Bing **Guo**

Ph.D in High Energy Physics · Data Scientist from Millennium

(+1)803-665-5452 | bingguo0625@gmail.com | bguo.me | bingguo1

Work Experience

Data Scientist — *Millennium*

Broker GPT – Model Fine Tuning *2023 Feb - Now*

- 500k+ broker emails from more than 45 global investment banks: GoldmanSachs, UBS, Societe Generale etc, clustering and tagging them into curated datasets for portfolio managers to access
- Use ***GPT2*** to preprocess email title & content to get embeddings, then use ***K-means*** to do clustering, achieved superior result which can take into production
- Use existing tagged emails, finetune with ***T5-3b***, achieved consistent result with the clustering approach

Design and maintain a distributed scheduling & monitoring system: ADAMS (Advanced Data Monitoring System) *2022 Oct - Now*

- Key product in IT department: Deployed 58k checks on corporate level data warehouse, serving as fundamental role for monitoring datasets used by quantitative and equity portfolio managers.
- Multi-master, multi-agent across multiple windows and linux servers, load balancer built to regulate how and which agent master send check request to.
- A flexible check type system covering raw file existence check, database query, AWS cloud query, heartbeat monitoring, and content related sophisticated adhoc check. Mutiple alert levels set up: to Opsgenie, email alerts to vendors/subscribers/users.

Postdoctoral Researcher — *Fermilab - U of South Carolina*

Detector segmentation with Submanifold sparse convolutional networks with LArTPC Data *2021*

- Simulate particle interactions with Geant4 and LArsoft, energy depositions are recorded in cubic voxels every 3mm, smeared.
- Final Data: 100k images for each 192px, 512px and 768px size.
- Five particle classes segmentation: Michael electrons, delta ray, showers, HIP, MIP. Used ***Sparse CNN*** and ***Dense U-ResNet*** as baseline.
- Use ADAM optimizer and softmax cross-entropy loss averaged over all the voxels
- Trained on Nvidia V100 GPUs (32Gb), spent 10hours(sparse) and 212h(dense) for 3D images of size 192px.
- Comparing to dense net, our sparse CNN allowed to use larger batch (64 to 4), and 2% higher accuracy, smoother learning curve, reduced memory footprint by factor of 495 and wall-time 119(same batch size)

Particle and Neutrino events Identification with a Context-Rich Convolutional Neural Network *2020*

- One of the pioneers applying CNN to high energy physics community, first implementation of a four-tower siamese-type architecture both for separation of independent inputs and inclusion of context information
- Use ***Monte Carlo*** simulated events as training and test data, hits clustered with a fuzzy K-means algorithm, XY and YZ matched with a Kuiper Test. 2.95 Million particles after quality cuts.
- Trained with ***Caffe***, evaluation produced in ROOT format with Caffe C++ API
- Treat XY and YZ planes topological information and energy deposition as image, achieved 83.3% efficiency and 83.5% purity.
- Incoporated in reconstructing the energy of electron neutrino interactions, energy resolution achieved 11%, 20% improvement over previous no-context CNN.

Projects

Stable Diffussion reversed (Image to Prompts)

- Data: [Generation] Image Generation with SD v2 with GPT2 generating prompts. [existing]Used diffusionDB 2M and 14M datasets and extra 900k image-prompt pairs
- Data cleansing: high-correlation removal, filter too short & long prompts
- models: ***ConvNeXt-XXLarge***, ***clip-ViT-Large***, ***Blip2***, ***Clip Interrogator***, ***KNN*** Regression(with CLIP embeddings), training: Linear Probing and Fine tuning
- Ensembled with first 3 models achieved best score: 0.591

Gomoku (*AlphaZero* Algorithm implementation) with reinforced learning

- Policy-value network built with pytorch, with policy and value networks sharing top 3 Covolutional neural networks.
- Use ***Monte Carlo Tree Search*** to do non-exhaustive exploration(P) and exploitation(Q)

Education

University of South Carolina — *712 Main St., Columbia, SC, 29208*

Ph.D. in High Energy Physics *Aug 2012 - Jul 2020*

East China Normal University — *Minhang district, Shanghai, China*

B.S. in Physics *Sep 2008 - May 2012*

Skills

Programming & Tools	C++, Python, Java, R, Javascript, SQL, Bash, LaTeX, PHP, Linux
Big Data & Cloud	Spark, Kafka, AWS Sagemaker, Amazon Athena, Hadoop, MapReduce, Hive, HBase
Machine Learning	Pytorch, TensorFlow, Keras, Sklearn, lightgbm, xgboost, CNN, RNN, Reinforced Learning, Transformers, GPT, LangChain