AI-Community Technical Lead Assignment

Deadline: ***31/03/2024 11:59PM***

Note: *You are expected to do all the problems. Doing optional problems reward you with brownie points!*

Problem 1:

Problem 1 has two parts, you have to do ***either one of them,*** *but again, note that doing the optional ones will give you brownie points.*

Improving Accuracy and Robustness of Convolutional Neural Networks Using Innovative Attention Mechanisms:

Convolutional Neural Networks (CNNs) remain one of the most successful and prevalent deep learning architectures used today for Computer Vision tasks. Despite impressive results achieved by classical CNN designs, there is always room for improvement when dealing with real-world challenges such as occlusion, noise, varying scales, and low contrast images. In order to address these limitations, we require you to propose solutions involving attention mechanisms integrated within established CNN structures.

Your task includes designing and implementing attention mechanisms, drawing inspiration from existing literature or developing original ones if desired. Clearly cite any referenced work influencing your proposed approach.

The goal of the project is threefold:

1. To design efficient attention mechanisms compatible with common CNN architectures (e.g., Simple CNN, ResNet, DenseNet, MobileNet, EfficientNet) demonstrating improved generalizability and universality across multiple models. Demonstrating on all isn't necessary, showcase your accuracies on any 3 of the above.
2. To empirically validate and quantify the benefits brought about by your proposed attention mechanisms compared to vanilla counterparts on various benchmark image classification datasets including, but not limited to, CIFAR-10, CIFAR-100, SVHN, MNIST, FASHION-MNIST, STL-1000, Imagenette, Caltech-101, and Oxford Flower Dataset. You are not forced to use these datasets, feel free to choose any 3 image datasets of your choice to showcase the performance of your attention mechanism.
3. To investigate how well each selected attention mechanism performs under adverse conditions (occlusion, noise, varying scales, and low contrast) resulting in enhanced robustness and resilience against unpredictable real-life scenarios.

Throughout development, ensure your solution remains scalable, computationally feasible, and modular enough to facilitate seamless integration with minimal modifications required to host CNN models. Furthermore, you may consider comparing and analyzing your attention-enhanced models alongside alternative prominent architectures like Transformers, Graph Convolutional Networks, or Siamese Networks on similar computer vision tasks. Ultimately, the quality of submissions shall be gauged based on the following criteria:

- Implementation of the proposed attention mechanism
- Quantitative improvements in validation and testing accuracy
- Computational efficiency, memory footprints, and convergence rates
- Quality of experimental analyses conducted on various datasets and adversarial settings
- Adaptability to multiple CNN architectures and compatibility with alternative architectures

OPTIONAL PROBLEM STATEMENT FOR PROBLEM 1:

Developing an Efficient Recommender System Using Collaborative Filtering or Content-Based Filtering Methods

In today's digital world, personalized recommendations greatly influence user experiences across various online services like e-commerce, entertainment, finance, healthcare, and education. However, many existing recommender systems face issues such as high computational requirements, difficulties with new user or item data, and lack of enough data for accurate predictions, which can lead to user dissatisfaction. This project aims to address these challenges by creating a lightweight yet powerful recommendation system based on proven collaborative filtering or content-based filtering principles.

Goal of the Project:

The main goal of this project is to develop a recommender system that is easy to deploy, computationally efficient, and provides accurate and diverse recommendations. You guys can choose either collaborative filtering (user-user/item-item CF) or content-based filtering techniques, and they are encouraged to explore hybrid, ensemble, transfer learning, meta-learning, or cascading approaches. You are free to implement any research paper/article given that cite the resources.

Specific milestones for the project include:

1. Gathering, cleaning, and combining diverse data sources relevant to the project.
2. Using techniques like Singular Value Decomposition (SVD), Latent Dirichlet Allocation (LDA), Non-negative Matrix Tri-Factorization (NTF), or Autoencoders to handle data scarcity and extract meaningful information.
3. Developing effective methods for forming neighborhoods, normalizing ratings, handling negative samples, calculating similarity, and establishing recommendation reliability.
4. Investigating different feedback types, addressing biases, and ensuring fairness in recommendations.

5. Balancing exploration and exploitation to adapt to user preferences and changing environments.

Proposals will be evaluated based on:

- Clarity of the proposed solution's concepts, mathematics, and algorithms.
- Empirical validation using synthetic or real-world data, showing comparable accuracies to baselines and current SOTA algorithms
- Comparisons highlighting the trade-offs between accuracy, coverage, timeliness, and maintainability.

<u>Problem 2:</u>

InstiGPT has been widely adopted across various platforms to provide human-like responses to user queries. However, one significant challenge we face is the occurrence of hallucinations, where the model generates responses that are factually incorrect or irrelevant. This issue poses a threat to the reliability and trustworthiness of InstiGPT, especially in critical applications such as educational assistance.

candidates are tasked with devising a **comprehensive problem-solving strategy** to address the issue of hallucinations in InstiGPT. This strategy should involve:

Challenges with Large Datasets in RAG:
The usage of large datasets, particularly in Retrieval-Augmented Generation (RAG) models like InstiGPT, can lead to computational inefficiencies and increased risk of hallucinations. candidates should anticipate and address potential problems arising from the scalability of datasets, such as data sparsity, overfitting, or model drift.

Continuous Data Updating and Efficient Embedding Generation:
To ensure InstiGPT remains up-to-date with the latest information, candidates must propose strategies for continuous data updating. Given that generating embeddings for large datasets can be time-consuming, candidates should explore alternative approaches to minimize computational overhead while maintaining embedding quality. This could involve incremental updating techniques, pre-trained embeddings, or caching mechanisms.

Alternative to Google Search and Performance Optimization:
Considering the limitations of traditional search engines like Google in providing contextually relevant responses, candidates are encouraged to propose alternative approaches for information retrieval. This could involve leveraging specialized knowledge graphs, domain-specific corpora, or curated datasets. Additionally, candidates should devise strategies to optimize the search process for speed and efficiency, such as indexing, caching, or parallelization.

Fine-Tuning and Deployment:

If fine-tuning is deemed necessary to improve InstiGPT's performance, candidates should outline a fine-tuning pipeline, including model selection, dataset creation, and deployment strategies. They should estimate the computational resources required for fine-tuning, consider cost implications, and propose mechanisms for monitoring and updating the deployed model in production environments.

Cost Management and Resource Allocation:
candidates need to consider the financial implications of their proposed solutions, including the cost of compute resources, data storage, and model deployment. They should provide recommendations for cost-effective resource allocation, scalability, and budget optimization to ensure the sustainability of the proposed solution.

Compute and Fine-Tuning Requirements:
Estimating the computational requirements for fine-tuning and deployment is crucial. candidates should provide insights into the computational resources (e.g., GPU hours, memory, storage) and time required for fine-tuning different models, along with strategies for optimizing resource utilization and minimizing training time.

Overall, candidates are expected to present a holistic strategy that addresses the technical, computational, and cost-related challenges associated with mitigating hallucinations in InstiGPT and enhancing its performance as a reliable conversational AI system.

Problem 3:

This is a non-technical problem. Write a complete schedule of your tenure starting off from April and ending in March. You should write a detailed plan of implementation of various hackathons and competitions that you aim to participate in or the projects that you plan to do along with InstiGPT. Write an implementation plan of work distribution of various engineers given that you have 10 junior engineers.

Submission Instruction:

Submission links will be sent out a day prior to the deadline. You are advised to submit a single GitHub repo link of all three questions. You are required to submit a report (in the same GitHub repo), explaining everything about your approach, the reasons of selecting this approach and how can this be relevant in the context of this problem statement, your results of experiments, and other cited relevant work that you used.