**Benchmarking LLMs on Summarization Task using Rouge and Bleu**

**Introduction**

Transformers pipelines from Hugging Face can be used for a wide range of tasks in natural language processing, computer vision, and audio processing. In NLP, tasks include text classification, named entity recognition, question answering, text generation, translation, etc. For computer vision, pipelines support image classification, object detection, depth estimation, and image segmentation. Audio processing tasks encompass automatic speech recognition and audio classification.

In the ‘summarization’ notebook, the "google-t5/t5-small" model from huggingface is evaluated on summarization task using Rouge and Bleu score.

**Summarization**

**Definition**

The summarization task in transformer pipelines involves creating a shorter version of a document or article while retaining its essential information. Summarization is formulated as a sequence-to-sequence task, where the model processes the input text and produces a concise summary as output.

**BillSum dataset (**[**link**](https://paperswithcode.com/dataset/billsum)**)**

The BillSum dataset is the first dataset for summarization of US Congressional and California state bills. It contains a collection of legislative texts along with their corresponding summaries, making it a valuable resource for training and evaluating summarization models. The dataset features fields such as the full text of the bill, its summary, and the title.

The BillSum dataset consists of three parts: US training bills, US test bills and California test bills. Each part has three main features:

1. Text: This field contains the full text of the bill, which serves as the input for summarization models.

2. Summary: This field provides a human-written summary of the bill, representing the target output that models aim to generate.

3. Title: Each bill has an associated title, which can be useful for context or generating summaries.

**ROUGE metric (**[**blog**](https://medium.com/nlplanet/two-minutes-nlp-learn-the-rouge-metric-by-examples-f179cc285499)**)**

ROUGE (Recall-Oriented Understudy for Gisting Evaluation), is a set of metrics and a software package specifically designed for evaluating automatic summarization, but that can be also used for machine translation. The metrics compare an automatically produced summary or translation against reference (high-quality and human-produced) summaries or translations.

**BLEU score (**[**blog**](https://medium.com/nlplanet/two-minutes-nlp-learn-the-bleu-metric-by-examples-df015ca73a86)**)**

BLEU, or the Bilingual Evaluation Understudy, is a metric for comparing a candidate translation to one or more reference translations. Although developed for translation, it can be used to evaluate text generated for different natural language processing tasks, such as paraphrasing and text summarization.

**Evaluate model on BillSum dataset using ROUGE metric**

Using transformers pipeline and datasets library, we evaluate the performance of pre-trained "google-t5/t5-small" model from huggingface on the first 10 examples of ‘ca\_test’ split of billsum dataset using ROUGE metric.

**Evaluate model on BillSum dataset using BLEU metric**

Using transformers pipeline and datasets library, we evaluate the performance of pre-trained "google-t5/t5-small" model from huggingface on the first 10 examples of ‘ca\_test’ split of billsum dataset using BLEU score.

**Summarization task from** [**huggingface blog**](https://huggingface.co/docs/transformers/en/tasks/summarization)

Following the blog, we load a pre-trained "google-t5/t5-small" model from huggingface and further train it on the billsum dataset for 4 epochs using Seq2SeqTrainer. The default ‘AdamW’ optimizer and cross-entropy loss is used. Evaluating the model after training for 4 epochs shows an improvement in the ROUGE score.