Project Report of DL Implementation

S.No	Name	Roll Number
1	Anurag Saraswat	M20CS066
2	Parsa Revanth	M20CS058

COLAB Link:
https://colab.research.google.com/drive/1aQv5I4JshLz39gtV8_vQGTyAeMUxMExb?usp=sharing

LIBRARIES USED

- Transformers module and datasets module from the Hugging face transformers
- Sentencepiece tokenization module: This is used as the authors of the papers found this tokenization technique with vocabulary size of 96K works better
- From the transformers module imported the PegasusForConditionalGeneration, PegasusTokenizer, Trainer, TrainingArguments
- From datasets imported load_metric, which is used to load the Rouge metric

DATASET

- The dataset is the news-summary dataset created on the news articles from Hindu, Indian times and Guardian from Time period ranges from february to august 2017.
- Summaries for that news are taken from Inshorts.
- The dataset consists of 4515 examples and contains Author_name, Headlines, Url of Article, Short text, Complete Article
- The columns which are useful for us are the Short text and the Complete article.
- Short text can be used as the reference text or summary corresponding to the news article
- Complete article is the document containing the news article
- After removing the columns which doesn't have the entries in the either of Short text or Complete article we have a total of 4396 examples

HYPERPARAMETERS

- Number of epochs as 2 because more epochs training is taking more time to train
- Number of documents per batch is 1, because to optimise the memory usage
- Weight decay is 0.01

IMPLEMENTATION

- We used the pre-trained PEGASUS model from the HuggingFace transformer
- Firstly, we read the data and clean the data by removing the columns which doesn't have the entries in the either of Short text or Complete article
- Secondly, split the data approximately as 80% for training data, 10% for validation data, 10% for testing data
- Thirdly, we tokenize the train, validate and test data sets by passing into the prepare_data function, which in turn passes to the tokenize_data function where the embeddings for the training data and corresponding summaries are generated
- Fourthly, these generated embeddings are passed to PegasusDataset class which will create the dictionary of input ids and attention mask and the training, validation and test datasets are now in the required format which can be used for fine tuning the PEGASUS model
- Fifthly, we download the pre-trained $PEGASUS_{large}$ model and call the prepare_fine_tuning function, which will pass the hyperparameters in the TrainingArguments object which will be in turn passed into the Trainer object and returns the base fine tuned model. As we added "do_train = True" argument it will train the model as well after creating base finetune model
- Sixthly, we evaluate the fine tuned model on the test dataset
- Seventhly, we calculate the Rouge-1, Rouge-2, and Rouge-L scores for the test dataset and take the average of the all Rouge scores
- Finally, we show some example summaries generated and reference summaries

EVALUATION METRICS

- **PERPLEXITY:** Perplexity is defined as the exponentiated average negative log-likelihood of a sequence

$$\mathrm{PPL}(X) = \exp\left\{ -\frac{1}{t} \sum_{i}^{t} \log p_\theta(x_i | x_{<i}) \right\}$$

- **ROUGE:** It is the Recall-Oriented Understudy for Gisting Evaluation, there are three variants namely used in the papers which are Rogue-1, Rouge-2, Rouge-L
    1. Rouge-1: This will check the unigram overlaps in the target sentence to that of the reference sentence given
    2. Rouge-2: This will check the bigram overlaps in the target sentence to that of the reference sentence given
    3. Rouge-L: This will check the longest common subsequence in the target sentence to that of the reference sentence

ANALYSIS

- The evaluation loss (perplexity) generated for the test dataset is **4.04**
- Perplexcity here is the exponential of the cross entropy loss
- The Rouge scores calculated for the test dataset are

S.NO	ROUGE	Precision	Recall	F1 measure
1	Rouge-1	37.09	41.43	35.73
2	Rouge-2	17.01	19.09	16.21
3	Rouge-L	25.28	27.78	24.03

- Rouge scores are calculated for each document in the test dataset and they are average of the Rouge scores for the whole corpus is shown in the table
- The Rouge F1 scores comparison to that of state of the art which are tested on the architecture which is trained on the C4 dataset
- R1 is the Rouge-1 F1 measure, R2 is the Rouge-2 F1 measure, RL is the Rouge-L F1 measure

S.NO	Dataset	R1/R2/RL
1	XSum	45.20/22.06/36.99
2	CNN/DailyMail	43.90/21.20/70.76
3	NEWSROOM	45.07/33.39/41.28
4	Multi-News	46.74/17.95/24.26
5	Gigaword	38.75/19.96/36.14
6	**news-summary**	**35.73/16.21/24.03**

- These Rouge scores for the benchmark datasets are calculated with the architecture trained on the C4 dataset
- The pertained model which we downloaded and used in the code is trained on the mixture of both C4 and Hugenews datasets with some proportion of data from each dataset.
- From the above table we can see that the Rouge scores of our dataset is close to the Gigaword dataset

SAMPLE EXAMPLES

Example 1:

Original summary = ['Actress-turned-author Twinkle Khanna, while speaking at an event, said that sex is important at every stage of life. "The things I found extremely attractive in Akshay have changed over time," she added. Twinkle and Akshay, who got married in 2001, completed 16 years of marriage in January this year. They have a 14-year-old son Aarav and a 4-year-old daughter Nitara.']

Generated summary = ["From tracing her story as a little girl who was asked to smile a lot to be liked by everyone to her first kiss resulting in a lot of Maths homework to the funny hashtags running on social media by men asking for equality--Swara Bhaskar nails the society's hypocrisy towards women.She then speaks as Sexism to the woman, about impure grapes being sour."]

Example 2:

Original summary = ['Singer Shreya Ghoshal is set to get a wax figure at the Madame Tussauds in New Delhi. Ghoshal said, "it is an honour to be featured among such talented stars, artists, historians and renowned celebrities." The wax museum, which will open later this year, will also feature wax figures of Bollywood actors Amitabh Bachchan and Shah Rukh Khan. ']

Generated summary = ['With Madame Tussauds all set to come to Delhi, one of the unexpected names of celeb statues that has surfaced is that of singer Shreya Ghoshal.The statue will be created in a distinctive singing pose, and will be open to the public when the museum opens at Regal Palace, in the heart of Delhi later this year.']

Example 3:

Original summary = ['A 26-year-old who has been sitting on an indefinite hunger strike for nearly 10 days for Special Backward Classes quota got married at the protest site in Rajasthan. Devraj Gujjar continued with his "fast unto death" after the rituals while his wife left with her in-laws. Meanwhile, his wife said she would join him if the demands were not met.']

Generated summary = ['Jaipur, Feb 24 (PTI) The site of an indefinite hunger strike for Special Backward Class quota turned into the marriage venue for 26-year-old Devraj Gujjar, who tied the nuptial knot here, as he chose not to leave the protest.']

Example 4:

Original summary = ["The Delhi Metro is planning to play instrumental music in its stations on the New Delhi-Dwarka Airport Line. An official said the decision was taken after a public survey revealed that 80% of people wanted light music in stations. The Metro, which has applied for the required license, will introduce music on other stations and inside trains based on users' feedback. "]

Generated summary = ['Delhiites may soon look forward to a soothing Metro commute during rush hours.The Delhi Metro Rail Corporation has decided to start playing music in stations on the New Delhi-Dwarka airport line and plans to extend it inside trains and across its network gradually, depending on user feedback.The decision to roll out instrumental music on the airport line? Depending on feedback, the music may soon be extended inside trains.?These two companies give license if a public transport facility wants to play music.']

References

- How to Perform Abstractive Summarization with PEGASUS, https://towardsdatascience.com/how-to-perform-abstractive-summarization-with-pegasus-3dd74e48bafb
- PEGASUS: Google's State of the Art Abstractive Summarization Model, https://towardsdatascience.com/pegasus-google-state-of-the-art-abstractive-summarization-model-627b1bbbc5ce
- Pegasus, https://huggingface.co/transformers/v3.2.0/model_doc/pegasus.html
- Fine-tuning with custom datasets, https://huggingface.co/transformers/custom_datasets.html
- The Ultimate Performance Metric in NLP, https://towardsdatascience.com/the-ultimate-performance-metric-in-nlp-111df6c64460
- Perplexity of fixed-length models, https://huggingface.co/transformers/perplexity.html
- Using a Metric, https://huggingface.co/docs/datasets/using_metrics.html
- https://ai.googleblog.com/2020/06/pegasus-state-of-art-model-for.html
- https://en.wikipedia.org/wiki/ROUGE_(metric)
- https://pypi.org/project/rouge-score/