Running an Inference Service for M*DEL Experts

Multi-Domain Expert Learning (M*DEL) is an approach to training LLMs for expertise in knowledge domains.

This process involves branching from a base model and training the branch on specific domain data in order to establish an expert, which routing logic is then able to activate in serving inference requests. The result is a framework for domain expertise that is easily-extensible, modular, and efficient.

Contents

1 Prerequisites
2 Launch the TGI Container Instance
3 Run Inference Commands

Prerequisites

Setting up the inference system will use M*DEL's Aurora model (aurora-m) and will rely on RunPod compute resources, similar to the model training process. If the prerequisites for Training an M*DEL Expert have already been completed, then continue to the next section.

Otherwise, follow the instructions in that document for setting up HuggingFace and RunPod accounts before continuing.

Launch the TGI Container Instance

A template exists at RunPod to launch a text generation inference (TGI) container for the Aurora model.

To initialize the service, log in at RunPod and perform the following steps:

1. Start from the text-generation-inference template and set the organization Account, as appropriate, from the top-right profile drop-down list.
2. Click the Filters icon and change "Allowed CUDA Versions" to 12.2 only.
3. Scroll down to the Previous Generation section and click Deploy on the 1xA100 80GB GPU. Click "Customize Deployment":
    - Update "Container Start Command" to point to the expert uploaded via the huggingface-cli during training; e.g.: `--model-id stillerman/aurora-mathematica`
    - Expand "Environment Variables" and set `HUGGING_FACE_HUB_TOKEN` to your read token from huggingface.co/settings/tokens
    - Click "Set Overrides" and then Continue Deploy
4. After the instance has started up, wait for the following line to appear in its Container Logs (this might take approximately 5 minutes):

    ```
    WARN text_generation_router: router/src/main.rs:327: Invalid hostname, defaulting to 0.0.0.0
    ```

    Close the log and click Connect Connect to HTTP Service [Port 80]; this will open a page in a web browser. Copy the URL to use for inference requests.

Run Inference Commands

HTTP requests can now be made to the running TGI system using an HTTP client or library.

For example, cURL can be used to make an inference request starting with the text "Name:" by running the following:

```
$ curl https://<YOUR BASE URL>/generate \
    -X POST \
    -d '{"inputs":"Name:","parameters":{"max_new_tokens":1024, "do_sample":true, "repetition_penalty":
1.1, "top_p":0.95, "top_k":40, "temperature": 0.9, "stop":["###"], "return_full_text": true}}' \
    -H 'Content-Type: application/json'
```

Replace `<YOUR BASE URL>` with the URL of the HTTP Service that was determined after launching the TGI container instance. This command will connect to the instance and submit an inference request with the given inputs and parameters.

The parameters of inference have a significant effect on the quality of output, so should be adjusted to discover the most appropriate values for a given use case.