AKSHAY JOSHI

akshay@akjo.tech | +49 155 661 76540 | https://akjo.tech | Munich, Germany | EU Blue Card & US B1

Experience

Senior Applied AI Scientist Oct 2023 - Present
Blinkin GmbH, Germany

- Lead applied researcher & manager of the Agentic RAG & Multimodal Vision Language Model R&D team.
- Optimized a 40B parameter Sparse Mixture-of-Expert VLM using curriculum learning for cross-modal grounding, multi-hop reasoning, and reduced hallucination using SSL, PEFT & RL alignment methods. Attained an MMMU val. score of ~51.4, surpassing Claude Haiku, LLaVa NEXT 70B, Llama 3.2 11B & Phi 3.5 V.
- Co-leading the optimization of Qwen 2 VL 72B-based Video Understanding model for Multimodal Search & RAG. Achieved an initial 3.1% avg. rel. Rank@1 gain across MSRVTT & MSVD text-to-video benchmarks.
- Increased ARR from €80K to €350K yoy by delivering AI PoCs & Multimodal pilots to Wilo, Bosch, Telekom.
- Core contributor to key technical sections of the investor data room for Seed/Series A fundraising, following a successful €2.5 million pre-seed from European Innovation Council (EIC) as part of its deep tech grant.

Machine Learning Researcher (HiWi) Nov 2020 - Jul 2023
German Research Center for Artificial Intelligence, Germany Advisors: Prof. Peter Loos, Peter Pfeiffer

- Developed novel SSL training methods for discriminative Transformer Language Models & Subword Tokenizers for Biomedical NLU. The models are pre-trained on ~40 million clinical research reports from PubMed.
- Achieved ~2% avg. abs. improvement over state-of-the-art transformer models of similar scale in PubMedQA (Question Answering), EBM NLP (PICO), BC5-disease (Medical NER) & BIOSSES (Sentence Similarity) downstream tasks in the Microsoft Biomedical Language Understanding and Reasoning Benchmark.
- Built highly parallel & computationally efficient Semantic Search (retrieves & ranks ~2.5 million documents in <3 sec) & Recommendation System for Smart Vigilance in Medical Product Research & Development.
- Improved model explainability by 6.4% using XAI techniques like Integrated Grads, Grad L2 Norm & MSA.

Joint Research Fellow Feb 2022 - May 2023
Google Research, Switzerland Advisors: Dr. Alessio Tonioni, Dr. Henry Rebecq

- Collaborated with the Machine Perception - AR/VR team to develop DiagramNet, a novel self-supervised (SSL) multitask, multimodal Vision Language Model for Representation Learning over Diagrams. Evaluated the synergy of various SSL objectives & proposed an optimal multitask formulation for pretraining.
- Built a novel multimodal multitask pretraining dataset with 1.87 million elementary diagram samples scraped from the web & paired with corresponding descriptions from curated sources like DBPedia, WikiData, CK12.
- Pretrained DiagramNet outperforms the supervised baseline DQA-Net & sota Pix2Struct models of similar scale, achieving a 28.39% & 22.36% rel. improvement on the AI2D multiple-choice diagram QA benchmark.

Joint Research Fellow Feb 2022 - May 2023
German Research Center for Artificial Intelligence, Germany Advisor: Dr. Simon Ostermann

- Devised novel methods to effectively encode long-form diagram captions & multimodal fusion strategies to mitigate strong priors imposed by language, leading to a 14.3% relative decrease in hallucinations & biases.
- Achieved a 5.36% avg. rel. accuracy increase across multiple chart comprehension benchmarks by developing ViT/CLIP-based methods for Entropy Minimization & Consistency Regularization in SSL pretraining.

Graduate Teaching Assistant Oct 2020 - Mar 2021
UdS Algorithmic Business Research Group, Germany Advisor: Prof. Jana Koehler

- Tutored & graded the 'Architectures for Intelligent Systems' course, which had a cohort of ~45 M. Sc. students from Computer/Data Science, Embedded Systems, Visual Computing, and Bioinformatics majors.
- Performed 2 iterations of development of a reference architecture & corresponding architectural design documents for a cloud-powered Conversational Question Answering Smart Digital Assistant in ~4 months.

Software Engineer Aug 2018 - Sep 2019
AMD R&D, India

- Implemented platform initialization routines of the off-chip phase of Platform Security Processor (PSP) firmware for Ryzen 3000 (Matisse & Castle Peak architecture) processors in ~5 months.
- Extended & validated the support for Microsoft PlayReady DRM protection technology in PSP firmware for Ryzen Pinnacle & Raven Ridge family of desktop & mobile x86 processors in 3 months.

Software Engineering Intern Feb 2018 - Aug 2018
AMD R&D, India

- Developed & validated the AMD Ryzen Master Software Development Kit for CPU/Memory Overclocking (Frequency, Voltage, Timing), Core Parking, and Simultaneous Multithreading utilities.
- In a span of 6 months, delivered ~70 multi-platform Windows SDK APIs. Further, established & documented >100 unit test cases & a detailed test plan for Ryzen Master CLI tool validation.

Education

ETH Zurich, Switzerland Feb 2022 - May 2023
Master of Science Research Thesis GPA: 1.2/5.0 (Best: 1.0)
Research Area: Vision & Language, Visual Reasoning Advisor: Prof. Mrinmaya Sachan

University of Saarland, Germany Oct 2019 - Jul 2023
Master of Science in Data Science & Artificial Intelligence GPA: 1.7/5.0 (Best: 1.0)
Research Area: NLU, Computational Semantics Advisor: Prof. Josef van Genabith

Visvesvaraya Technological University, India Aug 2013 - Jun 2017
Bachelor of Engineering in Computer Science & Engineering GPA: 8.03/10 (Best: 10.0)

Publications

- **Akshay Joshi**, Ankit Agrawal, Sushmita Nair, *"Art Style Classification with Self-Trained Ensemble of AutoEncoding Transformations"*, arXiv:2012.03377, 2020.
- **Akshay Joshi**, Peter Pfeiffer, Annalena Kohnert, Philip Hake, Peter Fettke, Peter Loos, *"PubMedSMBERT: A Pretrained Biomedical Language Model for Medical Smart Vigilance"*, Review, 2023.
- Annalena Kohnert, **Akshay Joshi**, Peter Pfeiffer, *"Evaluating Medical LM Pretraining"*, Review, 2023.

Skills

- **Languages & DB**: Python, C/C++, SQL, MongoDB, Redis, Elasticsearch, Qdrant & Milvus Vector Stores
- **ML Libraries**: PyTorch, Huggingface Transformers, Tokenizers, Accelerate, NumPy, Langchain, JAX/FLAX, PEFT, SciPy, LangGraph, vLLM, Nvidia TensorRT, LlamaIndex, DeepSpeed, FAISS, Detectron2, XGBoost, Pandas, Diffusers, OpenCV, Streamlit, Matplotlib, NLTK, MLFlow, OpenAI, SFT, WandB, NeMo Guardrails
- **Tools & Technologies**: Docker, Kubernetes, PyTest, LoREFT, DPO, RLHF, Causal Inference, Kafka, Spark, LoRA, FastAPI, Captum XAI, Presidio Anonymizer, Reranker, Chain-of-Thoughts, Git, CI/CD, GCP, AWS

Projects

- Low-rank Linear Subspace Representation Finetuning & Multi-Teacher Quantized Distillation of Llama 3/3.1.
- Open-domain Question Answering over KG & Text with Residual Memory Networks & Longformer Attention.
- Exploiting Point-level Correspondences for Self-supervised Dense 3D Point Cloud Understanding.
- RSNA-MICCAI Radiogenomic Classification of 3D fMRI sequences to detect latent MGMT Methylation.
- 3D Pose and Shape Estimation with Stitched Puppet Model & Max Product Belief Propagation.

Achievements

- Achieved a rank in the **top 5**% of the graduating class of Bachelors in Computer Science & Engineering. Total number of graduating students: **106**.