CS 6320.501 assignment4

Humza Salman, Kevin Tak Hay Cheung

TOTAL POINTS

100 / 100

QUESTION 1

1 Intro and Data 10 / 10

✓ **- 0 pts** *Correct*

- 0 pts Briefly describe the project and your main **experiments** and **results**, including mentioning the **data** you use.

- 1 pts not mentioned Question Answering task is a task {blablabla}. In this task, we utilize {dataset} on {model} and the overall {performance}.
no main **results**. eg. f1 score 90.XX
no mentioning the **data** you use. (Penn Treebank)

- 1 pts Abstractive Summarization task is a task {blablabla}. In this task, we utilize {dataset} on {model} and the overall {performance}.
no main **results**. eg. f1 score 90.XX
no mentioning the **data** you use. (Penn Treebank)

- 1 pts Multi-Document Summarization task is a task {blablabla}. In this task, we utilize {dataset} on {model} and the overall {performance}.
no main **results**. eg. f1 score 90.XX
no mentioning the **data** you use. (Penn Treebank)

- 10 pts this part empty

- 0 pts Although your intro and data part is empty. But you have detailed report. And I will adjust your points by Point Adjustment

QUESTION 2

2 Question Answering 30 / 30

✓ **- 0 pts** *Correct*

- 1 pts 2.1. Question Answering is a task {blablabla} (more detailed explanation).
deduct because: no task detailed explanation

- 1 pts 2.2. {dataset} show an example (e.g., Question, Context, Answer)
deduct because: not showing dataset example

- 2 pts 2.3. {model} more detailed explanation of the model
deduct because: no detailed explanation of the model

- 2 pts 2.4.1 {performance} high-level performance results
(in table, compare with different model, showing your accuracy score):

You pick up 10 examples. You have your own judgment that the predicted output is correct or not. Finally, compute the accuracy score.

deduct because: no high-level performance results

- 2 pts 2.4.2 {performance} low-level

performance results

low-level performance results(think of error analysis):
You pick an example; let's say you pick datapoint403, you have the judged answer (correct answer), which is answer403, and you compare it to the predictions by a different model.

datapoint403, answer403, bert-base-cased-squad2p_predict403, roberta-base-squad2_predict403, roberta-base-squad2-distilled_predict403. Try to analyze why different models have different predicted results. You try to find an example where the model differs in its predictions.

deduct because: no low-level performance results

- **5 pts** Task empty
- **5 pts** dataset empty
- **5 pts** model empty
- **5 pts** high level performance empty
- **5 pts** low level performance empty

QUESTION 3

3 Abstractive Summarization 30 / 30

✓ **- 0 pts** *Correct*

- **1 pts** 3.1. Task

deduct because: no task detailed explanation

- **1 pts** 3.2. {dataset}

deduct because: not showing dataset example

- **2 pts** 3.3. {model} more detailed explanation of the model

deduct because: no detailed explanation of the model

- **2 pts** 3.4.1 {performance} high-level performance results

deduct because: no high-level performance results

- **2 pts** 3.4.2 {performance} low-level performance results

deduct because: no low-level performance results

- **5 pts** Task empty
- **5 pts** dataset empty
- **5 pts** model empty
- **5 pts** high level performance empty
- **5 pts** low level performance empty

QUESTION 4

4 Multi-Document Summarization 30 / 30

✓ **- 0 pts** *Correct*

- **1 pts** 4.1. Task

deduct because: no task detailed explanation

- **1 pts** 4.2. {dataset}

deduct because: not showing dataset example

- **2 pts** 4.3. {model} more detailed explanation of the model

deduct because: no detailed explanation of the model

- **2 pts** 4.4.1 {performance} high-level performance results

deduct because: no high-level performance results

- **2 pts** 4.4.2 {performance} low-level performance results

deduct because: no low-level performance results

- **5 pts** task empty
- **5 pts** dataset empty
- **5 pts** model empty
- **5 pts** high level performance empty
- **5 pts** low level performance empty

gradescope

CS6320: Assignment 4

https://github.com/cs6320-501-fall22/assignment3-group18/tree/master

Group 18 Kevin Tak Hay Cheung (kxc220019) Humza Salman (mhs180007)

Introduction and Datasets

In this assignment, we are going to learn several state-of-the-art NLP models for three NLP tasks, including Question Answering, Abstractive Summarization, and Multi-Document Summarization. The goal of this assignment is to evaluate the accuracies of these models and compare their performances.

Question Answering

Question Answering is a task where the NLP model will provide an answer for the questions posed by the users. There are three main types of question answering models, including:

- Extractive Question Answering: The model will extract a short phrase or sentence from the document provided by the users as the answer.
- Open Generative Question Answering: Instead of extracting the answer, the NLP model will generate an original answer based on the given document.
- Closed Generative Question Answering: The model will generate an answer without any given document.

For this assignment, we are going to experiment with three pre-trained NLP models (trained with SQuAD2.0 dataset) to perform an extractive question answering task. We will feed the validation set into these three models and compare the performance.

Data Preprocessing

To test the model, we will be using the SQuAD2.0 dataset. It consists of more than 100,000 reading comprehension questions, posed by crowdworkers based on articles in Wikipedia. The answers to these questions are always a segment of the document. On top of that, there are over 50,000 questions that are deemed unanswerable, which means the questions cannot be found in the given document. To perform well in this dataset, the model should be able to provide the correct span of text for the answerable questions, and also identify the unanswerable problems.

Before feeding the question and document into the three pre-trained models, we will tokenize them using WordPiece. The tokenizer will split each word into either the full word or several word pieces. For example, the word “deterministically” will be splitted into 5 pieces ('de' , '##ter', '##mini', '##stic', '##ally'), while the word ‘machine’ will be intact after the tokenization. Also, the tokenizer will convert each text token into three id tokens, including inputs_ids token (an unique id for each word with same spelling), token_type_ids token (used to identify if the token belongs to the document or question) and attention_mask token (used to identify the padded text token so that no attention will be applied to them).

After the preprocessing, we are ready to feed the tokenized inputs into the models for the prediction, which should return a span of text as an extractive QA.

```
Question:
  When was the drainage basin of the Amazon believed
  to have split in the middle of South America?
Context:
  During the mid-Eocene, it is believed that the
  drainage basin of the Amazon was split along the
  middle of the continent by the Purus Arch. Water
  on the eastern side flowed toward the Atlantic,
  while to the west water flowed toward the Pacific
  across the Amazonas Basin. As the Andes Mountains
  rose, however, a large basin was created that
  enclosed a lake; now known as the Solimões Basin.
  Within the last 5–10 million years, this
  accumulating water broke through the Purus Arch,
  joining the easterly flow toward the Atlantic.
Answer:
  During the mid-Eocene
```

Sample Entry Number 3411 from SQuAD2.0:

Model Explanation

The three candidate models are variations of the Bidirectional Encoder Representation for Transformer (BERT), a very popular model choice for the pre-training/fine-tuning paradigm. The goal of the pre-training of a model is to accomplish the

1 Intro and Data 10 / 10

✓ **- 0 pts** *Correct*

- 0 pts Briefly describe the project and your main **experiments** and **results**, including mentioning the **data** you use.

- 1 pts not mentioned Question Answering task is a task {blablabla}. In this task, we utilize {dataset} on {model} and the overall {performance}.

no main **results**. eg. f1 score 90.XX

no mentioning the **data** you use. (Penn Treebank)

- 1 pts Abstractive Summarization task is a task {blablabla}. In this task, we utilize {dataset} on {model} and the overall {performance}.

no main **results**. eg. f1 score 90.XX

no mentioning the **data** you use. (Penn Treebank)

- 1 pts Multi-Document Summarization task is a task {blablabla}. In this task, we utilize {dataset} on {model} and the overall {performance}.

no main **results**. eg. f1 score 90.XX

no mentioning the **data** you use. (Penn Treebank)

- 10 pts this part empty

- 0 pts Although your intro and data part is empty. But you have detailed report. And I will adjust your points by Point Adjustment

gradescope

CS6320: Assignment 4

https://github.com/cs6320-501-fall22/assignment3-group18/tree/master

Group 18 Kevin Tak Hay Cheung (kxc220019) Humza Salman (mhs180007)

Introduction and Datasets

In this assignment, we are going to learn several state-of-the-art NLP models for three NLP tasks, including Question Answering, Abstractive Summarization, and Multi-Document Summarization. The goal of this assignment is to evaluate the accuracies of these models and compare their performances.

Question Answering

Question Answering is a task where the NLP model will provide an answer for the questions posed by the users. There are three main types of question answering models, including:

- Extractive Question Answering: The model will extract a short phrase or sentence from the document provided by the users as the answer.
- Open Generative Question Answering: Instead of extracting the answer, the NLP model will generate an original answer based on the given document.
- Closed Generative Question Answering: The model will generate an answer without any given document.

For this assignment, we are going to experiment with three pre-trained NLP models (trained with SQuAD2.0 dataset) to perform an extractive question answering task. We will feed the validation set into these three models and compare the performance.

Data Preprocessing

To test the model, we will be using the SQuAD2.0 dataset. It consists of more than 100,000 reading comprehension questions, posed by crowdworkers based on articles in Wikipedia. The answers to these questions are always a segment of the document. On top of that, there are over 50,000 questions that are deemed unanswerable, which means the questions cannot be found in the given document. To perform well in this dataset, the model should be able to provide the correct span of text for the answerable questions, and also identify the unanswerable problems.

Before feeding the question and document into the three pre-trained models, we will tokenize them using WordPiece. The tokenizer will split each word into either the full word or several word pieces. For example, the word “deterministically” will be splitted into 5 pieces ('de' , '##ter', '##mini', '##stic', '##ally'), while the word ‘machine’ will be intact after the tokenization. Also, the tokenizer will convert each text token into three id tokens, including inputs_ids token (an unique id for each word with same spelling), token_type_ids token (used to identify if the token belongs to the document or question) and attention_mask token (used to identify the padded text token so that no attention will be applied to them).

After the preprocessing, we are ready to feed the tokenized inputs into the models for the prediction, which should return a span of text as an extractive QA.

```
Question:
  When was the drainage basin of the Amazon believed
  to have split in the middle of South America?
Context:
  During the mid-Eocene, it is believed that the
  drainage basin of the Amazon was split along the
  middle of the continent by the Purus Arch. Water
  on the eastern side flowed toward the Atlantic,
  while to the west water flowed toward the Pacific
  across the Amazonas Basin. As the Andes Mountains
  rose, however, a large basin was created that
  enclosed a lake; now known as the Solimões Basin.
  Within the last 5–10 million years, this
  accumulating water broke through the Purus Arch,
  joining the easterly flow toward the Atlantic.
Answer:
  During the mid-Eocene
```

Sample Entry Number 3411 from SQuAD2.0:

Model Explanation

The three candidate models are variations of the Bidirectional Encoder Representation for Transformer (BERT), a very popular model choice for the pre-training/fine-tuning paradigm. The goal of the pre-training of a model is to accomplish the

pre-training objectives and obtain a set of trained parameters. In the fine-tuning stage, for any given downstream task, the relevant dataset will be loaded to the model with the weights initialized as the pre-trained parameters. The training process will carry on with these parameters.

With BERT model, during the pre-training process, corpus will be loaded into the network to perform two pre-training tasks:

1) Masked Language Modeling: The contextual representation would be more useful if the model is trained bidirectionally. However, the self-attention nature of the transformer would allow the model to "see the future", resulting in a zero training loss. As a result, the BERT authors decided to mask part of the inputs randomly and asked BERT to predict these tokens as the first pre-training objective.
2) Next Sentence Prediction: Wanting the BERT model to better understand the relationships between sentences, the authors asked BERT to perform a classification task - feeding two sentences to the model and identify if they belong to the same document.

After the pre-training, the BERT model is now with the pre-trained hidden state parameters. The next step is to fine-tune the model with the SQuAD2.0 dataset. The training dataset (the document and the question) is loaded to the BERT model along with the target (the starting and ending indices of the answer). At the end, we would receive a BERT model specifically trained for the extractive question answering task, which would be our first candidate model.

The second candidate model is the RoBERTa, which is an optimized BERT model. The RoBERTa authors believe that the original BERT model is highly undertrained. As a result, they proposed several improvements during the pre-training process, including:

- Dynamic Masking for Masked Language Modeling: Instead of masking the text tokens at the preprocessing stage, the masking would be completed randomly in each epoch.
- Next sentence prediction improvement: Several documents will be concatenated together and sentence pairs will be sampled from this packed document. The pre-trained task is still to predict whether these sentence pairs belong to the same document.
- Training with larger batch sizes: After experiments, it is discovered that larger batch sizes can help with the downstream task.
- Text Encoding: RoBERTa is trained with a larger byte-level Byte-Pair Encoding (BPE). Originally, BERT used a character-level BPE vocabulary with size = 30k. The RoBERTa modelers propose using a larger BPE vocabulary with size = 50K.

The third model is the distilled RoBERTa. A more simple model, the student model, is trained to reproduce the behavior of a more complex model, the teacher model. In this case, the student, the distilled RoBERTA, is derived from the complex RoBERTa by removing some unnecessary embedding, such as the token-type embedding in the input. Also, the calculation inside the hidden layers are replaced by a more efficient linear algebra framework. The authors claim that the student model can retain about on average 97% of the knowledge from the teacher model.

Performance

Sample #	Index	Question	Context	Answer	BERT Prediction	RoBERTa Prediction	Distilled RoBERTa Prediction
1	339	What is an example of a machine model that deviates from a generally accepted multi-tape Turing machine?	Many machine model	random access machines	random access machines	random access machines	random access machines
2	1729	What nation did the most Huguenots flee to from France?	The bulk of Huguenot	No Answer Provided	[CLS]	<s>	<s>
3	3411	When was the drainage basin of the Amazon believed to have split in the middle of South America?	During the mid-Eocen	During the mid-Eocene	[CLS]	<s>	<s>
4	6540	What is the 16th century known as the start of?	The area of the mode	the historical era	historical era	the historical era	the historical era
5	10395	Who was made rich and prosperous prior to World War 1	During the 20th centu	many imperial powers	[CLS]	imperial powers	many imperial powers
6	5641	Civil disobedience can occur when people speak about a certain topic that is deemed as?	In cases where the cri	criminalized behavior	forbidden speech	pure speech	pure speech
7	11165	Where did British settlers live?	British settlers outnu	from Nova Scotia and Newfoundland in the north, to Georgia in the south	Nova Scotia and Newfoundland in the north, to Georgia in the south	along the eastern coast of the continent	along the eastern coast of the continent
8	7568	Who is the current Chairman of President Barack Obama's Council of Economic Advisors?	Current faculty includ	No Answer Provided	Austan Goolsbee	<s>	<s>
9	5678	Why is the need for acceptance of punishment needed?	Some civil disobedien	validity of the social contr	[CLS] Why is the need for acceptance of punishment needed? [SEP] Some civil disobedients feel it is incumbent upon them to accept punishment because of their belief in the validity of the social contract	belief in the validity of the social contract	their belief in the validity of the social contract
10	6797	What motivation is opportunity-based entrepreneurship driven by?	On the other hand, hi	achievement-oriented	achievement - oriented motivations	achievement-oriented motivations	achievement-oriented motivations ("pull")

10 sample predictions of the BERT, RoBERTa, and Distilled RoBERTa models

We find that the accuracy rating for the models is as follows:

Model	Accuracy
BERT	60%
RoBERTa	70%
Distilled RoBERTa	70%

When performing the predictions, we found that simpler questions paired with simpler context passages where the answer was within a single sentence, all models performed well. However, BERT failed to perform well with context passages where the answer was not within a single sentence. An example of this would be Sample #7568. In the context passage, it only details the *former Chairman*, however BERT disregards the word *former* and predicts *Austan Goolsbee* as the current Chairman. RoBERTa and Distilled RoBERTa are able to recognize this difference. Similarly, we see this occurrence in Sample #10395. One reason for this happening could simply be the pre-trained differences in the models. RoBERTa and Distilled RoBERTa are successors of the BERT models and improvements in the pre-training process using different dynamic masking or sentence prediction improvement techniques can be seen to make a marked improvement. In Sample #11165 BERT is the only model to be precisely correct. Although the respective context passage details that the English settlers lived along the eastern coast of the continent, BERT details which places, hence we chose to award points to BERT as it provided the more specific answer closest to the original given answer.

Abstractive Summarization

Abstractive Summarization is a task where the NLP model generates a short and concise summary for a given document. The state of the art model is called Pre-training with Extracted Gap sentences for Abstractive Summarization (PEGASUS). In this following section, we will experiment with this model pre-trained by the CNN Daily Mail dataset.

Data Preprocessing

To test the model, we will be using the CNN/Daily Mail 3.0.0 dataset. It consists of more than 300,000 articles from CNN and Daily news. Each article is paired with a summary / highlight.

Before feeding the article into PEGASUS, we will divide the document into unigram using SentencePiece. The tokenizer will also include an underscore in front of every token to represent spacing. For some word abbreviations, such as “didn’t”, they will be tokenized as “_didn”, “‘“, “t”. Note that there are no underscores prior to the last two tokens since there is no space in front of them from the original text token. Also, the tokenizer will convert each text token into two id tokens, including inputs_ids token (an unique id for each word with same spelling) and attention_mask token (used to identify the padded text token so that no attention will be applied to them).

```
Document:
  Ivan Rakitic hit the beach with his wife to
  prepare for a crucial week in Barcelona’s season
  as Luis Enrique’ side pursues a treble. The La
  Liga leaders face Manchester City in the Champions
  League last 16 second-leg, having won the first-
  leg 2-1 last month, before hosting Real Madrid in
  El Clasico on Sunday. Rakitic has become an
  integral part of the Barcelona midfield ranks
  since joining from Sevilla last summer, making 36
  appearances so far this season for the Catalan
  giants. Barcelona star Ivan Rakitic enjoys a sunny
  afternoon at the beach with his wife ahead of a
  crucial week . The Croatian international posted a
  sunny snap on his Instagram with his Spanish wife
  Raquel Mauri ahead of the Manchester City clash
  with the caption ‘familiar Sunday’. Seemingly
  Rakitic is a regular on Barcelona’s beach but it’s
  not all leisure for the Croatian, having posted
  another picture on his Instagram account on Sunday
  from the team gym as he prepares for a ‘big week’.
  Luis Enrique’s side remained top of La Liga on
  Saturday with a 2-0 victory at Eibar, courtesy of
  a brace from Lionel Messi. Enrique is hoping to
  lead Barcelona to a treble, having reached the
  Copa del Rey final, the Champions League knockout
  stages and are currently leading La Liga. The
  Croatian (left) celebrates after Lionel Messi
  (right) fires Barcelona to victory at Eibar on
  Saturday . VIDEO We were clinical - Enrique .
  Rakitic is all smiles ahead of a busy week, facing
  Manchester City and then Real Madrid in El Clasico
  .
Annotated Summary:
  Barcelona beat Eibar 2-0 at the weekend to remain
  top of La Liga by a point . Luis Enrique's side
  face Manchester City in the Champions League last
  16 on Wednesday before hosting Real Madrid in El
  Clasico on Sunday . Ivan Rakitic has become an
  integral part of the Barcelona midfield .
```

Sample Entry #3411 from CNN_DailyMail

2 Question Answering 30 / 30

✓ **- 0 pts** *Correct*

- 1 pts 2.1. Question Answering is a task {blablabla} (more detailed explanation).

deduct because: no task detailed explanation

- 1 pts 2.2. {dataset} show an example (e.g., Question, Context, Answer)

deduct because: not showing dataset example

- 2 pts 2.3. {model} more detailed explanation of the model

deduct because: no detailed explanation of the model

- 2 pts 2.4.1 {performance} high-level performance results

(in table, compare with different model, showing your accuracy score):

You pick up 10 examples. You have your own judgment that the predicted output is correct or not. Finally, compute the accuracy score.

deduct because: no high-level performance results

- 2 pts 2.4.2 {performance} low-level performance results

low-level performance results(think of error analysis):
You pick an example; let's say you pick datapoint403, you have the judged answer (correct answer), which is answer403, and you compare it to the predictions by a different model.

datapoint403, answer403, bert-base-cased-squad2p_predict403, roberta-base-squad2_predict403, roberta-base-squad2-distilled_predict403. Try to analyze why different models have different predicted results. You try to find an example where the model differs in its predictions.

deduct because: no low-level performance results

- 5 pts Task empty

- 5 pts dataset empty

- 5 pts model empty

- 5 pts high level performance empty

- 5 pts low level performance empty

gradescope

We find that the accuracy rating for the models is as follows:

Model	Accuracy
BERT	60%
RoBERTa	70%
Distilled RoBERTa	70%

When performing the predictions, we found that simpler questions paired with simpler context passages where the answer was within a single sentence, all models performed well. However, BERT failed to perform well with context passages where the answer was not within a single sentence. An example of this would be Sample #7568. In the context passage, it only details the *former Chairman*, however BERT disregards the word *former* and predicts *Austan Goolsbee* as the current Chairman. RoBERTa and Distilled RoBERTa are able to recognize this difference. Similarly, we see this occurrence in Sample #10395. One reason for this happening could simply be the pre-trained differences in the models. RoBERTa and Distilled RoBERTa are successors of the BERT models and improvements in the pre-training process using different dynamic masking or sentence prediction improvement techniques can be seen to make a marked improvement. In Sample #11165 BERT is the only model to be precisely correct. Although the respective context passage details that the English settlers lived along the eastern coast of the continent, BERT details which places, hence we chose to award points to BERT as it provided the more specific answer closest to the original given answer.

Abstractive Summarization

Abstractive Summarization is a task where the NLP model generates a short and concise summary for a given document. The state of the art model is called Pre-training with Extracted Gap sentences for Abstractive Summarization (PEGASUS). In this following section, we will experiment with this model pre-trained by the CNN Daily Mail dataset.

Data Preprocessing

To test the model, we will be using the CNN/Daily Mail 3.0.0 dataset. It consists of more than 300,000 articles from CNN and Daily news. Each article is paired with a summary / highlight.

Before feeding the article into PEGASUS, we will divide the document into unigram using SentencePiece. The tokenizer will also include an underscore in front of every token to represent spacing. For some word abbreviations, such as “didn’t”, they will be tokenized as “_didn”, “‘“, “t”. Note that there are no underscores prior to the last two tokens since there is no space in front of them from the original text token. Also, the tokenizer will convert each text token into two id tokens, including inputs_ids token (an unique id for each word with same spelling) and attention_mask token (used to identify the padded text token so that no attention will be applied to them).

```
Document:
  Ivan Rakitic hit the beach with his wife to
  prepare for a crucial week in Barcelona’s season
  as Luis Enrique’ side pursues a treble. The La
  Liga leaders face Manchester City in the Champions
  League last 16 second-leg, having won the first-
  leg 2-1 last month, before hosting Real Madrid in
  El Clasico on Sunday. Rakitic has become an
  integral part of the Barcelona midfield ranks
  since joining from Sevilla last summer, making 36
  appearances so far this season for the Catalan
  giants. Barcelona star Ivan Rakitic enjoys a sunny
  afternoon at the beach with his wife ahead of a
  crucial week . The Croatian international posted a
  sunny snap on his Instagram with his Spanish wife
  Raquel Mauri ahead of the Manchester City clash
  with the caption ‘familiar Sunday’. Seemingly
  Rakitic is a regular on Barcelona’s beach but it’s
  not all leisure for the Croatian, having posted
  another picture on his Instagram account on Sunday
  from the team gym as he prepares for a ‘big week’.
  Luis Enrique’s side remained top of La Liga on
  Saturday with a 2-0 victory at Eibar, courtesy of
  a brace from Lionel Messi. Enrique is hoping to
  lead Barcelona to a treble, having reached the
  Copa del Rey final, the Champions League knockout
  stages and are currently leading La Liga. The
  Croatian (left) celebrates after Lionel Messi
  (right) fires Barcelona to victory at Eibar on
  Saturday . VIDEO We were clinical - Enrique .
  Rakitic is all smiles ahead of a busy week, facing
  Manchester City and then Real Madrid in El Clasico
  .
Annotated Summary:
  Barcelona beat Eibar 2-0 at the weekend to remain
  top of La Liga by a point . Luis Enrique's side
  face Manchester City in the Champions League last
  16 on Wednesday before hosting Real Madrid in El
  Clasico on Sunday . Ivan Rakitic has become an
  integral part of the Barcelona midfield .
```

Sample Entry #3411 from CNN_DailyMail

Model Explanation

The goal of the model is to obtain a set of trained parameters in the PEGASUS network after the pre-training process. For any given downstream task, the fine-tuning process will carry on with a PEGASUS model initialized with these pre-trained weights.

During the pre-training process, data will be loaded into the PEGASUS to perform two pre-training objectives:

1) Gap Sentences Generation (GSG): For any given document in the training corpus, sentences will be masked randomly. The decoder is required to generate the masked sentence as predictions. The authors believe this pre-training objective would better align to the summarization downstream task, which also requires autoregressive text generation.
2) Mask Language Modeling: This training objective is similar to the one in BERT and it will be implemented in conjunction with the GSG.

After the pre-trained tasks, the model is now with the hidden state parameters. The next step is to fine-tune the model with the CNN/DailyMail dataset. The training dataset (the document and the highlight) is loaded to the PEGASUS model along with the annotated summary. After processing the input in the hidden layers, the predicted summary will be generated in the output layer with softmax activation. There are several approaches for the generation process, including:

- Greedy Search: During the summary generation process, the model will select the next word with the maximum probability. i.e. $w_t = \text{argmax } P(w_t \mid w_{t-1})$. The drawback is that the high probability words may be hidden after some lower probability word in the earlier position of the sequence, and these tokens will not be selected.
- Beam Search: This approach aims at resolving the drawback of greedy algorithms - In each position, this algorithm will keep n sequences with the highest likelihood and continue developing these n sequences in the next time stamp.
- Sampling: During the generation process, the next word will be randomly selected conditioned on the current word's softmax probability distribution.
- Top-k sampling: Similar to the sampling method, however, only the top k words with highest probability conditioned on the current word will be selected.
- Top-p sampling: It is also a variation of the sampling method. The smallest possible set of words with the cumulative softmax probability (conditioned on the current word) exceeds p will be used in the sampling process

Performance

Sample #	Index	Document	Annotated	Greedy	Beam	Sample	Top-K	Top-P	Top-P-K
1	4863	Wojciech Szczesny can regain his place as Arsenal number one if he dedicates himself completely to the task, according to former	Former Arsenal goa	Former Arsenal goalkeeper Bob Wilson says Wojciech Szczesny must work hard to regain his place in the team.	Former Arsenal goalkeeper Bob Wilson says Wojciech Szczesny must work hard to regain his place in the team.	Arsenal goalkeeper Wojciech Szczesny has been ousted by David Ospina but is likely to play on Monday night	Former Arsenal goalkeeper Bob Wilson says Wojciech Szczesny must work hard to regain his place in the team.	you need to dedicate yourself to being the best you can be, to the exclusion of a lot of other things in your life".	All images are copyrighted.
2	1728	Steve Clarke afforded himself a few smiles on the touchline and who could blame him? This has been a strange old season for Read	Reading are now 13	Reading manager Steve Clarke was furious with the referee after his side beat Brighton.	Reading manager Steve Clarke was furious with the referee after his side beat Brighton.	Reading manager Steve Clarke was furious with the referee after his side beat Brighton at the Madejski Stadium.	Brighton at the Madejski Stadium to end a run of five games without a win and their manager was furious with the decision to award the Seagulls a late penalty.	Reading beat Brighton at the Madejski Stadium despite being reduced to 10 men in the second half when Gareth McCleary was sent off.	Reading beat Brighton at the Madejski Stadium and their manager was furious with the decision to award the Seagulls a late penalty.
3	3411	Ivan Rakitic hit the beach with his wife to prepare for a crucial week in Barcelona's season as Luis Enrique' side pursues a tri	Barcelona beat Bil	the beach with his wife to prepare for a crucial week in Barcelona's season as Luis Enrique' side pursues a treble.	All images are copyrighted.	Barcelona midfielder Ivan Rakitic hit the beach with his wife Olga Maori ahead of a busy week.	All images are copyrighted.	midfielder Ivan Rakitic enjoyed a sunny afternoon at the beach with his Spanish wife Olga Maori ahead of a busy week.	midfielder Ivan Rakitic enjoyed a sunny afternoon at the beach with his Spanish wife Olga Maori ahead of a busy week.
4	6540	He was born a healthy little boy, with his whole life ahead of him. But at the age of just one, Zach Parnaby's bright future was	Zach Parnaby has m	A terminally ill toddler's parents have compiled a bucket list of things he must see and do in the time he has left.	A terminally ill toddler's parents have compiled a bucket list of things he must see and do in the time he has left.	County Durham have compiled a bucket list for their 20-month-old son, who is unlikely to live beyond his fourth birthday, of things he must see and do in the time he has left.	A family have compiled a bucket list of things they want to see and do in the time they have left with their terminally ill son.	The parents of a 20-month-old boy with a rare disease have compiled a bucket list of things he must see and do in the time he has left.	A couple have compiled a bucket list for their terminally ill son, of things he must see and do in the time he has left.
5	10395	Angelina Jolie enjoyed a fun night out with her daughters at the Nickelodeon Kids' Choice Awards on Saturday night - in first ap	Actress recently	enjoyed a fun night out with her daughters at the Nickelodeon Kids' Choice Awards on Saturday night - in first appearance since she revealed she had had mastectomy	All images are copyrighted.	All images are copyrighted.	It's been a busy few weeks for the Jolie-Pitt family.	All images are copyrighted.	All images are copyrighted.
6	5641	Grim details have emerged about how death row inmates are executed by firing squad at Indonesia's 'Death Island' of Nusa Kambangan	In January, dozen	recently executed prisoners were strapped to wooden planks with a type of rubber used to make the inside of car tyres.	All images are copyrighted.	All images are copyrighted.	recently executed prisoners were strapped to wooden planks with a type of rubber used to make the inside of car tyres.	All images are copyrighted.	All images are copyrighted.
7	11165	A federal judge has ordered the US military to release photographs of detainees being abused in Iraq and Afghanistan . U.S. Distr	Follows a long-run	A federal judge has ordered that more pictures of detainee abuse must be shared by the US military.	A federal judge has ordered that more pictures of detainee abuse must be shared by the US military.	Civil Liberties Union has been seeking to make them public in the name of holding government	A federal judge has ordered that more pictures of detainee abuse must be shared by the US military.	A federal judge has ordered that more pictures of detainee abuse must be shared by the US military.	A federal judge has ordered that more pictures of detainee abuse must be shared by the US military.
8	7568	To his adoring sons, Jonathan Weekes was the perfect dad - lavishing them with attention, treating them to trips away and spendin	Jonathan Weekes, 4	He was a devoted father who devoted his life to his two young sons.	He was a devoted father who devoted his life to his two young sons.	He was a devoted father who devoted his life to his two young sons.	of-two has been jailed for 12 years for downloading indecent images of children.	of-two was so obsessed with child pornography that he blamed his own sons for his crimes.	of-two was so obsessed with child pornography that he blamed his own sons for his crimes.
9	5678	As the classic Australian invention of the cask wine turns 50 this year, the best goon bags on offer are revealed.	As cask wine turns	Australian invention of the cask wine turns 50 this year, the best goon bags on offer are revealed.	Australian invention of the cask wine turns 50 this year, the best goon bags on offer are revealed.	Australian invention of the cask wine turns 50 this year, the best goon bags on offer are revealed.	Australian invention of the cask wine turns 50 this year, the best goon bags on offer are revealed.	Australian invention of the cask wine turns 50 this year, the best goon bags on offer are revealed.	Australian invention of the cask wine turns 50 this year, the best goon bags on offer are revealed.
10	6797	There wasn't much chance of any of the pooches at Crufts today being called a "dog"s dinner". The canines who had arrived at the	Hundreds of dogs h	All images are copyrighted.	All images are copyrighted.	All images are copyrighted.	All images are copyrighted.	All images are copyrighted.	All images are copyrighted.

10 sample predictions of the PEGASUS model

We found that varying strategies had different accuracies:

Model	Accuracy
Greedy	60%
Beam	30%
Sample	40%
Top-K	60%
Top-P	30%
Top-P-K	50%

To determine if the model summary was correct for a given search, we compared it to the given annotated summary. If it did not match the summary we continued to the document and manually determined if the given output summary was a decent summarization of the document. For this we looked to see if it contained relevant information from the document such that it was one of the main talking points. Analyzing Sample #6540, we found that all searches did a good job at summarizing the document, indicating that the terminally ill children's parents planned a bucket list to go through. For Sample #3411 we found that Beam and Top-K search failed to produce a good summary as they focused on the fact that the document contained an image and did not summarize the document. This is likely due to selecting the most probable sequence. Overall, different models have different results because of how the generation process differs.

Multi-Document Summarization

Multi-Document Summarization is a task where the NLP model generates a short and concise summary for several documents. The state of the art model is called Pyramid-based Masked Sentence Pre-training for Multi-document Summarization (PRIMERA). In this following section, we will experiment with this model pre-trained by the multi-x_science_sum dataset.

Data Preprocessing

To test the model, we will be using the multi-x_science_sum dataset. The documents are collected from arXiv.org and the target summaries are the abstract of these scientific papers.

Before feeding the example into PRIMERA, the documents under one sample will be tokenized. Each sentence will be padded with a pair of sentence tokens (<s>, </s>). Also, there will be a special token <doc-sep> added in between each document. During the pre-training process, each example (with multiple processed documents) along with the annotated summaries will be loaded into the PRIMERA model to obtain the pre-trained weights.

```
Documents:
  One of the key concepts in testing is that of adequate test sets. A test selection criterion decides
  which test sets are adequate. In this paper, a language schema for specifying a large class of test
  selection criteria is developed; the schema is based on two operations for building complex criteria
  from simple ones. Basic algebraic properties of the two operations are derived. In the second part
  of the paper, a simple language-an instance of the general schema-is studied in detail, with the
  goal of generating small adequate test sets automatically. It is shown that one version of the
  problem is intractable, while another is solvable by an efficient algorithm. An implementation of
  the algorithm is described.

  Testing has long been in need of mathematical underpinnings to explain its value as well as its
  limitations. This paper develops and applies a mathematical framework that 1) unifies previous work
  on the subject, 2) provides a mechanism for comparing the power of methods of testing programs based
  on the degree to which the methods approximate program verification, and 3) provides a reasonable
  and useful interpretation of the notion that successful tests increase one's confidence in the
  program's correctness.

Annotated Summary:
  Gourlay cite presents a precise framework for the discussion of issues in testing. In his
  terminology, our test selection criteria are a special form of the test methods for the set-choice
  construction testing system. Gourlay reinterprets previously published discussions about the
  suitability of various test selection criteria. In our approach, we do not attempt to decide a
  priori which criteria are sufficient --- we leave that decision to the test designer. That is why we
  emphasize the importance of a language in which criteria are specified.
```

Sample Entry #1 from CNN_DailyMail

Model Explanation

The goal of the model is to obtain a set of trained parameters in the PRIMERA network after the pre-training process. For any given downstream task, the PRIMERA model will be initialized with these weights to start the fine-tuning stage.

During the pre-training process, data will be loaded into the PRIMERA to perform Gap Sentences Generation (GSG). There is a subtle difference between the GSG in PEGASUS and PRIMERA: PRIMERA will utilize the algorithm called Entity Pyramid Masking to select the more representative sentences to mask. The term or token that contains the representative information is called Summary Content Unit (SCU). If a SCU appears in different sentences, SCU is more important. A representative sentence should always contain such SCUs and they should be masked to obtain better pre-training

3 Abstractive Summarization 30 / 30

✓ **- 0 pts** *Correct*

- 1 pts 3.1. Task

deduct because: no task detailed explanation

- 1 pts 3.2. {dataset}

deduct because: not showing dataset example

- 2 pts 3.3. {model} more detailed explanation of the model

deduct because: no detailed explanation of the model

- 2 pts 3.4.1 {performance} high-level performance results

deduct because: no high-level performance results

- 2 pts 3.4.2 {performance} low-level performance results

deduct because: no low-level performance results

- 5 pts Task empty

- 5 pts dataset empty

- 5 pts model empty

- 5 pts high level performance empty

- 5 pts low level performance empty

gradescope

We found that varying strategies had different accuracies:

Model	Accuracy
Greedy	60%
Beam	30%
Sample	40%
Top-K	60%
Top-P	30%
Top-P-K	50%

To determine if the model summary was correct for a given search, we compared it to the given annotated summary. If it did not match the summary we continued to the document and manually determined if the given output summary was a decent summarization of the document. For this we looked to see if it contained relevant information from the document such that it was one of the main talking points. Analyzing Sample #6540, we found that all searches did a good job at summarizing the document, indicating that the terminally ill children's parents planned a bucket list to go through. For Sample #3411 we found that Beam and Top-K search failed to produce a good summary as they focused on the fact that the document contained an image and did not summarize the document. This is likely due to selecting the most probable sequence. Overall, different models have different results because of how the generation process differs.

Multi-Document Summarization

Multi-Document Summarization is a task where the NLP model generates a short and concise summary for several documents. The state of the art model is called Pyramid-based Masked Sentence Pre-training for Multi-document Summarization (PRIMERA). In this following section, we will experiment with this model pre-trained by the multi-x_science_sum dataset.

Data Preprocessing

To test the model, we will be using the multi-x_science_sum dataset. The documents are collected from arXiv.org and the target summaries are the abstract of these scientific papers.

Before feeding the example into PRIMERA, the documents under one sample will be tokenized. Each sentence will be padded with a pair of sentence tokens (<s>, </s>). Also, there will be a special token <doc-sep> added in between each document. During the pre-training process, each example (with multiple processed documents) along with the annotated summaries will be loaded into the PRIMERA model to obtain the pre-trained weights.

```
Documents:
  One of the key concepts in testing is that of adequate test sets. A test selection criterion decides
  which test sets are adequate. In this paper, a language schema for specifying a large class of test
  selection criteria is developed; the schema is based on two operations for building complex criteria
  from simple ones. Basic algebraic properties of the two operations are derived. In the second part
  of the paper, a simple language-an instance of the general schema-is studied in detail, with the
  goal of generating small adequate test sets automatically. It is shown that one version of the
  problem is intractable, while another is solvable by an efficient algorithm. An implementation of
  the algorithm is described.

  Testing has long been in need of mathematical underpinnings to explain its value as well as its
  limitations. This paper develops and applies a mathematical framework that 1) unifies previous work
  on the subject, 2) provides a mechanism for comparing the power of methods of testing programs based
  on the degree to which the methods approximate program verification, and 3) provides a reasonable
  and useful interpretation of the notion that successful tests increase one's confidence in the
  program's correctness.

Annotated Summary:
  Gourlay cite presents a precise framework for the discussion of issues in testing. In his
  terminology, our test selection criteria are a special form of the test methods for the set-choice
  construction testing system. Gourlay reinterprets previously published discussions about the
  suitability of various test selection criteria. In our approach, we do not attempt to decide a
  priori which criteria are sufficient --- we leave that decision to the test designer. That is why we
  emphasize the importance of a language in which criteria are specified.
```

Sample Entry #1 from CNN_DailyMail

Model Explanation

The goal of the model is to obtain a set of trained parameters in the PRIMERA network after the pre-training process. For any given downstream task, the PRIMERA model will be initialized with these weights to start the fine-tuning stage.

During the pre-training process, data will be loaded into the PRIMERA to perform Gap Sentences Generation (GSG). There is a subtle difference between the GSG in PEGASUS and PRIMERA: PRIMERA will utilize the algorithm called Entity Pyramid Masking to select the more representative sentences to mask. The term or token that contains the representative information is called Summary Content Unit (SCU). If a SCU appears in different sentences, SCU is more important. A representative sentence should always contain such SCUs and they should be masked to obtain better pre-training

performance. However, these SCUs are normally annotated by humans. As a result, a larger scale pre-training to select the important sentence to mask is nearly impossible. To overcome the issue, named entities are used as proxies for the SCUs.

After the pre-trained tasks, the model is now with the trained parameters. The next step is to fine-tune the model with the dataset. After processing the input in the hidden layers, the predicted summary will be generated in the output layer with softmax activation. The approaches of autoregressive generation would be similar to the ones we introduced in the PEGASUS section.

Performance

Sample #	Index	Documents	Annotated	Greedy	Beam	Sample	Top-K	Top-P	Top-P-K
1	2270	Nonlinear opt	Another class	Guided Policy Search (GPS) cite	Guided Policy Search (GPS) cite	Guided policy searchful algorit	Our approach uses model-based R	There have been several approac	Guided Policy Search cite learn
2	1729	Online media	It was not u	The clickbait detection task ha	The application of neural netwo	The first approach to distingui	Several methods have been propo	Other work has focused on impro	Clickbait detection has been an
3	3411	Object detec	Given object	The cascade structure was first	Our work is also related to the	There are several other fundame	Methods to solve this problem a	For classification, a popular p	Object detection is a fundament
4	3312	Trying to co	To the best	The Cell processor, which is th	To the best of our knowledge, t	The era of modular architecture	In this paper we present our im	In the case of complementary pr	Hera-JVM cite is an implementat
5	837	During the l	In cite a co	The first approach to detecting	cite propose an approach to det	We rely heavily on the work of	More closely related to our wor	Making predictions based on a d	cite a framework is proposed to
6	1974	Understandin	There are se	A number of methods have been p	Gradient-based methods compute	A batch of studies cite cite ci	Previous works on generating he	Before the era of deep learning	A broad range of approaches hav
7	2501	Action recog	More recentl	In cite, the authors propose to	cite propose a function-based t	At the @math dissolution of the	In the second category cite pro	A lot of work has been done to	Prior to deep learning, several
8	4102	At ICSE'2013	In the resea	A taxonomy of the problem space	The problem space for self-heal	The question, What is a. Detail	Regarding questions of fairness	Many authors provide general de	There is a large body of work o
9	1234	Analyzing la	The works mo	Other approaches use a tree lay	Extensive research has been don	GTreePlus cite and GOTreePlusca	cite and cite propose interacti	Various approaches have been pr	A number of approaches have bee
10	2198	In the natur	Model compre	There is a large body of work o	Our work is also related to oth	" Compressing a Network Go To L	There is also much research per	More recently, Yang al cite emp	In order to reduce the amount o

10 sample predictions of the PRIMERA model

We found that varying strategies had different accuracies:

Model	Accuracy
Greedy	90%
Beam	90%
Sample	20%
Top-K	80%
Top-P	50%
Top-P-K	90%

To determine if the model summary was correct for a given search, we compared it to the given annotated summary. If it did not match the annotated summary we continued to the document and manually determined if the given output summary was a decent summarization of the document. As such we wanted to notice if key topics were discussed in the summary, or if a main point was summarized. Interestingly enough, Sample Search did not produce the most coherent summaries and this incoherence led to the predicted summary missing many main points. We noticed that the data in documents and the provided annotation summary often used the word “cite” to refer to a paper, and the models picked up on this. Sample #2198 allows each type of search to produce a consistently accurate summary which mentions the efforts to reduce the number of parameters in neural networks.

Contributions of Each Team Member

Kevin: Question-Answering (50%), Abstractive Summarization(50%), Multi-Document Summarization(50%), Report(50%)
Humza: Question-Answering (50%), Abstractive Summarization(50%), Multi-Document Summarization(50%), Report(50%)

4 Multi-Document Summarization **30 / 30**

✓ **- 0 pts** *Correct*

- 1 pts 4.1. Task

deduct because: no task detailed explanation

- 1 pts 4.2. {dataset}

deduct because: not showing dataset example

- 2 pts 4.3. {model} more detailed explanation of the model

deduct because: no detailed explanation of the model

- 2 pts 4.4.1 {performance} high-level performance results

deduct because: no high-level performance results

- 2 pts 4.4.2 {performance} low-level performance results

deduct because: no low-level performance results

- 5 pts task empty

- 5 pts dataset empty

- 5 pts model empty

- 5 pts high level performance empty

- 5 pts low level performance empty

gradescope