Ranking Attention Heads by Operational Contribution Using Few-Shot In-Context Learning

Objective:

You are required to identify and rank the attention heads in the **last layer** of a language model based on their **operational importance**. The ranking will be determined by measuring the impact each head has on model performance when subjected to controlled perturbations. The evaluation should follow a **Few-shot In-Context Learning (ICL)** setup on a given dataset.

Detailed Steps:

Step 1: Dataset Access

- Obtain the specified dataset that will be used to evaluate the language model in a few-shot ICL setting. Ensure the dataset is formatted properly for few-shot prompts.

Step 2: Load a Small Language Model

- Choose and load a lightweight transformer-based language model (e.g., **GPT-Neo**) that supports access to individual attention head outputs.

Step 3: Baseline Evaluation

- Run inference on the dataset using the unmodified model in a few-shot ICL setup.
- Record the model's **baseline performance metric** (e.g., accuracy, loss, F1 score). This will serve as the reference and be referred to as result_original.

Step 4: Perturbation Analysis of Last Layer Attention Heads

- Iterate through **each attention head** in the **final layer** of the model.
  For a given attention head i, perform the following:
  Step 4.1: Inject a small, controlled amount of **random noise** into the output of attention head i.
  Step 4.2: Forward propagate the noisy attention head output through the rest of the model.
  Step 4.3: Evaluate the perturbed model on the same dataset under identical conditions.
  Step 4.4: Record the new performance metric, referred to as result_atth_i.

Step 5: Calculate Operational Contribution

- For each attention head i, compute the **deviation** between result_atth_i and the result_original.
  - The greater the performance degradation, the more **operationally important** that attention head is considered to be.

Step 6: Rank Attention Heads

- Sort the attention heads in **descending order of their performance impact**, i.e., heads causing the largest drop in performance when perturbed should rank highest.
- Print or return the sorted list along with their contribution scores.

Expected Output

- A ranked list of attention heads in the final layer along with their respective operational impact scores.