
- Technical Skills: Python, PyTorch, Matplotlib, Pandas, NumPy

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 [\[NeurIPS Workshop Paper Link\]](#), and pruning to make neural networks more efficient, and faster. 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 a sophisticated deep learning-based model for keyword extraction, utilizing application descriptions from App Store. Successfully commercialized the model for Samsung's flagship smartphones, where it was triggered daily to generate keywords for newly developed apps. The search index stored the extracted keywords, which led to a significant 25% increase in recall for application search on mobile devices.

- Developed and integrated a machine learning model using Apache OpenNLP for Name Entity Recognition and Stanford Core NLP for processing temporal expressions within the Gallery App, enabling natural language query search functionality. Awarded the Best Demo prize at Samsung’s Annual Technical Event for the Proof of Concept feature developed within the Gallery App.
- Published two research papers showcasing innovative approaches in the field. Presented a novel method for app clustering, classification, and retrieval using app-embeddings at CICLing 2019[\[Paper Link\]](#), resulting in improved end-user experience with mobile apps. Also, published a paper at NLDB 2019 [\[Paper Link\]](#) on a multi-task neural architecture that predicts categorical parameters like app category and ratings by jointly modeling app descriptions and reviews.
- Technical Skills: Machine Learning, Deep Learning, Information Retrieval, NLP, Python

EDUCATION

University of Tübingen

Master of Science in Machine Learning

Oct 2019 - Oct 2021

Tubingen, Germany

Indian Institute of Technology (IIT)

Integrated Master of Technology in Mathematics & Computing

Jul 2013 - May 2018

Dhanbad, India

PUBLICATIONS

- Schneider S*, **Krishna S***, et al. “Generalized Invariant Risk Minimization: relating adaptation and invariant representation learning.” NeurIPS pre-registration workshop, 2020. [\[Paper Link\]](#)
- **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. [\[Paper Link\]](#)
- **Krishna S**, Bajaj A, et al. “RelEmb: A relevance-based application embedding for Mobile App retrieval and categorization.” Computacion y Sistemas 23.3, 2019. [\[Paper Link\]](#)
- **Krishna S**, Billot R, Jullien N. “A clustering approach to infer Wikipedia contributors’ profile.” International symposium on open collaboration, 2018. [\[Paper Link\]](#)

SELECTED PROJECTS

- Developed a 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 blog post on deploying Transformer models and tokenizers in production using Nvidia Triton’s Ensemble Model, which has been read by over 3.8k individuals, highlighting the advantages of server-side tokenization in terms of flexibility, artifact management and ease of use. [\[Link\]](#)