Viraj Bagal

Machine Learning Research Fellow

ADDRESS
64, Safalya Bungalow, Dnyaneshwar Nagar
Ekvira Chowk, Pipeline Road
Ahmednagar, Maharashtra - 414003

7219652039
virajbagal12@gmail.com
Viraj Bagal
Viraj Bagal

ABOUT MYSELF

I am currently a final year BS-MS student pursuing major in **Physics** and minor in **Mathematics**. I am working on my Master's thesis at IIIT-Hyderabad as a Research Fellow in the **Healthcare with AI (HAI)** team. **Click here for the blog**. I am interested to work on challenging problems in **Computer VIsion (CV) and Natural Language Processing (NLP)**. I am experienced in **multimodal data representation learning (CV + NLP), generative networks, interpretability, multi-GPU DDP training and Docker.** Click here for my webage

WORK EXPERIENCE

CCNSB Lab, International Institute of Information Technology (IIIT), Hyderabad
Research Fellow (May 2020 - April 2021)

- **Viraj Bagal et al., 'LigGPT: Molecular generation using transformer decoder'. Shorter version accepted at AAAI-SDA 2021 workshop. The longer version under review at ACS Central Science Journal.** Link to the paper
- Aim was to generate molecules conditioned on multiple physicochemical properties as well as scaffolds. Such models can act as catalysts in the drug discovery process.
- Initially, I approached the problem with graph neural networks. Implemented **vanilla GNNs, GCNs, and GATs using Pytorch and Geometric Pytorch**. Trained models to maximize the likelihood of co-occurrence of the molecule and its properties. Properties were obtained using **RDKiT**. Validity, Uniqueness, Novelty, Internal Diversity and Mean Absolute Difference score of the generated molecules were the metrics used.
- Implemented **multi-GPU DDP training on slurm using sbatch scripts and Pytorch Lightning. Collaborated with my partner using Docker.**
- Moved on to custom small **transformer decoder model similar to GPT and SMILES representation** of molecules as conditional generation using GNNs wasn't producing desirable results. Models trained on next token prediction. Trained on ~3 Million molecules.
- Our model achieved **new state-of-the-art results** in terms of the above metrics on the GuacaMol dataset and competitive results to graph-based approaches on the MOSES dataset.
- Our model can generate molecules having particular values of certain properties like **QED score, logP, TPSA, SAS**. Moreover, our model can generate molecules of certain scaffolds as well. This is extremely useful for **one-shot lead optimization.**
- Interpretability of the generative process addressed by **saliency maps**. Gradients of output wrt input tokens calculated to obtain their importance.

CVIT Lab, International Institute of Information Technology (IIIT), Hyderabad
Research Fellow (May 2020 - April 2021)

- **Viraj Bagal, Yash Khare, et al., MMBERT: Multimodal BERT for Improved Medical VQA. Paper accepted at ISBI 2021.**Link to the paper
- Aim was to build an **interpretable medical visual question answering system** to answer medically relevant questions on radiology images.
- Due to scarcity of labelled data, I implemented **self-supervised training** with pretext tasks such as **Image-Text Matching and Masked Visual- Language Modelling**.
- For it, I modified BERT to take text as well as image features as input. **Hugging Face transformers library and Pytorch used. Mutli-GPU DDP training carried on slurm using sbatch scripts and Pytorch Lightning. Collaborated with my partner using Docker.**
- Finetuned the pre-trained model on downstream VQA. Achieved **new state-of-the-art (SOTA) performance on ImageClef 2019 and VQA-RAD datasets.**
- Our single model **outperforms the ensemble** of previous SOTA models
- Interpretability addressed via **attention maps**. Collaborated with a doctor for qualitative interpretation of the attention maps.

Indian Institute of Science Education and Research (IISER), Pune
Research Student (December 2019 - May 2020)

- Aim was to build models for **distinguishing fake electrons** at the Large Hadron Collider, CERN.
- Analysed Drell Yan process obtained from CERN Opens Source using **nanoAODs in C++ and ROOT**.
- Wrote **C++ code for grouping different particle collections** and analyzing their properties such as momentum, energy using distribution plots in ROOT.
- Created dataset of the images of **Calotower collection** within a cutoff dR of electrons from collision data using C++ and ROOT.
- Implemented end to end pipeline for faster experimentation of training **CNNs** for the identification of fake electrons using **Pytorch.**
- Analysed results using **probability histograms and ROC curves using seaborn, matplotlib, sklearn.**
- Achieved **81% accuracy in identifying fake electrons.**

EDUCATION

Indian Institute of Science Education and Research (IISER), Pune
Integrated BS-MS Physics and Mathematics (July 2016 - April 2021)
GPA - 9.3

- **Courses in Physics**
  - Classical & Quantum Physics, Statistical Physics, Condensed Matter Physics, Quantum Field Theory, Atomic & Molecular Physics, Particle Physics, Optics
- **Courses in Mathematics**
  - Linear Algebra, Single & Multivariable Calculus, Probability, Statistics, Set theory

PERSONAL PROJECTS

Mixed Sample Data Augmentations (MSDAs)
https://github.com/ViraJBagal/FMix-Paper-Implementation (May 2020 - June 2020)

- Reproduced, Ethan Harris et al. **FMix: Enhancing Mixed Sample Data Augmentation** paper.
- Compared the performance of **FMix, Cutmix , Mixup and Baseline** on the Fashion MNIST dataset. Coded in **Pytorch and Colab.**
- Wrote a **medium article that has received > 150 claps**. Click here

Efficient Resizing & Highly Imbalanced Multilabel Classification of ChestX-rays
https://github.com/ViraJBagal/ChestXRay14-Reimplementation (June 2020 - July 2020)

- Reproduced, Ekagra et al.**Jointly Learning Convolutional Representations to Compress Radiological Images and Classify Thoracic Diseases in the Compressed Domain**. ICVGIP 2018.
- Used **latent space of AutoEncoders** for compressing radiological images.
- Implemented **DenseNets and ResNets** on these latent vectors for **mulit-label classification.**
- Going one step further, tried **FMix augmentation** to improve the performance.
- Added **Grad-CAM** in the pipeline. Model not only predicts but even highlights the decisive region.
- **Click here for the written report.**

Kaggle: Prostate cANcer graDe Assessment (PANDA)
(July 2020 - August 2020)

- Task was to classify prostate whole slide images (WSIs) in **5 ISUP classes.**
- Created **Stratifiled K-Fold** splits. Trained varieties of **ResNets, DenseNets, Se-ResNexts, EfficientNets** and compared their performances. Quadratic Weighted Kappa was the metric used.
- Implemented **weighted losses and focal loss** for tackling class imbalance and harder samples.
- Tried **Macenko, Reinhard and Vahadane stain normalization** techniques to normalize the difference in staining in different WSIs. Also tried **Stain Augmentors.**
- Tried building **two-stage pipeline** of first getting **ROIs using segmentation models like UNet** and then classification. Masks for training images were provided.
- The final system was a robust **ensemble of 1-stage and 2-stage models** for predicting Gleason scores and ISUP grades for WSIs of prostate biopsy.
- The system achieved 0.928 Cohen Kappa score and secured **16th (top 2%) position on the leaderboard among 1010 participants.** I thus achieved Kaggle Silver medal.

SIIM-ISIC Melanoma Classification
(July 2020 - August 2020)

- Task was to identify melanoma (skin cancer) in lesion images.
- Entered the competition particularly for practicing **TensorFlow, TFRecords, and TPU training on Kaggle.**
- Created **Stratifiled K-Fold** splits. Implemented variations of **EfficientNets with the Imagenet and Noisy-Student** pre-trained weights.
- Implemented scaling, translation, rotation augmentations for TPU.
- Tried exotic **hair augmentation and microscopy augmentations** that randomly insert hair like black stripes and simulate a microscope image by creating a black region around the center circle respectively.

SKILLS

Python
Expert

Pytorch
Expert

Hugging Face
Expert

OpenCV
Expert

Computer Vision
Expert

NLP
Expert

Scikit Learn
Expert

Object Detection
Expert

Matplotlib
Expert

Seaborn
Expert

RDKiT
Expert

TensorFlow
Intermediate

Semantic & Instance Segmentation
Intermediate

Geometric Pytorch
Intermediate

Latex
Intermediate

Scipy
Intermediate

NLTK
Intermediate

FastAPI
Familiar

Docker
Familiar

Amazon Web Services
Familiar

C++
Familiar

TPU Training
Familiar

Multi-GPU DDP
Familiar

ACHIEVEMENTS

KVPY Scholar
Indian Institute of Science (IISc), Bangalore

KVPY program aims to identify and support talented and motivated students in research. I secured **All India Rank 69** in their written exam+interview. My research will be supported by them until 06/2021.

National Top 1% in National Graduate Physics Examination (NGPE) 2019
Indian Association of Physics Teachers (IAPT))

Selected in Mathematics Madhava Competition 2018
Homi Bhabha Centre for Science Education, TIFR

Exam included topics like Calculus, Algebra, Combinatorics. On account of selection, I attended the prestigious Madhava Camp in Indian Statistical Institute, Bangalore, India

Silver Medal (Top 2%) in PANDA Competition on Kaggle
Kaggle

The system achieved 0.928 Cohen Kappa score and secured 16th (top 2%) position on the leaderboard among 1010 participants.

2 x Kaggle Expert
Kaggle

Only 8% of total Kaggle competitors are at this or above this rank.

LANGUAGES

English
Professional Working Proficiency

Hindi
Professional Working Proficiency

Marathi
Native