Build an end-to-end QA system in machine learning approach

Wenxin Han
8701130
whan035@uottawa.ca

CSI 5180 Topics in AI: Virtual Assistants

Project Summary

- Build a QA system to help users to extract information from a long news article
- Easy for users to find answers
- Implement a extracting QA system in machine learning models

Resources

- Microsoft newsQA dataset: contain question-answer pairs from news article
- Transformer Library: provide thousands of pre-trained models and multiple datasets
- Pycharm: Python IDE, easy for understanding
- GitHub: the project code repository

Methodology

- Data cleaning
  - Extend character ranges to have complete words
  - Select a single answer for a question
- Baseline model (Cosine Similarity)
  - Compare cosine similarity in each sentence
  - Each sentence is tokenized and any stop words or URLs are removed
  - Taking average the Glove embedding to calculate an embedding for each sentence

Methodology

- Advanced model (Machine learning approach)
  - The data require pre-processing to store details
  - Use BERT model for training
  - Only linear layers were fine-tuning because the BERT model is huge
- Evaluation Metric
  - F1 score: harmonic mean of precision and recall
    - Calculate the number of overlapping characters between actual and predicted answers
  - Accuracy: the percentage of answers that are correctly predicted
    - At least one token overlapped, which is seen as correct

Activity table

- Processing data in right format spent more time
- Training models was time consuming

Activity	Why	Time Planned	Time Taken	Deliverable
Find related work	Find multiple models and make comparison	3h	3h	
Explore dataset	For know the format and find ways to testing	2h	3h	Choose newsQA dataset
Project environment setup	To create environment to implement	3h	3h	
Training baseline model	To get the performance results on baseline model	4h	4h	Using cosine similarity approach
Training advanced model	To get the performance results on advanced model	3h	7h	The Bert model is huge
Evaluation between two approaches	Get results for comparison	2h	3h	
Analyze the results	Analyze the results	6h	6h	Make the result tables
Writing report	Make the presentation video and write report	7h	7h	
	Total	30h	36h	

Results I

- Baseline model
  - F1 Score is extremely low because the actual answer does not span entire sentence
  - Predict the sentence with correct answer in 12.7% of the time

Cosine Similarity	F1 Score	Accuracy
	0.0501	0.127

Results II

- Advanced model
  - BERT model has better F1 Score and higher accuracy
  - After fine-tuning model has better performance
  - BERT model can predict correct answers in more than 50% of the time

BERT model	F1 Score	Accuracy
Before fine-tuning	0.2750	0.425
After fine-tuning	0.345	0.536

Challenges

- Training models are time-consuming
- Data cleaning is required, otherwise it is hard to implement
- In processing data, the data exceeds the BERT maximum length => find the parts of text has answers

What have you learned ?

- How to build an end-to-end QA system

- Learn Transformer Library in NLP
  - Transformer library has many pretrained models to perform tasks, such as text classification, image detection and speech recognition

- Learn fine-tuning in NLP
  - Re-train a pre-trained language model
  - The weights of the original model are updates and results may perform well

Conclusion

- How to build an end-to-end QA system
- How to compare models in two approaches to train based on the same dataset
- Machine learning approach in QA system
- Fine-tuning of models could achieve better performance

References

1. Microsoft's NewsQA dataset: https://www.microsoft.com/en-us/research/project/newsqa-dataset/

2. BERT model for QA: https://huggingface.co/docs/transformers/model_doc/bert?highlight=bertforquestionanswering#transformers.BertForQuestionAnswering

3. Data processing:https://github.com/smitkiri/news-qa

4. My GitHub Link: https://github.com/han0807/CSI-5180-Project