

GenAI HandsOn-1

Name: Avani Ajith

SRN: PES2UG23CS110

Section: B

1. Model Comparisons: Standard vs. Distilled

We explored the differences between standard Large Language Models (LLMs) and their distilled versions.

- **Main Model (e.g., GPT-2):** These are larger and typically more coherent. They are trained to predict the next word in a sequence based on probability (using a Softmax function).
- **Distilled Model (e.g., DistilGPT-2):** A smaller, faster version optimized for speed and lower memory usage. However, this distillation often leads to a decrease in accuracy and coherence compared to the main model.

2. The Role of Seed Values

A seed value is used to ensure that random results are reproducible. By setting a seed, the random number generator starts from the same point, producing the same sequence of values every time the code is run.

Experimenting with Seed Values:

- **Seed 42:**

For distilgpt:

```
... /usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:  
  The secret 'HF_TOKEN' does not exist in your Colab secrets.  
 To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as secret in your Google Colab and restart your session.  
 You will be able to reuse this secret in all of your notebooks.  
 Please note that authentication is recommended but still optional to access public models or datasets.  
 warnings.warn(  
 config.json: 100% [██████████] 762/762 [00:00<00:00, 36.2kB/s]  
 model.safetensors: 100% [██████████] 353M/353M [00:02<00:00, 218MB/s]  
 generation_config.json: 100% [██████████] 124/124 [00:00<00:00, 13.3kB/s]  
 tokenizer_config.json: 100% [██████████] 26.0/26.0 [00:00<00:00, 2.31kB/s]  
 vocab.json: 100% [██████████] 1.04M/1.04M [00:00<00:00, 15.1MB/s]  
 merges.txt: 100% [██████████] 456K/456K [00:00<00:00, 20.5MB/s]  
 tokenizer.json: 100% [██████████] 1.36M/1.36M [00:00<00:00, 27.6MB/s]  
 Device set to use cpu  
 Truncation was not explicitly activated but 'max_length' is provided a specific value, please use 'truncation=True' to explicitly truncate examples to max length. Defaulting to 'longest_first' truncation strategy. If you enc  
 Setting 'pad_token_id' to 'eos_token_id':50256 for open-end generation.  
 Both 'max_new_tokens' (=256) and 'max_length' (=50) seem to have been set. 'max_new_tokens' will take precedence. Please refer to the documentation for more information. (https://huggingface.co/docs/transformers/main/en/main)  
 Generative AI is a revolutionary technology that is designed to work with existing AI systems. It has been developed by the University of California, Berkeley. Its research team is the leading developer of AI software and it
```

The research team led by Professor Daniel Kranz, from the University of California, Berkeley, has developed a program to learn how to use the AI to improve the performance of the software. It has been developed by the University of California, Berkeley, and is a top-selling research computer. The research team developed the program to learn how to use the AI to improve the performance of the software. It has been developed by the University of California, Berkeley, Berkeley, and is a top-selling research computer. The research team developed the program to learn how to use the AI to improve the performance of the software. It has been developed by

For GPT2:

The AI is a model of human intelligence, and has many aspects that are similar to artificial intelligence. It can learn from humans, and it can adapt to the environment. It can learn by experimenting with new ways of thinking. It is the main driving force behind the new Artificial Intelligence, and the AI is very important to the success of AI. The new AI is designed to work out problems that need to be solved in a way that is easy to understand. The AI is designed to be scalable and adaptable to different environments. It can be used to solve complex problems without relying on humans. It can be used to build a solution that is very quickly scalable, scalable, inexpensive. The new AI is designed to work out problems that need to be solved in a way that is

- **Seed 123:**

Step 3: Fast Model (`distilgpt2`)

Let's see how the smaller model performs.

```
# Initialize the pipeline with the specific model
fast_generator = pipeline('text-generation', model='distilgpt2')

# Generate text
output_fast = fast_generator(prompt, max_length=50, num_return_sequences=1)
print(output_fast[0]['generated_text'])
```

Device set to use cpu
Truncation was not explicitly activated but 'max_length' is provided a specific value, please use 'truncation=True' to explicitly truncate examples to max length. Defaulting to 'longest_first' truncation strategy. If you enc
Setting 'pad_token_id' to 'eos_token_id':50256 for open-end generation.
Both 'max_new_tokens' (=256) and 'max_length' (=50) seem to have been set. 'max_new_tokens' will take precedence. Please refer to the documentation for more information. (<https://huggingface.co/docs/transformers/main/en/main>)
Generative AI is a revolutionary technology that will make artificial intelligence a reality, but only for a few years.

So what do you think? Let us know in the comments below.

For GPT2:

Device set to use cpu
Truncation was not explicitly activated but 'max_length' is provided a specific value, please use 'truncation=True' to explicitly truncate examples to max length. Defaulting to 'longest_first' truncation strategy. If you enc
Setting 'pad_token_id' to 'eos_token_id':50256 for open-end generation.
Both 'max_new_tokens' (=256) and 'max_length' (=50) seem to have been set. 'max_new_tokens' will take precedence. Please refer to the documentation for more information. (<https://huggingface.co/docs/transformers/main/en/main>)
Generative AI is a revolutionary technology that will revolutionize the human brain. The concept of AI is based on the notion that machines can do all sorts of things. They can predict events, predict how they will affect th

In the next couple of years, we will see a lot of new innovations in AI technology. But at the same time, we need to keep the revolution going. In this sense, the revolution is happening now.

The Future of Information Technology

If you are looking at the future of knowledge technology, it is very exciting to see that there are a lot of people who are looking at it from the perspective of a future in which the technology of the future will be increas
One of the challenges we face today is that we are constantly changing the way we think about information. Information is not just a technology, it is a set of beliefs, a kind of beliefs. The first step in the evolution of i

- **Seed 98**

For DistilGPT:

```
... Device set to use cpu
Truncation was not explicitly activated but 'max_length' is provided a specific value, please use 'truncation=True' to explicitly truncate examples to max length. Defaulting to 'longest_first' truncation strategy. If you enc
Setting 'pad_token_id' to 'eos_token_id':50256 for open-end generation.
Both 'max_new_tokens' (=256) and 'max_length' (=50) seem to have been set. 'max_new_tokens' will take precedence. Please refer to the documentation for more information. (https://huggingface.co/docs/transformers/main/en/main)
Generative AI is a revolutionary technology that advances in the field of robotics, and it's in the making-s progress. It's a new way of learning, and it's a new way of learning.
```

For GPT2:

```
Device set to use cpu
Truncation was not explicitly activated but 'max_length' is provided a specific value, please use 'truncation=True' to explicitly truncate examples to max length. Defaulting to 'longest_first' truncation strategy. If you enc
Setting 'pad_token_id' to 'eos_token_id':50256 for open-end generation.
Both 'max_new_tokens' (=256) and 'max_length' (=50) seem to have been set. 'max_new_tokens' will take precedence. Please refer to the documentation for more information. (https://huggingface.co/docs/transformers/main/en/main)
Generative AI is a revolutionary technology that is disrupting the way we manage and collect data. It's creating a new era in machine learning, the use of deep learning to improve training and performance, and the democratiz
Machine Learning is a new paradigm for machine learning. From the beginning of machine learning, some researchers have been using the concept of a neural network to learn and use human-like behaviors to improve learning. In
Machine Learning is an exciting and exciting time to be in the industry. Machine learning is a new paradigm for machine learning. From the beginning of machine learning, some researchers have been using the concept of a neur
Machine Learning is a new paradigm for machine learning. From the beginning of machine learning, some researchers have been using the concept of a neural network to learn and use human-like behaviors to improve learning. In
```

3. NLP Fundamentals: Under the Hood

Before a model can generate text, several processing steps occur:

- **Tokenization:** Models cannot read text directly. Tokenization breaks text into "Tokens" and assigns each a unique numerical ID.
- **POS Tagging:** This identifies the grammatical parts of speech (e.g., NNS for plural nouns, VBD for past tense verbs) to help the model understand sentence structure.
- **Named Entity Recognition (NER):** Used to extract structured data like names, organizations (ORG), or miscellaneous (MISC) entities from raw text.

4. Key Model Architectures

We differentiated between the "reading" and "writing" components of Transformers:

- **Encoders:** Used to "read" and understand the context of the input. Models like BERT are bidirectional, meaning they read text from both left-to-right and right-to-left to fully grasp context (e.g., distinguishing if "Apple" is a fruit or a company).
- **Decoders:** Used to "write" or generate text. GPT-2 is a **decoder-only** model, trained specifically to predict the next word based on preceding ones.
- **Combined Architectures:** Tasks like Question Answering or filling in the blanks (Masked Language Modeling) often utilize both encoders and decoders.

5. Advanced Decoding Concepts

To control the variety and quality of generated text, we use specific parameters:

- **Temperature:** Controls randomness. A higher temperature makes the output more diverse/random, while a lower temperature makes it more predictable.

- **Top-K:** The model selects the top k most likely next words.
- **Top-P:** The model selects the smallest set of words whose cumulative probability is greater than or equal to p .

A major focus for the upcoming year in LLM development includes:

- **Latency:** Reducing the time it takes for a model to generate a response.
- **Context Engineering:** Improving how models handle and remember long-range dependencies within a large text context.

6. Linguistics: Parts of Speech (POS)

We used NLTK to tag the parts of speech in the sentence generated by GPT-2 ("*Generative AI is a revolutionary technology...*") to prove the model understands grammar.

- Generative (JJ - Adjective)
- AI (NNP - Proper Noun, Singular)
- is (VBZ - Verb, 3rd person singular)
- a (DT - Determiner)
- revolutionary (JJ - Adjective)
- technology (NN - Noun, Singular)
- that (WDT - WH-determiner)
- has (VBZ - Verb)
- the (DT - Determiner)
- potential (NN - Noun)