**MatchSum源代码使用手册**

目录

[目录结构 1](#_Toc83476330)

[环境配置 2](#_Toc83476331)

[重要代码说明 3](#_Toc83476332)

[数据读取相关类 3](#_Toc83476333)

[模型实现相关类 4](#_Toc83476334)

[模型训练部分 5](#_Toc83476335)

[客户端服务器架构部分 5](#_Toc83476336)

[输入与输出 7](#_Toc83476337)

[输入数据格式 7](#_Toc83476338)

[输入数据案例 7](#_Toc83476339)

[输出数据格式 8](#_Toc83476340)

[输出数据案例 8](#_Toc83476341)

# 目录结构

│ End2End\_Client.py

│ End2End\_Server.py

│ MatchSum源代码使用手册.docx

│ model.py

│

├─data

│ test\_CNNDM\_bert.jsonl

│ test\_CNNDM\_roberta.jsonl

│ train\_CNNDM\_bert.jsonl

│ train\_CNNDM\_roberta.jsonl

│ val\_CNNDM\_bert.jsonl

│ val\_CNNDM\_roberta.jsonl

│

├─Demo

│ input.summary

│ input.txt

│

├─MatchSum\_cnndm\_model

│ BertSum-Parameter.pkl

│ MatchSum\_cnndm\_roberta.ckpt

│

├─Request

│

├─Result

│

├─Train

BertSumExperiment.py

callback.py

dataloader.py

MatchSumExperiment.py

metrics.py

tools.py

utils.py

1. Data文件夹用来存放数据集。
2. MatchSum\_cnndm\_model文件夹存放已经训练完毕的神经网络参数。
3. Request文件夹存放Client提交以及Server处理的请求。
4. Result文件夹存放Client下载以及Server处理完毕的结果。
5. Train文件夹为模型进行训练的主程序

# 环境配置

1. python==3.6
2. cuda 10.2（注：请结合实际显卡驱动以及程序进行对应版本的安装）
3. python依赖包：

os

sys

time

pyrouge

tempfile

datetime

subprocess

rouge\_score

json==2.0.9

rouge==1.0.0

tqdm==4.57.0

argparse==1.1

numpy==1.19.5

paramiko==2.7.2

logging==0.5.1.2

fastNLP==0.6.0

torch==1.9.0+cu102

transformers==2.10.0

# 重要代码说明

## 数据读取相关类

该部分主要介绍读取数据时的类

MatchSumLoader

所属文件：Train/dataloader.py

主要功能：读取放在data文件夹下的文件。

\_\_init\_\_ 构造函数

参数列表：

candidate\_num：提取的前candidate\_num数量的候选集（int）

encoder：所选择的encoder类型（”bert” or “roberta”）

max\_len：截取的最大长度（int）

load 读取函数

功能：从文件之中进行读取。

参数列表：

paths：读取文件的路径（str）

MatchSumPipe

所属文件：Train/dataloader.py

主要功能：读取放在data文件夹下的文件。

\_\_init\_\_ 构造函数

参数列表：

candidate\_num：提取的前candidate\_num数量的候选集（int）

encoder：所选择的encoder类型（”bert” or “roberta”）

process\_from\_file 处理函数

功能：调用MatchSumLoader对文件进行处理。

参数列表：

paths：读取文件的路径（str）

## 模型实现相关类

该部分用来介绍项目之中实现模型的主要类

MatchSum

所属文件：model.py

主要功能：MatchSum结构的实现，计算输入的text, candidate, summary之间的相似性程度。

\_\_init\_\_ 构造函数

参数列表：

candidate\_num：输入的摘要候选集candidate的数量（int）

encoder：使用的预训练模型的名字（”bert” or “roberta”）

hidden\_size：预训练模型所产生的向量尺寸大小（int）

forward 前向运算函数

功能：计算输入的text, candidate, summary之间的相似性程度。

参数列表：

text\_id：输入的原始文本id（torch.LongTensor）

candidate\_id：输入的摘要候选集id（torch.LongTensor）

summary\_id：输入的摘要标签id（torch.LongTensor）

返回值：

{‘score’:score, ‘summary\_score’: summary\_score}：

score：计算出的摘要候选集与输入文本之间的相似性程度结果（float）

summary\_score：计算出的摘要标签与输入文本之间的相似性程度结果（float）

BertSum

所属文件：model.py

主要功能：BertSum结构的实现，计算输入句子之中信息量程度的大小。

\_\_init\_\_ 构造函数

参数列表：

model\_name：使用的预训练模型的名字（”bert-base-uncased” or “bert-large-uncased”）

forward 前向运算函数

功能：计算输入句子之中信息量程度的大小，来决定是否保留所对应的句子。

参数列表：

text\_id：输入的原始文本id（torch.LongTensor）

text\_position：输入的原始文本下每一个句子的起始位置（list）

label：输入的是否保留该句的标签（torch.LongTensor Or None）

返回值：

predict：若label为None，计算出每一句进行保留的概率（torch.FloatTensor）

loss：若label不为None，计算出预测结果与实际结果之间的交叉熵。

## 模型训练部分

该部分用来介绍模型训练部分的相关代码

MatchSumExperiment

所属文件：Train/MatchSumExperiment.py

主要功能：进行MatchSum模型的训练。

参数列表：

mode：进行处理的模式（'train’ or ‘test’）

save\_path：保存训练文件结果的路径（str）

gpus：训练所使用的gpu编号（str）

encoder：进行微调使用的预训练模型（'bert’ or ‘roberta’）

batch\_size：一次进行处理的batch大小（int）

accum\_count：经过多少个batch进行一次update（int）

candidate\_num：进行比较的candidate的数量（int）

max\_lr：Warm Up过程中的最大学习率（float）

margin：MarginRankingLoss的参数（float）

warmup\_steps：进行Warm Up过程的步数（int）

n\_epochs：进行训练的epoch的数量（int）

valid\_steps：进行多少次update之后进行验证（int）

BertSumExperiment

所属文件：Train/BertSumExperiment.py

主要功能：进行BertSum模型的训练。

参数列表：

save\_path：保存训练文件结果的路径（str）

gpus：训练所使用的gpu编号（str）

encoder\_name：进行微调使用的预训练模型（str）

batch\_size：一次进行处理的batch大小（int）

learning\_rate：学习率的大小（float）

epoch\_number：进行训练的epoch的数量（int）

## 客户端服务器架构部分

MatchSumServer

所属文件：End2End\_Server.py

主要功能：客户端服务器架构的服务器部分，进行端对端的摘要任务的处理：对给出的请求进行摘要任务的处理，并保存结果。

\_\_init\_\_ 构造函数

参数列表：

bertsum\_path：训练完毕的BertSum模型的路径（str）

matchsum\_path：训练完毕的MatchSum模型的路径（str）

gpu\_used：进行处理所使用GPU的编号（str）

reveal\_top\_n：最后显示的Top-N个摘要结果（int）

bertsum\_select\_number：BertSum模型处理之中所选择的句子数量（int）

batch\_size：MatchSum之中一次处理的Candidate的数量（int）

summarize\_loop

功能：接受请求（Request），并对请求进行处理。

summarize\_bertsum

功能：使用BertSum模型来计算出句子的重要性程度，并给出要保留句子的编号。

参数列表：

text：输入进行摘要的句子（[str, str, …, str]）

返回值：

selected\_sentence\_id：计算出需要保留的N个句子的编号（list）

summarize\_matchsum

功能：使用MatchSum模型计算出各个候选集（Candidate）与文本（Text）的相似性程度。

参数列表：

text：输入进行摘要的句子（[str, str, …, str]）

candidate：输入作为摘要候选集的段落（[str, str, …, str]）

返回值：

total\_score：计算出的摘要候选集与文本的相似性程度（list）

MatchSumClient

所属文件：End2End\_Client.py

主要功能：客户端服务器架构的客户端部分，将本地文件打包为请求发送到服务器上。待服务器处理完毕之后将摘要结果下载，并展示结果。

\_\_init\_\_ 构造函数

参数列表：

ip：所连接服务器的ip地址（str）

username：所连接服务器的用户名（str）

password：所连接服务器的密码（str）

port：所连接服务器的端口号（str）

request\_path：所连接服务器上保存请求的路径（str）

result\_path：所连接服务器上保存结果的路径（str）

\_\_search\_file\_\_

功能：搜索服务器上结果文件夹中是否产生结果。

参数列表：

filename：进行搜索的文件名（str）

package\_request

功能：将输入文件以及摘要标签打包为请求。

参数列表：

file\_name：进行摘要的文件路径（str）

summary\_name：作为结果的摘要标签路径（str or None）

返回值：

packaged\_request：打包完毕的结果（{‘Time’, ‘Text’, ‘Summary’}）

upload\_file

功能：将处于请求提交文件夹的请求上传到服务器，并在服务器处理完毕之后进行下载再显示。

# 输入与输出

## 输入数据格式

输入MatchSumClient. package\_request的文件格式应当为：

Sentence1

Sentence2

…

SentenceN

经过请求封装之后的文件格式应当为：

{

“Time”: 时间戳,

“Text”: [文本第一句, 文本第二句, …, 文本第N句],

“Summary”: [摘要第一句, 摘要第二句, …, 摘要第M句]

}

## 输入数据案例

以下为对Demo文件夹下进行处理，封装之后的Request内容：

{

"Time": "2021\_09\_25\_13\_13\_19",

"Text": [

"turkey has blocked access to twitter and youtube after they refused a request to remove pictures of a prosecutor held during an armed siege last week .",

"a turkish court imposed the blocks because images of the deadly siege were being shared on social media and ` deeply upset ' the wife and children of mehmet selim kiraz , the hostage who was killed .",

…,

"figures provided by twitter revealed that turkey filed more requests to remove content from the social network than any other nation between july and december 2014 ."

],

"Summary": [

"turkish court imposed blocks as images of siege shared on social media .",

"images ` deeply upset ' wife and children of hostage mehmet selim kiraz .",

"prosecutor , 46 , died in hospital after hostages stormed a courthouse .",

"two of his captors were killed when security forces took back the building ."

]

}

## 输出数据格式

Server对请求进行处理之后，所输出的格式为：

{

“Time”: 时间戳,

“BertSumSelect”: BertSum所选择的句子编号,

“Score”: MatchSum计算出各个候选集与文本的相似性程度,

“Top-1”: Top-1的摘要结果,

“Top-2”: Top-2的摘要结果,

…

“Top-N”: Top-N的摘要结果,

}

## 输出数据案例

以下为对Demo文件夹下Request进行处理，并返回的结果内容：

{

"Time": "2021\_09\_25\_13\_12\_58",

"BertSumSelect": [0, 1, 2, 3, 4, 11, 18],

"Scores": [

0.6858487129211426,

0.7707276940345764,

…,

0.5284197330474854,

0.40226835012435913

],

"Top-1": "turkey has blocked access to twitter and youtube after they refused a request to remove pictures of a prosecutor held during an armed siege last week . a turkish court imposed the blocks because images of the deadly siege were being shared on social media and ` deeply upset ' the wife and children of mehmet selim kiraz , the hostage who was killed . the 46-year-old turkish prosecutor died in hospital when members of the revolutionary people 's liberation party-front ( dhkp-c ) stormed a courthouse and took him hostage .",

"Top-2": "a turkish court imposed the blocks because images of the deadly siege were being shared on social media and ` deeply upset ' the wife and children of mehmet selim kiraz , the hostage who was killed . the 46-year-old turkish prosecutor died in hospital when members of the revolutionary people 's liberation party-front ( dhkp-c ) stormed a courthouse and took him hostage . a turkish court has blocked access to twitter and youtube after they refused a request to remove pictures of prosecutor mehmet selim kiraz held during an armed siege last week",

"Top-3": "the 46-year-old turkish prosecutor died in hospital when members of the revolutionary people 's liberation party-front ( dhkp-c ) stormed a courthouse and took him hostage . the dhkp-c is considered a terrorist group by turkey , the european union and us . a turkish court has blocked access to twitter and youtube after they refused a request to remove pictures of prosecutor mehmet selim kiraz held during an armed siege last week"

}