**Assignment 3 – News headline generation using a causal transformer-based LM**

**Objective:**
The objective of the assignment is to fine-tune an transformer model(DistilGPT-2) to generate news headline using a NewsRoom dataset. The predicts the headline based on the given article and summary. Here I have used multiple decoding strategies are compared (Greedy, Top-K,Top-p, and temperature) to analyze their quantitative and qualitative effects on the quality of generation.

**About the Dataset:**
The dataset used is a NewsRoom dataset which is combined of 10,000 training samples and 1000 testing samples. Each record contains two text fields summary and title. Here each example was formatted with a custom instruction prompt Article: [summary], headline:[target headline] and within 1024-token limit of the DistilGPT-2, each summary was cutoff to nearly 500 tokens.

**Data Preprocessing and Masking:**
I have converted each input sample into a structured instruction format to help the model understand the task "Article: [summary text]\nHeadline: [headline text]".

After that, I have tokenized the text using the DistilGPT-2 Byte-pair Encoding tokenizer, which converts words into subword tokens within the model token limit (1024). Then to avoid exceeding the model's input limit, each article summary was truncated to 500 tokens. And leaving sufficient space for headline generation.

"DataCollatorForLanguageModeling(mlm=False)" This will dynamically pads each batch size to length of the longest sequence. The it is ensured that each token only attends to previous tokens i.e., from left to right, preventing information leakage for future words.

The prompt tokens are the masked out from loss computation, which then allows only the headline tokens to contribute to training loss. Although padding masking is handled automatically prompt tokens can optionally excluded from loss computation by explicitly setting their labels -100.

**Model Description – DistilGPT-2:**
**Model Overview:** DistilGPT-2 is a compressed version of GPT-2, which is created by using knowledge distillation. It achieves 97% of GPT-2's performance while being 40% smaller and 60% faster that GPT-2 which makes it ideal for finetuning.

**Model Architecture:** It is a decoder-only transformer model which is designed for autoregressive language modelling in which the model predicts the next token based on the previous tokens.

**Number of transformer layers:** It contains 6 stacked transformer layers in which each are responsible for progressively refining the hidden representations of input tokens.
**Attention mechanism:** In this model each layer consists of 12 self-attention heads, which enabling the model to focus on multiple contextual relationships simultaneously within the text.

**Model Dimensionality:** The each token embedding of the model has a hidden size of 768, which is representing the internal vector dimension that encodes the semantic and syntactic information.

**Feed-Forward Layer:** Each transformer block of the model has a two layer feed-forward network with a dimension of 768→ 3072→ 768, which then allows nonlinear transformations and improving representation depth.

**Fine-Tuning Process:**
The model was initialized using the Hugging face transformer library with 'AutoModelForCasulaLM'. Fine tuning was performed using the TrainerAPI, with a learning rate of 8e-6 (used very low learning rate to prevent overfitting), weight decay of 0.15, and batch size as 2 for training and 4 for evaluation. The process is ran for 12 epochs with early stopping with patience as 5 to prevent overfitting.

The Casual Language model (CLM) objective was used (mlm=false), meaning the model learned to predict the next token given all previous token. The **DataCollatorForLanguageModeling** handled dynamic padding and it also ensured only the headline tokens contributed to loss.

The embedding matrix was frozen to preserve pre-trained lexical knowledge, while all output layers of the transformer were trainable. Which allows the model to specialize in the headline generation task with maintaining linguistic stability. The model was evaluated with four decoding strategies Greedy, Top-K (K=10), Top-P (p=0.9), and Temperature (T=1.2) to analyze the coherence in generation.

The loss of training and evaluation are consistent over epochs, with lowest validation loss at 3.4. Top-p and Temperature sampling scored the highest semantic similarity which is showing that they are contextual fluent whereas greedy decoding produced more repetitive results with lexical overlap.

**Training Progress: I Ran 12 epochs with early stopping.**

| Epoch | Training Loss | Evaluation Loss |
|---|---|---|
| 1 | 3.807000 | 3.642519 |
| 2 | 3.677900 | 3.545955 |
| 3 | 3.671300 | 3.516960 |

The loss is decreasing steadily which shows that stable convergence and minimal overfitting.

**Example Generated Headlines:**

```
 Example 1
Summary: NASA has announced the Artemis mission that aims to return astronauts to the Moon by 2025.
Reference Headline: NASA Plans Return to Moon by 2025 Under Artemis Program
Generated (Greedy): Artemis mission to return to Earth by 2025: NASA says it will
Generated (Top-K): 'Antarctica' mission to the moon is on schedule
Generated (Top-P): Artemis mission to return to Earth by 2025 : People.com
Generated (Temperature): Apollo mission is coming off a 3-year drought : People.

 Example 2
Summary: Germany’s soccer federation is facing scrutiny following reports of racist chants during the World Cup qualifiers.
Reference Headline: Racism Allegations Cloud Germany’s World Cup Campaign
Generated (Greedy): Germany‘s soccer federations face scrutiny after racist chants at
Generated (Top-K): Germany’scores’ clash with soccer federation amid growing
Generated (Top-P): Germany's soccer federation faces scrutiny following racist chant during World Cup qualifier
Generated (Temperature): Germany is up in arms over accusations against Poland during World Cup in

 Example 3
Summary: Apple unveiled its latest iPhone model featuring improved battery life and advanced AI-powered camera software.
Reference Headline: Apple Reveals New iPhone With Smarter Camera Features
Generated (Greedy): Apple unveils new iPhone 5s with improved battery-life
Generated (Top-K): iPhone 7's battery life improvement is impressive, but it
Generated (Top-P): Apple's new iPhone 7 Plus Plus Camera app is a good start
Generated (Temperature): iPhone 5, 4S, 5S...
```

**Observations:**

**Greedy decoding** often tends to produce more deterministic and repetitive outputs that often represents the structure of the reference but may cut off early. **Top-k sampling** is showing moderate diversity but sometimes it yield incomplete or incoherent phrases. **Top-p sampling** is showing a balanced results which are fluent, semantically coherent, and contextually close to the true headline. **Temperature Sampling** which is generating more creative but once in a while it gives less factual results, illustrating the trade-off between diversity and precision. Overall, top-p gave me the most semantically faithful and stylistically natural results which are giving better BertScore and Rouge scores.

**Evaluation Metrics:**

**ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) F1:**

The metric ROUGE-L which measures the longest common subsequence(LCS) between the generated and reference headline. It mainly focuses on lexical overlap which is how many words or phrases from the reference appear in the generated headline which are maintaining order. High rouge score indicated better word-level similarity and structural alignment. However it is not able to catch the sentence which are similar which makes that to limit its use for more flexible or creative generations.

**BERTScore-F1:**

BERTScore uses contextual embeddings from pre-trained bert models for evaluating the semantic similarity with the original and generated texts. It directly computes the cosine similarity between embedding tokens instead of matching words directly. Thereby it captures paraphrasing sentences. Usually F1 score balances precision and recall which represents the overall semantic faithfulness of the generated headline. High BERTScore indicates that is generating the headline with same meaning but with different words.

ROUGE-L measure the top level overlap whereas BERTScore evaluates the semantic similarity. By using these both evaluations metrics we can get a balanced understanding of the generated headline quality.

| Decoding Strategy | ROUGE-L F1 | BERTScore F1 |
|---|---|---|
| **Greedy** | 0.1384 | 0.8421 |
| **Top-K (k=10)** | 0.0678 | 0.8274 |
| **Top-P (p=0.9)** | 0.1214 | 0.8368 |
| **Temperature (T=1.2)** | 0.1161 | 0.8359 |

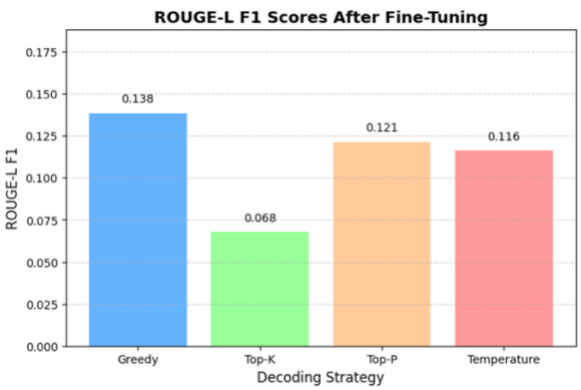
**Observations and Conclusions:**
Among all the decoding strategies greedy decoding achieved the highest overall scores for both the evaluation metrics i.e., ROUGE-L (0.1384) and BERTScore (0.8421) which indicates that deterministic tokens selection giving us more factually precise and semantically faithful headlines.

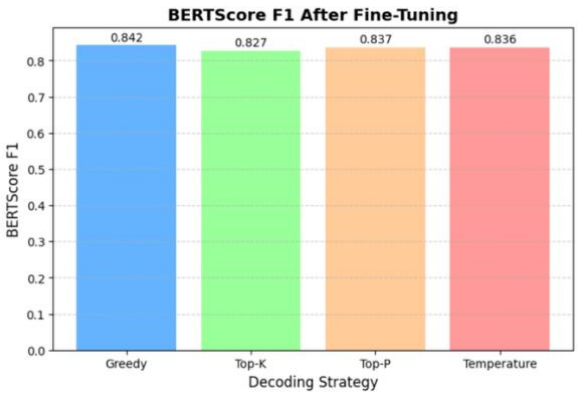
While ROUGE-L values are showing a noticeable variation across methods, BERTScore remains relatively stable around 0.83 - 0.84. This shows us even word-level overlap changes, the semantic meaning of the generated headlines stays largely preserved.

Whereas sampling based methods (Top-K, Top-P, and Temperature) gives us more creative and diverse outputs but at the cost of lower lexical accuracy. The lower ROUGE-L tells us that randomization can introduce paraphrasing or incomplete phrases that deviate slightly from the reference.

Among the three Top-P has produced a balanced performance which tells us it perfectly balances the trade-off between controlled randomness and semantic consistency.

Overall, Greedy decoding performed quantitatively best, but Top-P decoding produced qualitatively superior outputs.




**My Opinion on Best Strategy:**

Greedy achieved the highest ROUGE-L and BERTScore which shows a strong accuracy but limited diversity. Top- K and Temperature sampling introduced better randomness but often reducing coherence and factual alignment. Top-p has provided the best balance among creativity and precision with a good rouge and bertscore. Its headlines were more natural, fluent, semantically rich and realistic while compared to other strategies. Overall Top – P sampling came out as my most effective strategy for generating high quality, human like headlines.