Vasu Sharma

https://vasusharma.github.io/
sharma.vasu55@gmail.com | +1(412)-616-6880

EDUCATION

SCHOOL OF COMPUTER SCIENCE, CARNEGIE MELLON UNIVERSITY

Masters in Language Technologies
4.19/4.33 (Dept. Rank: 1)

INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

B.Tech. in Computer Science and Engineering
Cum. GPA: 9.9/10.0

ST. COLUMBA'S SCHOOL

AISSCE (Class XII, CBSE)
Percentage: 97%
IIT-JEE All India Rank 165

AREAS OF INTEREST

Deep Learning
Computer Vision
Natural Language Processing
Speech and Music Processing
Machine Learning Algorithm design

RELEVANT COURSEWORK

Deep Reinforcement Learning (A+)
Neural Networks for NLP (A)
Deep Learning (A+)
Advanced Machine Learning (A+)
Advanced Multi Modal Machine Learning (A+)
Recent Advances in Computer Vision (A)
Natural Language Processing (A)
Visual Recognition (A)
Machine Learning Techniques (A)
Human Centered Computing (A*)
Human Cognitive Processes (A*)
Advanced Algorithms (A)
Data Structures and Algorithms (A)
Digital Signal Processing (A)
Probability and Statistics (A*)
Linear Algebra (A*)
Data Science Specialization(Coursera)

WORK EXPERIENCE

APPLIED RESEARCH SCIENTIST

Meta AI (Facebook AI Research)
FAIR team| Aug 2022 - Present| Menlo Park, CA, USA

- Working on text conditioned aligned audio-video generation models for enabling automatic and scalable large scale content creation leveraging state of the art generative AI models
- Working on creating a multimodal version of chatGPT like model allowing incorporating multimodal generative responses including text, images, audio and video as a part of a multi turn dialog also supporting multimodal inputs
- Created a large scale Audio-Video self supervised learning based representation learning model called MaVIL which leverages masking and inter and intra modal reconstruction objectives to achieve state of the art performance beating even state of the art supervised approaches
- Worked on creating and benchmarking billion scale mutimodal datasets and enabling training of state of the art contrastive learning based CLIP like models
- Created a text quality aesthetic score pipeline to perform effective data pruning allowing large language models to obtain state of the art performance despite using only 50% of the original data

APPLIED SCIENTIST

Amazon Lab 126
Alexa AI team| Aug 2021 - Aug 2022| Sunnyvale, CA, USA

- Created a new benchmark for dialog enabled visual-language navigation as a part of the Alexa Prize Simbot challenge leveraging the multimodal data sources to faithfully navigate a virtual environment based on user instruction and designed benchmark models for the same
- Designing efficient multimodal transformers to speed up their training and deployment by improving the computational complexity of the self attention mechanism
- Video processing applications like video action recognition, video question answering, video summarization, moment retrieval etc working directly with compressed video streams
- Created a benchmark for cooperative heterogenous multi agent reinforcement learning platform including open sourcing the collected dataset and it's benchmark models
- Worked on creating a massively multimodal transformer pipeline capable of handling a wide range of input modalities with modality agnostic transformer blocks which work well for a several tasks leveraging a multitude of modalities

QUANTITATIVE RESEARCH ANALYST

Citadel LLC
Global Quantitative Strategies Team| Aug 2019 - Aug 2021| Chicago, USA

- Working on cross asset class Alpha Construction using hybrid linear and non-linear fitting techniques to best exploit statistical arbitrage opportunities in financial markets
- Building Automated fitting pipeline for incorporating advanced Machine Learning techniques into trading strategies
- Optimizing existing machine learning framework to scale upto large volumes of incoming data and improve performance and speed

INTERNSHIPS

CITADEL LLC

Summer Intern, Machine Learning Team

Global Quantitative Strategies | May 2018 – Aug 2018 | Chicago, USA

- Worked on "Deep Neural Networks for Time series modelling of financial markets" and "Effective Feature scalability for Machine Learning models". In this project I explored a variety of Deep Learning models and effective training techniques to perform time series analysis on the large scale and highly noisy financial markets data. I also ensured that the models scale to arbitrarily large dimension feature sets.

UNIVERSITY OF TORONTO

Summer Intern, Machine Learning Team

Raquel Urtasun, Sanja Fidler | May 2016 – Jul 2016 | Toronto, Canada

- "FlowSeg: A Deep Learning based approach for simultaneous semantic segmentation and flow estimation from videos"
- The project focused on building Deep Convolutional Neural Network architectures to study the problem of Instance and Semantic segmentation of videos. We experiment with fairly advanced and novel Deep CNN architectures to jointly estimate semantic segmentation and flow from videos. The approach shows promising results on various datasets.

ABZOOBA INC.

Technical Consultant

Labhesh Patel | Aug 2016 – Jul 2017 | California, USA (working remotely)

- Worked on building "A Smart E-commerce Virtual Assistant". Implemented features like cloth parsing from images, similar image retrieval from a huge fashion catalogue and a state of the art Deep Recommender system.
- Implemented a Multi Turn Conversational Voice Agent to facilitate user interaction. Involved the use of Memory Networks and a soft attention mechanism over previous queries and responses to figure out the best response to a given user query.
- Also worked on "Query based document retrieval" by learning rich semantic document embeddings using a deep LSTM pipeline and using these to find the match the queries to relevant documents
- "Abstractive summarization using Attention based encoder-decoder networks" : Worked on building a deep residual LSTM pipeline which used temporal attention over both encoder and decoder networks to generate an abstractive summary of documents.

CARNEGIE MELLON UNIVERSITY

Summer Intern, School of Computer Science

Bhiksha Raj, Rita Singh| May 2014 – Jul 2014 | Pittsburgh, USA

- "Deep Recurrent Gated Neural Networks for Dynamic Audio Denoising"
- The project focused on construction of a Deep Recurrent neural network to achieve signal reconstruction by denoising noise corrupted signals by dynamic spectral subtraction.

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE (EPFL)

Summer Intern, Machine Learning and Optimization Lab

Martin Jaggi | May 2017 – Jul 2017| Lausanne, Switzerland

- "Learning semantic sentence embeddings using Hierarchical Convolutional Neural Networks
- In this project I worked on creating Deep Hierarchical Convolutional Neural Networks to learn unsupervised semantic textual embeddings. The representations learnt capture both local and global textual information and hence perform competitively against major state of the art approaches on both supervised tasks like sentiment analysis and unsupervised ones like similarity matching.

XEROX RESEARCH LABS, EUROPE

Research Intern, Computer Vision Team

Diane Larlus, Albert Gordo | Sep 2015 – Dec 2015 | Grenoble, France

- Worked on "Large Scale Image Recognition using Deep Convolutional Neural Nets"
- The projects primarily focused on constructing Deep Learning frameworks for Image Recognition. Worked on designing some novel Deep Learning frameworks for the image recognition task on the ImageNet dataset. Also made extensive use of GPUs and the popular Caffe library for training Deep Convolutional Neural Nets.

XEROX RESEARCH LABS, INDIA

Research Intern, Speech Processing Team

Vivek Tyagi | May 2015 – Sep 2015 | Bangalore, India

- Worked on 3 projects during this internship: "Application of Deep Learning for Automatic Speech Recognition", "A comprehensive analysis of Activation Functions in Deep Nets" and "A new hashing technique to enhance Deep Net performance " . Also got the Best Project award for the same.

PUBLICATIONS

- "MAViL: Masked Audio-Video Learners"
  Po-Yao Huang*, **Vasu Sharma*** , Hu Xu, Chaitanya Ryali, Haoqi Fan, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, Christoph Feichtenhofer
  Under Review at ICML, Hawaii, USA, 2023

- "Attend, Attribute & Attack Model: Multimodal Adversarial Attacks to Investigate Vulnerability in Visual Question Answering Models"
  Vasu Sharma* , Ankita Kalra, Vaibhav, Simral Choudhary, Louis-Philippe Morency
  Under Review at CVPR, Vancouver, Canada, 2023

- "E-ViLM: Efficient Video-Language Model via Masked Video Modeling with Semantic Vector-Quantized Tokenizer"
  Vasu Sharma* , Jacob Fang, Skyler Zheng, Robinson Piramithu
  Under Review at ICCV, Paris, France, 2023

- "CHMARL: A Multimodal Benchmark for Cooperative, Heterogeneous Multi-Agent Reinforcement Learning"
  Vasu Sharma* , Prasoon Goyal, Kaixiang Lin, Govind Thattai, Qiaozi Gao, Gaurav Sukhatme
  Published at Robotics Science and Systems (RSS), New York, USA, 2022

- "Attend and Attack: Attention Guided Adversarial Attacks on Visual Question Answering Models"
  Vasu Sharma* , Ankita Kalra, Vaibhav, Simral Choudhary, Louis-Philippe Morency
  Published at NIPS 2018 (Security in Machine Learning Track), Montreal, Canada

- "Community Regularization of Visually-Grounded Dialog"
  Swaminathan Gurumurthy*, Akshat Agarwal*, **Vasu Sharma*** , Katia Sycara
  Published at International Conference on Autonomous Agents and Multiagent Systems (AAMAS), Montreal, Canada, 2019

- "Multimodal Behavioral Markers Exploring Suicidal Intent in Social Media Videos "
  Vasu Sharma* , Ankit Shah*, Vaibhav*, Mahmoud Al Ismail*, Jeffrey Girard, Louis Philippe Morency
  Published at 21st ACM International Conference on Multimodal Interaction (ICMI) 2019
- "BioAMA: Towards an End to End BioMedical Question Answering System"
  Vasu Sharma , Nitish Kulkarni, Srividya Pranavi, Gabriel Bayomi, Eric Nyberg, Teruko Mitamura
  Published at Annual Meeting of the Association for Computational Linguistics(ACL), BioNLP track, Melbourne, Australia 2018

- "Mind Your Language: Learning Visually Grounded Dialog in a Multi-Agent Setting"
  Akshat Agarwal*, Swaminathan Gurumurthy*, **Vasu Sharma*** , Katia Sycara
  Published at CVPR, VQA Challenge and Visual Dialog Workshop, Salt Lake City, USA, 2018

- "Cyclegen: Cyclic consistency based product review generator from attributes"
  Vasu Sharma , Harsh Sharma, Ankita Bishnu , Labhesh Patel
  Published at International Conference on Natural Language Generation (INLG 2018), Tilburg, Netherlands, 2018

- "CorrNet: Enhancing Facial Action Unit Detection using Inter Action Unit correlations"
  Vasu Sharma* , Satrajit Chatterjee, Amir Zadeh, Louis-Philippe Morency
  Accepted at FG 2020 : IEEE International Conference on Automatic Face & Gesture Recognition, Buenos Aires, Argentina, 2018

- "Induced Attention Invariance: Defending VQA Models against Adversarial Attacks"
  Vasu Sharma* , Ankita Kalra, Louis-Philippe Morency
  Published at NIPS 2019 (ViGIL workshop), Vancouver, Canada

- "Segmentation Guided Attention Networks for Visual Question Answering"
  Vasu Sharma , Ankita Bishnu, Labhesh Patel
  Published at Annual Meeting of the Association for Computational Linguistic, ACL-SRW, Vancouver, Canada, 2017

- "Automatic tagging and retrieval of E-Commerce products based on visual Features"
  Vasu Sharma , Harish Karnick
  Published at NAACL-SRW, Association for Computational Linguistics(ACL) conference, San Diego, 2016

- "A Deep Neural Network Based Approach For Vocal Extraction From Songs"
  Vasu Sharma (single authorship)
  Published at IEEE's International Conference on Signal and Image Processing Applications 2015

OTHER PROJECTS

- "Adversarial Attacks on Visual Question Answering Models"
  Course Project: Advanced MultiModal Machine Learning (Prof. Louis-Philippe Morency)
  Received the overall best project award . The project involved designing adversarial attacks against Visual Question Answering models where we use the implicit attention maps to focus our attacks on the crucial image regions. We also augment the visual noise addition based attacks with language distortion based attacks. We further extend these attacks to BlackBox model based attacks where the architecture of the model under attack is unknown by using a student model and distilling the knowledge of the teacher model onto the teacher by training it to mimic the teacher model's responses. We then use the knowledge obtained from these attacks to identify key vulnerabilities in these architectures. Based on this knowledge, we build learning algorithms more robust to such attacks and making them safe for AI critical applications.

- "3Dify: Automatically convert 2D images and videos to 3D using Deep Neural Networks"
  Course Project: Visual Learning and Recognition (Prof. Abhinav Gupta)
  Received the overall best project award . Created a Deep Convolutional Network based pipeline to automatically learn 3D anaglyph maps from 2D images. Unlike other models this is trained directly on 3D images and learns the depth maps as implicit representations rather than learning them explicitly.

- "Video Action Recognition from Compressed Video Streams"
  Independent Study with Prof. Bhiksha Raj and Caiming Xiong (AI director, Salesforce (Metamind))
  In this project we designed Deep Neural Network pipeline to perform video action recognition directly from videos compressed with H.264 video compression (almost 200x compression ration). Most prior works deal with uncompressed video inputs which is realistically infeasible. We design pipelines which can work with the anchor image frames (I frames) and the motion vector encoding frames (P frames) as used in the compressed video and perform action recognition directly from it.

- "Multi Agent Deep Reinforcement Learning for Co-operative Visual Dialog"
  Course Project: Deep Reinforcement Learning (Prof. Ruslan Salakhutdinov)
  The project involved building 2 conversational voice agents conversing autonomously with each other to play the "20 Questions Image guessing game". We casted the problem in a Multi Agent Dialog setup which allowed the bots to adhere to natural language and avoiding language divergence properties as experienced in the prior state of the art works. We outperform all state of the art methods on both numerical metrics and human evaluation.

- "Data is the New Oil: Learning to Answer Questions in an Active Learning Setting"
  Course Project: Neural Networks for NLP (Prof. Graham Neubig)
  In this project we design an Active Learning pipeline to generate hard examples to train a question answering model on the Movie QA dataset. We use a question generator network which generates questions which are further checked for grammatical and formational correctness by a discriminator. A Query by committee ensemble network is then used and the inter classifier confusion is used to select hard examples for active learning. Our pipeline managed to attain accuracy comparable to the state of the art with 75% lesser data.

- "End to End pipeline for Question Answering"
  Course Project: Question Answering (Prof. Eric Nyberg, Teruko Mitamura)
  In this project we design an end to end Question answering pipeline based on Joint Co-attention answer generation networks. We follow a pipeline of relevant snippet ranking, sentence selection and summary generation for the ideal answer type questions in the BioASQ challenge. **We achieved 1st position in the BioASQ challenge and were ranked on the highly competitive MSMarco and the SQuAD leaderboards**

- "Real Time Video Surveillance using Deep Convolutional Neural Networks"
  Course Project: Machine Learning Techniques (Prof. Harish Karnick)
  Built a real time surveillance system which included object and entity detection and localisation along with face recognition and abnormal action detection. In this project we extended the Faster RCNN model and added time recurrent connections to model context across the video frames. The Face recognition and and abnormal action detection networks were integrated into this Recurrent Faster RCNN model using a novel combination layer and the whole network was trained in a joint end to end manner.

- " Visual Storytelling"
  Course Project: Recent Advances in Computer Vision (Prof. Gaurav Sharma)
  This task entails producing story like descriptions for a sequence of images. I experimented with a unique GRU based decoder which looks at all the encoder states simultaneously which allows the model to peek into relevant parts of the encoder states using a soft attention mechanism. I also use a bidirectional encoder and also implemented my own custom version of the beam search algorithm in a more parrallelized fashion rather than the traditionally used sequential version.