Vasudev Gupta

Email: 7vasudevgupta@gmail.com **GitHub:** github.com/thevasudevgupta **Website:** thevasudevgupta.com
Phone: +91 9992526231 **LinkedIn:** linkedin.com/in/thevasudevgupta

Education

Program	Institution	CGPA / %	Completion
Dual Degree (M.Tech. in Data Science, B.Tech. in Mechanical Engineering)	Indian Institute of Technology, Madras	9.08 / 10	2023
XII (CBSE)	OS DAV Public School, Kaithal	93%	2018
X (CBSE)	OS DAV Public School, Kaithal	10 / 10	2016

Relevant Courses

Artificial Intelligence	Deep Learning | Pattern Recognition & Machine Learning | ML for Engineering & Science Applications
Computer Science	GPU Programming | Programming and Data Structures | Database Management System
Mathematics	Probability, Statistics & Stochastic Processes | Differential Equations | Calculus | Series & Matrices
Data Science	Math Foundations of Data Science | Data Analytics Lab | Introduction to Data Analytics | Big Data Lab
Tools: PyTorch, Triton, Transformers, Datasets, Accelerate, JAX, DeepSpeed, Wandb, Streamlit	**Languages:** Python, Rust, C++, MATLAB, LaTeX

Achievements and Honors

Google Developer Expert (2022)	Recognised as one of the youngest Google Developer Experts in JAX worldwide for open-source work.
TensorFlow (Google) Recognition	Featured in official TensorFlow blog post for outstanding work during Google Summer of Code 2021.
HuggingFace Fellowship	Awarded $3,000 fellowship & featured in HuggingFace’s newsletter for contributions to Transformers.
Open-Source Impact	Earned 200+ stars on GitHub, widely recognised by open-source community for contributions to AI.
Inter IIT Tech Meet (2021)	Secured ‘Gold’ in 9th Inter IIT Tech Meet organised by IIT Guwahati, representing the IIT Madras team.
JEE Advanced (2018)	Secured All India Rank (AIR) of 3331 with a percentile score of 98.89% in the JEE Advanced 2018 exam.

Professional Experience

AI Pre-training Lead, Unbox AI - CA, USA (Remote) June’21 - Present

- Handpicked by Prof. Gunnar Carlsson (Stanford) & Mr. Rickard Gabrielsson (Stanford & MIT) to help lead AI efforts at Unbox AI, an emerging leader in building advanced foundation models like BehaviorGPT that anticipate user needs, serving major retail clients, including Klarna.
- Helped scale company 5x, managing over $1M in compute. Promoted to partner in 2 years, played a key role in securing Klarna as a client.
- Led a cross-functional team to build a scalable codebase for large models and worked with Stanford PhDs to bring research into production.
- Scaled BehaviorGPT pre-training to billions of parameters & 1T tokens on 96 H100s, and studied emergent properties with increase in scale.
- Iteratively optimised training data and carefully refined learning objectives to align model pre-training with desired production use cases.
- BehaviorGPT-powered search & recommendations receive 70M+ visits monthly and boost sales by 20% for a leading e-commerce company.
- Implemented Triton kernels to increase memory efficiency and achieve linear scaling with nodes, resulting in a 5x speedup over PyTorch.

Open-Source Development

Google Summer of Code, TensorFlow May’21 - Aug’21

- Selected for the highly competitive Google Summer of Code, a global program that funds top students to work on open-source projects.
- Implemented Meta’s Wav2Vec2, built a library for training speech-to-text models on TPUs, and earned over 90 stars & 29 forks on GitHub.
- Trained the model on 300 GB of LibriSpeech data on TPU v3-8 and advanced data streaming, and achieved a 3% word error rate on test set.

HuggingFace (Transformers, Accelerate, Hub) Feb’21 - June’21

- Selected by HuggingFace, a $4.5B AI startup, to contribute BigBird to its widely-used Transformers library, impacting researchers globally.
- Implemented BigBird model in PyTorch and JAX, including training scripts, quickly adopted by over 200,000 users within months of release.
- Integrated Microsoft’s DeepSpeed with HuggingFace Accelerate and added support for distributed strategies such as ZeRO-3, CPU-offload.
- Contributed ModelHubMixin to enable the upload of PyTorch model to Hub. ModelHubMixin was presented at PyTorch Ecosystem Days.

Projects

Triton GPT	• Implemented GPT-2 model in Triton and achieved a 2x speed and memory reduction over PyTorch. • Implemented FlashAttention algorithm from scratch to achieve optimized and efficient execution.
BioBigBird: Leveraging Long Articles for Biomedical Understanding	**Guide:** Dr. Nirav Bhatt (Professor, IIT Madras) • Developed data pipelines to filter high-quality pre-training text data from 300 GB of raw text data. • Pre-trained BigBird on 42M articles from PubMed and achieved 86.46 BLURB score & 9.76/10 GPA.
Long document Question Answering	**Guide:** Mr. Patrick von Platen & Dr. Thomas Wolf (HuggingFace) • Fine-tuned BigBird on 100 GB of natural-questions on TPU v3-8 and earned over 47 stars on GitHub. • Achieved Exact Match score of 55% and surpassed BigBird paper’s reported score of 53% on test set.
Reproduced NeurIPS’20 paper	**Guide:** Dr. Anurag Mittal (Professor, IIT Madras) • Successfully implemented and reproduced results of NeurIPS’20 paper - ‘Incorporating BERT into parallel sequence decoding with Adapters’ for CS6910 (Deep Learning), and received a 9/10 grade.
Optimizing Adapters for Neural Machine Translation	**Guide:** Dr. Nishant Sinha (Founder, OffNote Labs) • Trained BART for translation tasks & achieved BLEU scores: 25.3 on HIN→ENG & 18.1 on GUJ→ENG. • Reduced memory footprint using adapters by over 76% without significant loss in translation quality.
Quick: PyTorch trainer	• Developed a framework using PyTorch & DeepSpeed for large-scale training with support of ZeRO-2.

Blog Posts

Understanding BigBird’s Block Sparse Attention	• Authored an official blog post on Hugging Face, providing an in-depth analysis of BigBird attention. • Reached Hugging Face’s global audience, highlighting key contributions to open-source AI research.
Optimizing Adapters for Neural Machine Translation	• Authored a comprehensive blog post with OffNote Labs on the effectiveness of adapters in Transformer models for multilingual machine translation, offering strategies for training adapters.

Extracurricular Activities

Google Summer of Code’2022 Mentor May’22 - Aug’22

- Selected by Google to mentor for Google Summer of Code 2022 project titled ‘Developing NLP Examples Using Hugging Face Transformers’.
- Successfully guided participants in implementing, optimizing and training state-of-the-art NLP models for impactful real-world applications.

Strategist, Analytics Club (CFI) Mar’20 - Mar’21

- Contributed to the AI community at IIT Madras by mentoring five projects and guided over 15 students in developing innovative solutions.
- Led a team of over 8 members, attracted more than 300 participants from various engineering disciplines for live sessions and workshops.