README

November 15, 2021

System Info: Tested on Ubuntu 20.04 or Ubuntu 18.04

1 Step 1: Set up environment

Note: The commands below in step 1 need to be run in a terminal. Then in the created anaconda environment, start jupyter notebook and open this README.ipynb, then the commands in the following steps can be directly run in this notebook

1.1 create an Anaconda environment, with a name e.g. memsum

Note: the symbol "!" is used to run command in jupyter notebook. In a terminal, "!" is not needed.

Note: Without further notification, the following commands need to be run in the working directory where this jupyter notebook is located.

[ ]:
```
!conda create -n memsum python=3.7
```

1.2 activate this environment

[ ]:
```
!source activate memsum
```

1.3 In the created anaconda environment, install jupyter notebook and ipython to run this code

[ ]:
```
!conda install ipython
!conda install -c anaconda jupyter
```

[ ]:
```

```

2 Setp 2: Install dependencies, download word embeddings and load them to pretrained model

2.1 Install dependencies via pip

[1]:
```
!pip install -r requirements.txt
```

2.2 Install pytorch (GPU version)

You need to specify your CUDA version correctly before installing pytorch. If CUDA version is 11.3 then we can use the following command to install pytorch

[2]:

```
!conda install pytorch cudatoolkit=11.3 -c pytorch -y
```

We provide a trained MemSum model on PubMed dataset. In order to use this model, we need to download the pretrained GLOVE word embedding from the official website and add them to MemSum using the following script.

This command takes time, as we need to first download and the unzip GloVe embeddings.

[3]:

```
!python download_and_load_word_embedding.py
```

```
--2021-11-15 03:30:18--  https://nlp.stanford.edu/data/glove.6B.zip
Resolving proxy.ethz.ch (proxy.ethz.ch)… 129.132.202.155
Connecting to proxy.ethz.ch (proxy.ethz.ch)|129.132.202.155|:3128… connected.
Proxy request sent, awaiting response… 301 Moved Permanently
Location: http://downloads.cs.stanford.edu/nlp/data/glove.6B.zip [following]
--2021-11-15 03:30:19--  http://downloads.cs.stanford.edu/nlp/data/glove.6B.zip
Connecting to proxy.ethz.ch (proxy.ethz.ch)|129.132.202.155|:3128… connected.
Proxy request sent, awaiting response… 200 OK
Length: 862182613 (822M) [application/zip]
Saving to: ‘glove.6B.zip’

glove.6B.zip        100%[===================>] 822.24M  5.10MB/s    in 2m 50s

2021-11-15 03:33:09 (4.84 MB/s) - ‘glove.6B.zip’ saved [862182613/862182613]

Archive:  glove.6B.zip
  inflating: glove.6B.50d.txt
  inflating: glove.6B.100d.txt
  inflating: glove.6B.200d.txt
  inflating: glove.6B.300d.txt
All model loaded!
```

3 Step 3: Testing trained model on a given dataset

For example, the following command test the performance of the full MemSum model, on the Pubmed's test set. The model is evaluated by ROUGE 1/2/L's precision, recall and F1 scores.

[4]:

```
!python my_test.py -model_type MemSum_Final -summarizer_model_path model/
 ↪MemSum_Final/pubmed_full/200dim/best/model.pt -vocabulary_path model/glove/
 ↪vocabulary_200dim.pkl -corpus_path data/pubmed_full/test_PUBMED.jsonl -gpu 0␣
 ↪-max_extracted_sentences_per_document 7 -p_stop_thres 0.6 -output_file␣
 ↪results/MemSum_Final/pubmed_full/200dim/test_results.txt  -max_doc_len 500␣
 ↪-max_seq_len 100
```

```
Start Computation …
100it [00:09, 10.68it/s]
p_stop_thres: 0.6000, avg. # sentences: 6.24 ± 1.10, avg. extraction time: 69.72
± 6.16 ms, R-1 (p,r,f1): 0.4878, 0.5433, 0.4969 R-2: 0.2296, 0.2500, 0.2323
R-L: 0.4418, 0.4907, 0.4493
```

Due to the size limit of the appendix, we only provided the trained MemSum on the PubMed dataset, and we only provide the first 100 training/validation/testing examples for the PubMed, arXiv and GovReport datasets.

In the future, we will release all the trained models and full datasets used in our experiments.

4 Step 4: Use the pretrained summarizer as a module

4.1 load the full MemSum model

[5]:

```python
from my_summarizers import ExtractiveSummarizer_MemSum_Final


memsum_model = ExtractiveSummarizer_MemSum_Final(
                "model/MemSum_Final/pubmed_full/200dim/best/model.pt",
                "model/glove/vocabulary_200dim.pkl",
                gpu = 0,
                embed_dim = 200,
                max_doc_len  = 500,
                max_seq_len = 100
)
```

4.2 Get a document to be summarized

The format of the document to be summarized is a list of sentences

[6]:

```python
import json

database = [ json.loads(line) for line in open( "data/pubmed_full/test_PUBMED.
 ↪jsonl" ).readlines() ]
pos = 6
document = database[pos]["text"]
gold_summary =  database[pos]["summary"]
```

[7]:

```python
print(document[:5])
```

```
['the family is the cornerstone of human social support network and its presence
is essential in everyone s life . changes inevitably occur in families with
illness and hospitalization of a family member . in other words , among the
sources of stress for families are accidents leading to hospitalization
particularly intensive care unit ( icu ) .', 'statistics show that 8% of
hospital beds in the united states are occupied by the intensive care units .',
'stress in the family while the patient is in the icu can disrupt the harmony
power of the family members and finally , it may causes disturbances in the
support of the patient .', 'in addition to the various sources of stress in
intensive care units such as the patient s fear of death , financial problems ,
lack of awareness about the environment and etc . , their satisfaction level is
another important source of stress for the patient s family .', 'today , the
family needs of hospitalized patients in the icu are summarized in five sections
.']
```

The gold summary is the abstract of the corresponding paper to which the document belongs. Here we get an example from the test set of the PubMed dataset

[8]:
```
print(gold_summary)
```

```
['background : since the family is a social system , the impairment in each of
its component members may disrupt the entire family system .', 'one of the
stress sources for families is accidents leading to hospitalization particularly
in the intensive care unit ( icu ) . in many cases ,', 'the families needs in
patient care are not met that cause dissatisfaction . since the nurses spend a
lot of time with patients and their families , they are in a good position to
assess their needs and perform appropriate interventions .', 'therefore , this
study was conducted to determine the effectiveness of nursing interventions
based on family needs on family satisfaction level of hospitalized patients in
the neurosurgery icu.materials and methods : this clinical trial was conducted
in the neurosurgery icu of al - zahra hospital , isfahan , iran in 2010 .',
'sixty four families were selected by simple sampling method and were randomly
placed in two groups ( test and control ) using envelopes . in the test group
,', 'some interventions were performed to meet their needs . in the control
group ,', 'the routine actions were only carried out .', 'the satisfaction
questionnaire was completed by both groups two days after admission and again on
the fourth day.findings:both of the intervention and control groups were
compared in terms of the mean satisfaction scores before and after intervention
.', 'there was no significant difference in mean satisfaction scores between
test and control groups before the intervention .', 'the mean satisfaction score
significantly increased after the intervention compared to the control
group.conclusions:nursing interventions based on family needs of hospitalized
patients in the icu increase their satisfaction .', 'attention to family nursing
should be planned especially in the icus .']
```

4.3 Extractively summarize the document using MemSum

[9]:
```
extracted_summary = memsum_model.extract( [ document ], p_stop_thres=0.6,␣
 ↪max_extracted_sentences_per_document= 7, return_sentence_position= False )[0]
```

[10]:
```
print(extracted_summary)
```

```
['the purpose of the study was to determine the effectiveness of nursing
interventions based on family needs on family satisfaction level of hospitalized
patients in the neurosurgery intensive care unit of al - zahra hospital in 2010
.', 'in this study , it was shown that the use of nursing interventions based on
family needs ( confidence , support , information , proximity and convenience )
had significant impact on the family satisfaction of the patient hospitalized in
intensive care unit .', 'the statistical research community was the families of
hospitalized patients in neurosurgery intensive care unit of al - zahra ( sa )
hospital , isfahan , iran from may to september 2010 .', 'the aim of this study
was to analyze the satisfaction of the families of icu patients .', 'the results
of the present study showed that the nursing interventions based on the family
needs increased the patient s family satisfaction in the neurosurgery intensive
care unit of al - zahra hospital .', 'comparison of mean satisfaction score (
100 * ) of participants in the intervention and control groups the mean of
satisfaction score changes of the studied subjects in the intervention and
control groups after intervention', 'the mean satisfaction score in the
intervention group after the intervention was significantly higher than before
the intervention ( p < 0.001 ) .']
```

4.4 Evaluate the extracted summary via ROUGE scores

[11]:
```
from rouge_score import rouge_scorer
rouge_cal = rouge_scorer.RougeScorer(['rouge1','rouge2', 'rougeLsum'],␣
 ↪use_stemmer=True)
print(rouge_cal.score( "\n".join(gold_summary), "\n".join(extracted_summary)  ))
```

```
{'rouge1': Score(precision=0.6517412935323383, recall=0.4833948339483395,
fmeasure=0.5550847457627119), 'rouge2': Score(precision=0.36,
recall=0.26666666666666666, fmeasure=0.30638297872340425), 'rougeLsum':
Score(precision=0.6069651741293532, recall=0.45018450184501846,
fmeasure=0.5169491525423728)}
```

5 Step 5: Training model from script

For example, if we want to train the full MemSum model on the PubMed dataset, we first change working directory to "src/MemSum_Final/", then run the python script "train.py". The train.py takes one parameter: config_file_path, which is the path to the training configuration file.

In the configuration file there are detailed list of key-value pairs that configure the training procedure. For example, the number of GPU devices, batch size, learning rate, etc.

[ ]:
```
!cd src/MemSum_Final/; python train.py -config_file_path config/pubmed_full/
 ↪200dim/run0/training.config
```

[ ]:
```

```