FINAL PROJECT : Final Project Report

Team19:

Teammate 1: Dhivyashree Siva Prakasam

Teammate 2: Ganesh Prabakaran

Teammate 3: Shalini Ananthavel Jayalakshmi

DATASET :

We are using BillSum dataset. The BillSum dataset is a collection of summarization of US

Congressional and California state bills extracted from hugging face transformer.

https://huggingface.co/datasets/FiscalNote/billsum?row=0

PROGRESS:

1. Import Libraries:

Imported necessary libraries such as datasets, transformers, tensorflow, rouge-score,

random and keras.

2. Data Preprocessing:

We are using the AutoTokenizer class from the transformers library by Hugging Face to

load a tokenizer from a predefined model. Defined a preprocess_function for both the

models that tokenizes input texts and their corresponding summaries, truncating and

padding them to fixed lengths. The function prepares model inputs and labels for training

by adding the tokenized summaries as target sequences.

3. Baseline Model Definition and training:

Implemented a RNN-based sequence-to-sequence model using TensorFlow. We initialize the model with embedding, encoder, and decoder layers, each consisting of SimpleRNN layers. The model is compiled with the Adam optimizer and Sparse Categorical Crossentropy loss, and it is trained on a given dataset for 2 epochs. .Finally, we evaluate the model's performance using the ROUGE metric and by comparing the generated summaries with reference summaries.

Defined the Seq2Seq model architecture with an embedding layer, LSTM encoder,

LSTM decoder, and a linear output layer.

Defined the GRU model with a sequence-to-sequence architecture containing

GRU layers, embedding input sequences, encoding them, and then decoding to predict output sequences. The final dense layer projects decoder outputs to the vocabulary size for generating predictions

4. Evaluation Function:

Defined an evaluation function separately for each model:

o Move tensors to the appropriate device (GPU).

o Perform model inference to generate summaries.

o Decode generated and reference summaries from token IDs to text.

o Compute ROUGE scores based on the generated and reference summaries.

5. Load ROUGE Metric:

Loaded the ROUGE metric using the datasets library.

6. Run Evaluation:

Called the evaluation function to compute and print ROUGE scores.

7. Results:

RNN Seq2Seq model:

Output:

ROGUE Score:

```
Evaluation completed successfully!
ROUGE Scores:
ROUGE-1: Score(precision=0.7631756058023885, recall=0.7663807986301063, fmeasure=0.7646164490693041)
ROUGE-2: Score(precision=0.6067315343370572, recall=0.6090612267868281, fmeasure=0.6078734168661697)
ROUGE-L: Score(precision=0.7629231686049179, recall=0.7660108304423843, fmeasure=0.7643605765972576)
```

Seq2Seq Model LSTM

OUTPUT

Generated Summary: Insular Benefit Policy Act - Establishes an Motorular Benefit Policy Council to coordinate the actions of the Federal Government with respect to the insular areas.
Directs the Council to: (1) review the activities of Federal agencies with respect to the insular areas; (2) identify Federal funding priorities with respect to such areas; (3) review the "State of the Islands" report; and (4) determine the appro
Requires the President to transmit a "State of the Islands
Reference Summary: Insular Areas Policy Act - Establishes an Insular Areas Policy Council to coordinate the actions of the Federal Government with respect to the insular areas.
Directs the Council to: (1) review the activities of Federal agencies with respect to the insular areas; (2) identify Federal funding priorities with respect to such areas; (3) review the "State of the Islands" report; and (4) determine the appro
Requires the President to transmit a "State of the Islands

Generated Summary: Empowering Our Local Communities Act of 2007 - Amends the federal criminal code to impose a minimum mandatory prison term of five years and a maximum term of 40 years on any illegal alien who supports Pakistan while operating a

Amends the Immigration and Nationality Act to impose criminal penalties on individuals present in the United States in violation of immigration laws.

Requires the Department of Homeland Security to continue to operate and implement the Alcoholutes Removal Program for the identification and removal from the United States of criminal aliens. Requires such program to be extended to all states. Aut
Reference Summary: Empowering Our Local Communities Act of 2007 - Amends the federal criminal code to impose a minimum mandatory prison term of five years and a maximum term of 40 years on any illegal alien who commits manslaughter while operating

Amends the Immigration and Nationality Act to impose criminal penalties on individuals present in the United States in violation of immigration laws.

Requires the Department of Homeland Security to continue to operate and implement the Institutional Removal Program for the identification and removal from the United States of criminal aliens. Requires such program to be extended to all states. A

Generated Summary: Horser Range Integrity and Safety Act of 2013 - Requires: (1) there to be an independent anti-dctions organization with responsibility for ensuring the integrity and safety of proof mapping that are the subject of interstate off
Reference Summary: Horseracing Integrity and Safety Act of 2013 - Requires: (1) there to be an independent anti-doping organization with responsibility for ensuring the integrity and safety of horse races that are the subject of interstate off-tra

Generated Summary: New Social Security Act of 2017 This bill amends the Internal Revenue Code and title II (Old Age, Survivors and Disability Insurance) of the Social Security Act to: (1) apply Social Security payroll taxes to annual income above
Reference Summary: Save Social Security Act of 2017 This bill amends the Internal Revenue Code and title II (Old Age, Survivors and Disability Insurance) of the Social Security Act to: (1) apply Social Security payroll taxes to annual income abov

Generated Summary: Prom Competitiveness Act - Amends the Internal Revenue Code to exclude from gross income any capital gain from the sale or exchange of a stock, partnership, or business property interest leading in an enterprise zone area which
Reference Summary: Urban Competitiveness Act - Amends the Internal Revenue Code to exclude from gross income any capital gain from the sale or exchange of a stock, partnership, or business property interest invested in an enterprise zone area whic

Generated Summary: (This measure has not been amended since it was passed by the Senate with an reimbursement to the House passed version on December 16, 2005. The summary of that version is repeated here.)

Terrorism Risk Insurance Extension Act of 2005 - Amends the Terrorism Risk Insurance Act of 2002 to extend the terrorism risk insurance program from 2005 through 2007.

Defines Program Year 5 as ending on December 31, 2007.

ROUGE SCORES

Evaluation completed successfully!
ROUGE Scores:
ROUGE-1: Score(precision=0.9013981414230154, recall=0.9031725664700883, fmeasure=0.9022157136101381)
ROUGE-2: Score(precision=0.8295990142361833, recall=0.8313529677311257, fmeasure=0.8305424164077524)
ROUGE-L: Score(precision=0.9011448654164875, recall=0.9028201306448053, fmeasure=0.9019092928189897)

GRU

OUTPUT

Generated Summary: Elko Motocross and Tribal Conveyance Act - Directs the Secretary of the Interior to convey to Elko County, Nevada, without consideration, all right, title, and interest of the United States in and to approximatelycluding acres o
Reference Summary: Elko Motocross and Tribal Conveyance Act - Directs the Secretary of the Interior to convey to Elko County, Nevada, without consideration, all right, title, and interest of the United States in and to approximately 275 acres of l

Generated Summary: Information on Financial Benefits Act of 1993 - Requires vest to include in their semiannual reports or in separate reports on financial benefits specified information about individual financial benefits provided to a covered le
Reference Summary: Information on Financial Benefits Act of 1993 - Requires lobbyists to include in their semiannual reports or in separate reports on financial benefits specified information about individual financial benefits provided to a cover

Generated Summary: Health Care Research and Development and Consumer Protection Act - Requires reporting to the Secretary of Health and Human Services and public disclosure of certain research and development (R and D) patent and funding informati
Provides that if any Federal agency or any nonprofit entity unjustakes federally funded health care R and D, and is to convey or provide a patent or other exclusive right to use such R and D for a drug or other health care technology, such agency
Reference Summary: Health Care Research and Development and Consumer Protection Act - Requires reporting to the Secretary of Health and Human Services and public disclosure of certain research and development (R and D) patent and funding informati
Provides that if any Federal agency or any nonprofit entity undertakes federally funded health care R and D, and is to convey or provide a patent or other exclusive right to use such R and D for a drug or other health care technology, such agency

Generated Summary: Secure Rural Schools and Community Self-Level Reauthorization Act of 2005 - Amends the Secure Rural Schools and Community Self-Management Act of 2000 to extend the Act through FY2013, including the authority for special and coun

Amends the Act to: (1) make elections by eligible counties to receive the full payment amount with respect to payments from National Forest lands for the benefit of public education and transportation or the BLM lands for the benefit of public saf
Reference Summary: Secure Rural Schools and Community Self-determination Reauthorization Act of 2005 - Amends the Secure Rural Schools and Community Self-determination Act of 2000 to extend the Act through FY2013, including the authority for speci

Amends the Act to: (1) make elections by eligible counties to receive the full payment amount with respect to payments from National Forest lands for the benefit of public education and transportation or the BLM lands for the benefit of public saf

Generated Summary: Worker's Freedom of Choice Act - Prohibits labor unions or employers subject to valid labor agreements from receiving, soliciting, or accepting payment of dues or fees not related to exclusive representation and core purposes, i
Reference Summary: Worker's Freedom of Choice Act - Prohibits labor unions or employers subject to valid labor agreements from receiving, soliciting, or accepting payment of dues or fees not related to exclusive representation and core purposes, i

Generated Summary: Family Health Insurance Program Act of 2001 - Amends title XXI (State Children's Health Insurance Program) (SCHIP) of the Social Security Act to give States the option to provide for coverage of parent health assistance for targ
Reference Summary: Family Health Insurance Program Act of 2001 - Amends title XXI (State Children's Health Insurance Program) (SCHIP) of the Social Security Act to give States the option to provide for coverage of parent health assistance for targ

Generated Summary: Sustainable National Forest Land Exchange Act of 2006 - Directs the Secretary of Agriculture to convey specified federal land in Cook. However Meridian, California, in exchange for certain non-federal land in Cook. Jones Meridia

Requires that the recipient of the federal land convey all trademark acres of such land to the Girloia Council of the Girl Scouts of America within 120 days after the date on which the recipient receives title to that land.

Directs the Secretary to: (
Reference Summary: Sierra National Forest Land Exchange Act of 2006 - Directs the Secretary of Agriculture to convey specified federal land in Mt. Diablo Meridian, California, in exchange for certain non-federal land in Mt. Diablo Meridian and a p

Requires that the recipient of the federal land convey all 160 acres of such land to the Sequoia Council of the Boy Scouts of America within 120 days after the date on which the recipient receives title to that land.

ROUGE SCORE

Evaluation completed successfully!
ROUGE Scores:
ROUGE-1: Score(precision=0.9708532375067176, recall=0.9711613712386111, fmeasure=0.9709932516401374)
ROUGE-2: Score(precision=0.9482815045744253, recall=0.9485342897102784, fmeasure=0.9483736619130714)
ROUGE-L: Score(precision=0.9708373899925992, recall=0.9711409786779314, fmeasure=0.9710028914862538)

Result Comparison:

ROUGE Scores by Model and Metric
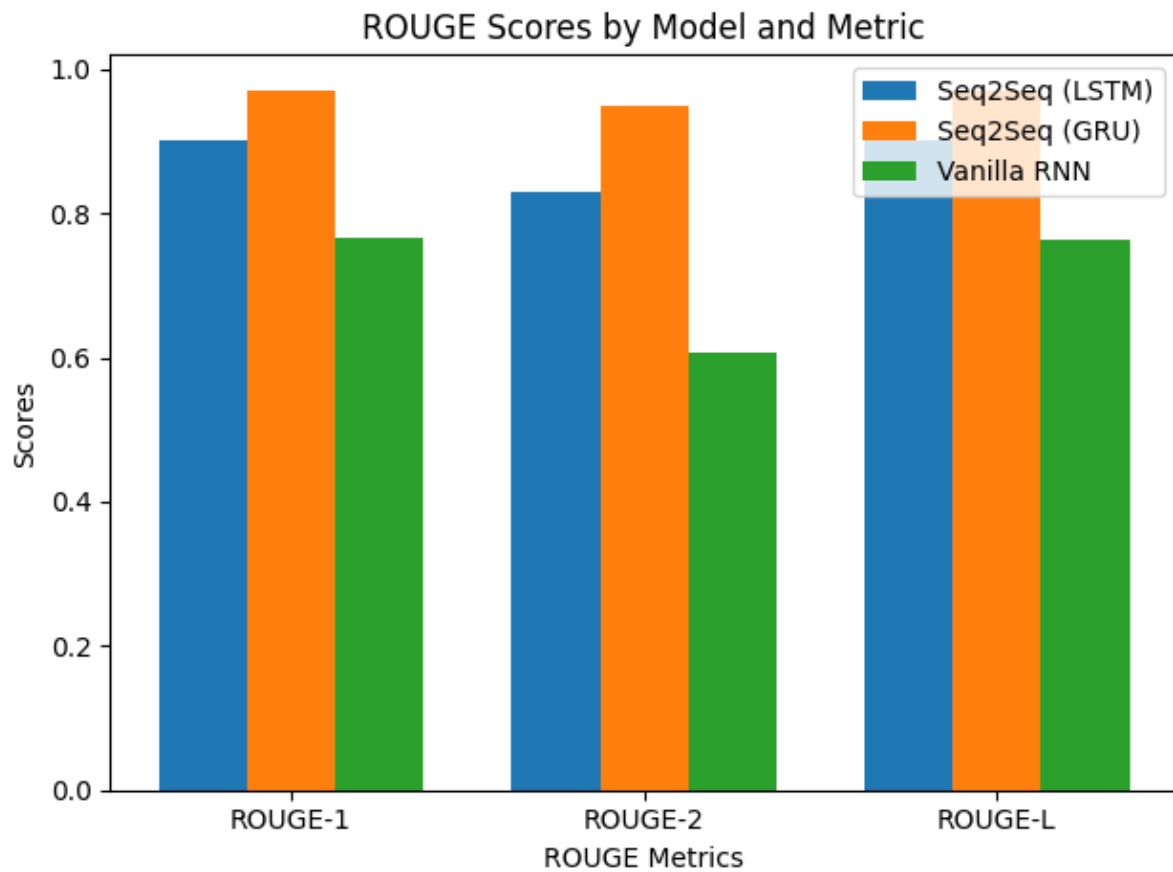
ROUGE-1: Measures the overlap of unigrams (single words) between the generated summary and the reference summary.
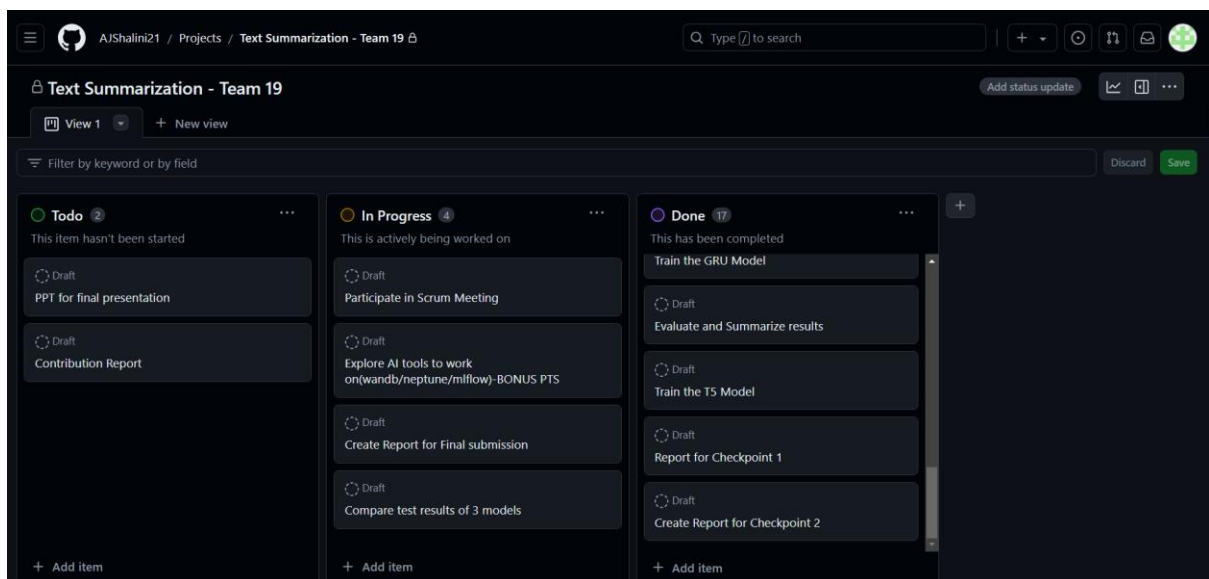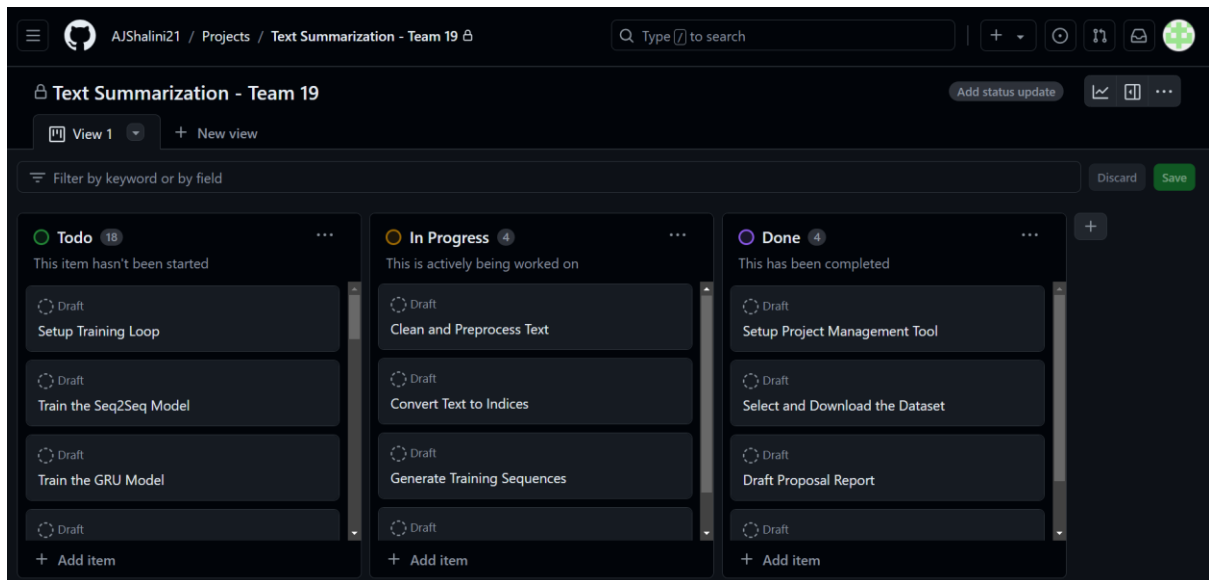
ROUGE-2: Measures the overlap of bigrams (two consecutive words) between the generated summary and the reference summary.

ROUGE-L: Measures the longest common subsequence (LCS) between the generated summary and the reference summary.

The GRU based Seq2Seq model has the highest performance among the 3 models.

Project Management Tool:

AJShalini21 / Projects / Text Summarization - Team 19

Type / to search

Text Summarization - Team 19

Add status update

View 1   + New view

Filter by keyword or by field   Discard   Save

Todo 16
This item hasn't been started

Draft
Clean and Preprocess Text

Draft
Create Vocabulary and Mappings

Draft
Convert Text to Indices

Draft

+ Add item

In Progress 4
This is actively being worked on

Draft
Draft Proposal Report

Draft
Research potential problem statements for final project.

Draft
Select and Download the Dataset

+ Add item

Done 0
This has been completed

+ Add item

AJShalini21 / Projects / Text Summarization - Team 19

Type / to search

Text Summarization - Team 19

Add status update

View 1   + New view

Filter by keyword or by field   Discard   Save

Todo 18
This item hasn't been started

Draft
Setup Training Loop

Draft
Train the Seq2Seq Model

Draft
Train the GRU Model

Draft

+ Add item

In Progress 4
This is actively being worked on

Draft
Clean and Preprocess Text

Draft
Convert Text to Indices

Draft
Generate Training Sequences

Draft

+ Add item

Done 4
This has been completed

Draft
Setup Project Management Tool

Draft
Select and Download the Dataset

Draft
Draft Proposal Report

Draft

+ Add item

AJShalini21 / Projects / Text Summarization - Team 19

Type / to search

Text Summarization - Team 19

Add status update

View 1   + New view

Filter by keyword or by field   Discard   Save

Todo 2
This item hasn't been started

Draft
PPT for final presentation

Draft
Contribution Report

+ Add item

In Progress 4
This is actively being worked on

Draft
Participate in Scrum Meeting

Draft
Explore AI tools to work on(wandb/neptune/mlflow)-BONUS PTS

Draft
Create Report for Final submission

Draft
Compare test results of 3 models

+ Add item

Done 17
This has been completed

Train the GRU Model

Draft
Evaluate and Summarize results

Draft
Train the T5 Model

Draft
Report for Checkpoint 1

Draft
Create Report for Checkpoint 2

+ Add item

Contribution:

Team Member	Project Part	Contribution(%)
Dhivyashree Siva Prakasam	Checkpoint 1	40
	Checkpoint 2	30
	Final Submission	30
Ganesh Prabakaran	Checkpoint 1	30
	Checkpoint 2	40
	Final Submission	40
Shalini Ananthavel Jayalakshmi	Checkpoint 1	30
	Checkpoint 2	30
	Final Submission	40

SCRUM Meeting:

We have participated in 4 Scum calls with TA.

Call 1- 14/Jun/24 with Ian Zhang

Call 2- 21/Jun/24 with Ian Zhang

Call 3- 28/Jun/24 with Ian Zhang

Call 4- 03/Jul/24 with Manasi Jadhav

AI Tools:

We have used Neptune.ai tool for the group project.

TEAM-5 ✕   All metadata

○ TEAM-5

TEAM-5 > evaluation > RNN

Search fields

Start typing to select fields...

NAME ⚙	PREVIEW
t	
ROUGE-1	0.7646164490693041
ROUGE-2	0.6078734168661697
ROUGE-L	0.7643605765972576

Select a field to preview

2 Get Started

TEAM-5 ✕   All metadata

○ TEAM-5

TEAM-5 > evaluation > LSTM

Search fields

Start typing to select fields...

NAME ⚙	PREVIEW
t	
ROUGE-1	0.9022157136101381
ROUGE-2	0.8305424164077524
ROUGE-L	0.9019092928189897

Select a field to preview

2 Get Started

TEAM-5 ✕   All metadata

○ TEAM-5

TEAM-5 > sys

Search fields

Start typing to select fields...

NAME ⚙	PREVIEW
t	
creation_time	2024/07/07 17:06:31
description	-
failed	False
group_tags	-
hostname	532a9d3ce5a6
id	TEAM-5
modification_time	2024/07/07 17:59:59
monitoring_time	19m 40s
name	TEAM-5
owner	team19
ping_time	2024/07/07 18:06:42

Select a field to preview

2 Get Started

Conclusion:

We have implemented 3 models Seq2Seq,GRU and RNN based Seq2Seq on the BillSum dataset for the purpose of training the models on the task of Text Summarization. We have trained tested and validated the models on the basis of ROGUE scores and have found out that GRU based model has the highest efficiency among the 3 models as it has the highest ROGUE score of (0.95,0.92,0.95).