Deep Learning Part 2

Section 2

## **Fine-Tuning a Language Model for Extractive Summarization**

**Understanding the Task:**

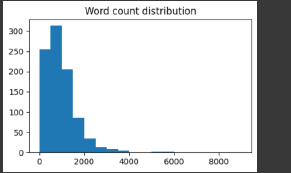
Extractive summarization involves selecting the most important sentences or phrases from a given text to create a concise summary. In this case, we're fine-tuning a model to generate suitable headings for paragraphs, which can be considered a form of extractive summarization.

**Choosing the Right Model:**

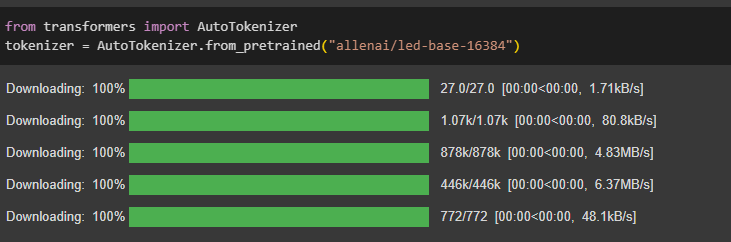
The Longformer Encoder-Decoder (LED) is a suitable choice for this task due to its ability to handle long sequences effectively. This is important because paragraphs can vary significantly in length.

**Steps Involved:**

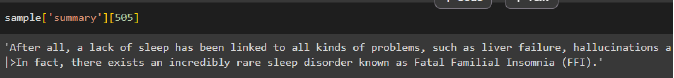
1. **Load and Preprocess Data:**
   * Load the CSV data using Pandas.
   * Clean the data by handling missing values, removing duplicates, and addressing outliers.
   * Calculate the word count for each paragraph.
   * Filter out outliers to ensure a more balanced dataset.
   * **Note:** Log transformation might not be necessary if the data distribution is positively skewed, as outliers can be removed directly.



1. **Tokenization:**
   * Initialize the tokenizer specific to the LED model.
   * Tokenize the text data to convert it into numerical representations.
   * Pad or truncate sequences to ensure consistent input lengths.
   * Create a global attention mask to guide the model's focus.

****

1. **Training Model:**
   * Load a pre-trained LED model.
   * Fine-tune the model on your specific dataset to adapt it to the heading generation task.
   * Experiment with different hyperparameters like batch size, learning rate, and number of epochs to optimize performance.



**Data Splitting:**

While the prompt didn't explicitly mention data splitting, it's a crucial step in machine learning. Typically, the dataset is divided into training, validation, and testing sets to evaluate the model's performance during training and prevent overfitting.

**Key Points:**

* **Data quality:** Ensure clean and relevant data for effective model training.
* **Model choice:** LED is well-suited for handling long sequences.
* **Hyperparameter tuning:** Experiment with different parameters to optimize performance.
* **Data splitting:** Divide the dataset into training, validation, and testing sets for proper evaluation.

By following these steps and carefully considering the specific requirements of your task, you can effectively fine-tune a language model for extractive summarization, such as generating suitable headings for paragraphs.

