Gen-AI Semester-6

LAB-1

Date:23/01/2026

Name : Anushka Mandal	SRN : PES2UG23CS083	SECTION : B

Documenting the Gen-AI Handson

Hugging Face can be viewed as the GitHub of Artificial Intelligence, hosting a vast collection of open-source models, datasets, and AI projects. These resources showcase how modern AI systems function and enable developers to leverage pre-trained models rather than building solutions entirely from the ground up.

The Transformers library serves as the connection between Hugging Face and our code. It offers easy-to-use APIs for downloading and applying pre-trained models in custom projects, while supporting multiple deep-learning frameworks.

One of the most useful features of this library is the `pipeline()` function. It abstracts the entire workflow—preprocessing input data, running model inference, and postprocessing outputs—into a single callable interface, allowing complex NLP tasks to be performed with minimal code.

During the hands-on session, we compared two generative language models: distilgpt2 and gpt2. The distilgpt2 model is a compressed and faster version of gpt2, designed to use less memory and deliver quicker responses. In contrast, the full gpt2 model is larger and generally produces higher-quality and more fluent text.

To ensure reproducibility, a random seed was set at the beginning of the experiment. This guarantees that identical inputs generate the same outputs across multiple runs, which is essential for consistent evaluation.

When both models were given the same prompt, the gpt2 model generated text that was more coherent and grammatically accurate. Although distilgpt2 performed reasonably well, it tended to repeat itself and lose contextual understanding as the generated text became longer.

Tokenization plays a crucial role in preprocessing. Since models cannot directly interpret raw text, sentences are divided into smaller units called tokens, which are then mapped to numerical IDs that the model can process efficiently.

The temperature parameter influences the randomness of text generation. Higher temperature values encourage more diverse and creative outputs by allowing less likely word choices, while lower values result in more predictable and focused text.

Part-of-Speech (POS) tagging is another important preprocessing technique, where grammatical categories are assigned to words. For example, distinguishing whether the word “will” functions as a noun or a modal verb helps the model better understand sentence structure.

Named Entity Recognition (NER) is used to detect and classify entities such as people, organizations, and locations. For instance, NER helps determine whether the word “Apple” refers to a fruit or a technology company by analyzing the surrounding context.

Documenting the Benchmark Assignment

The goal of the benchmark assignment was to analyze how different transformer architectures behave when they are applied to tasks they are not equally optimized for. This exercise helps illustrate why model architecture plays a critical role in Generative AI.

Three transformer models were selected for the benchmark:

- **BERT (bert-base-uncased)** – an encoder-only architecture
- **RoBERTa (roberta-base)** – an improved and optimized encoder-only architecture
- **BART (facebook/bart-base)** – an encoder–decoder architecture

All three models were evaluated on the same set of tasks using Hugging Face pipeline interfaces to ensure a fair comparison.

Tasks Conducted

Text Generation
The models were prompted to generate text. BERT and RoBERTa were unable to produce meaningful output because encoder-only models are not designed for next-token prediction. BART, which includes a decoder, was capable of generating text, although the quality was relatively limited.

Fill-Mask (Masked Language Modeling)
In this task, sentences with masked tokens were provided. BERT and RoBERTa performed effectively, accurately predicting appropriate words due to their training objective centered on masked language modeling. BART, however, showed weaker performance since this task is not its primary focus.

Question Answering
The models were asked to answer questions based on a given context. The results were inconsistent, as the base versions of these models were not fine-tuned specifically for question answering. Nevertheless, they occasionally produced partial or approximate answers.

Key Observations

The benchmark clearly showed that:

- Encoder-only models like BERT and RoBERTa excel at understanding-oriented tasks such as fill-mask.
- Encoder–decoder models like BART are more suitable for generation-based tasks.
- Applying a model to tasks outside its intended design leads to suboptimal or failed outputs, which is an expected and meaningful outcome.

Conclusion

This benchmark assignment emphasizes the significance of transformer architecture in Generative AI. It demonstrates that a model's effectiveness depends on how closely the task aligns with its architectural design, reinforcing the foundational concepts introduced in Unit 1.