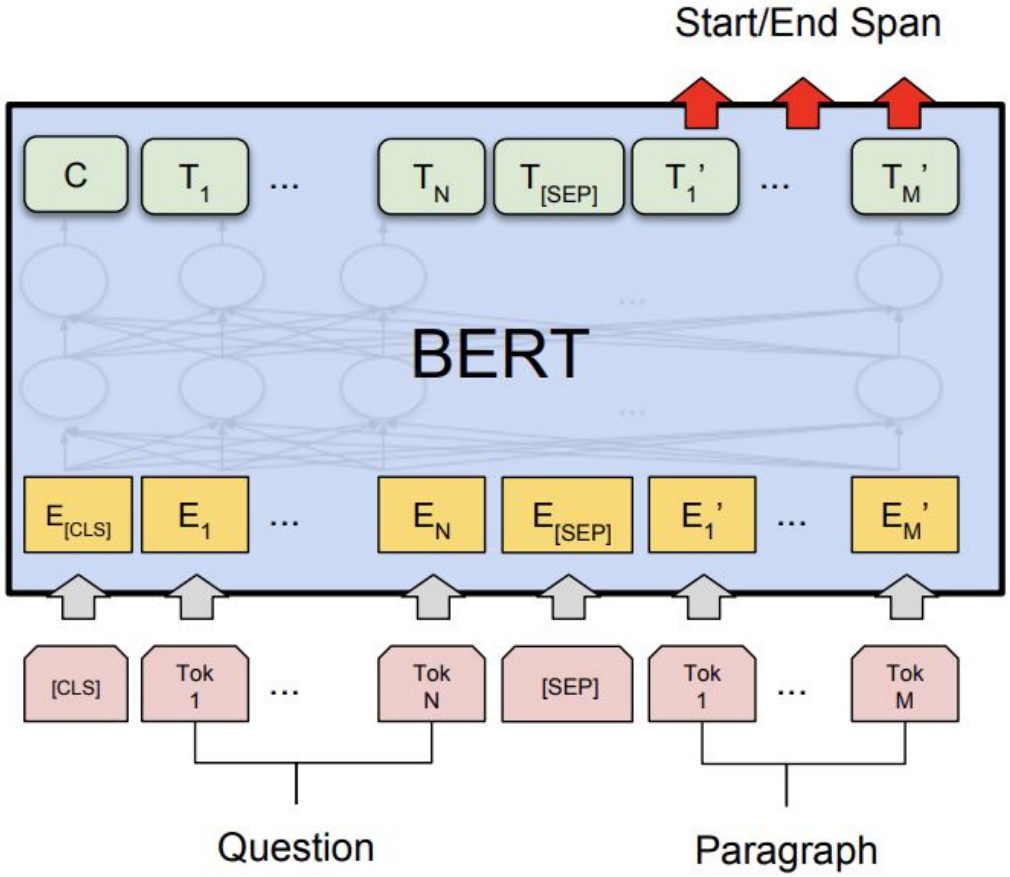
Chinese Question-Answering with BERT

EECS 496-7 Final Presentation

Edmond Chen, Michael Chen & Sebastian Pretzer

Github repo: https://github.com/edmondchensj/ChineseQA-with-BERT

Northwestern | McCORMICK SCHOOL OF ENGINEERING


Start/End Span
C
T1
...
TN
T[SEP]
T1'
...
TM'
BERT
E[CLS]
E1
...
EN
E[SEP]
E1'
...
EM'
[CLS]
Tok 1
...
Tok N
[SEP]
Tok 1
...
Tok M
Question
Paragraph

Data


```
{
  "question_id": 186358,
  "question_type": "YES_NO",
  "question": "上海迪士尼可以带吃的进去吗",
  "documents": [
    {
      'paragraphs': ["text paragraph 1", "text paragraph 2"]
    },
    ...
  ],
  "answers": [
    "完全密封的可以，其它不可以。",                                    // answer1
    "可以的，不限制的。只要不是易燃易爆的危险物品，一般都可以带进去的。",  //answer2
    "罐装婴儿食品、包装完好的果汁、水等饮料及包装完好的食物都可以带进乐园，但游客自己在家制作
  ],
  "yesno_answers": [
    "Depends",                    // corresponding to answer 1
    "Yes",                        // corresponding to answer 2
    "Depends"                     // corresponding to asnwer 3
  ]
}
```

Target Generation

Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an Ar record producer and actress. Born and raised in Houston, Texas, singing and dancing competitions as a child, and rose to fame in the of R&B girl-group Destiny's Child. Managed by her father, Mathew K one of the world's best-selling girl groups of all time. Their hiatus sa debut album, Dangerously in Love (2003), which established her as a so five Grammy Awards and featured the Billboard Hot 100 number-one s "Baby Boy".
Q: "In what city and state did Beyoncé grow up?" A: "Houston, Texas"
Q: "What areas did Beyoncé compete in when she was growing up?" A: "singing and dancing"
Q: "When did Beyoncé release Dangerously in Love?" A: "2003"

Figure (above): Sample QA from SQuAD dataset. Recall that reading comprehension models (e.g. BERT) relies on answer spans in the input paragraph as the target.

Problem with original dataset:
Answers are human-generated, **not** answer spans within input paragraph.

Solution:
Generate candidate answer spans based on the paragraph substring with maximum F1-score of real answers.

BERT Chinese Model

- 12-layer, 768-hidden, 12-heads, 110M parameters
- Tested to be 3% more accurate than the multilingual model on XNLI inference tasks
- Pre-training data: full Wikipedia dump for Chinese
- Character-tokenized instead of WordPiece

BLEU Scoring

- Compares n-grams between predicted text and reference texts
- Limits the $\text{gram}_{\text{pred}}$ count to the max count of a gram_{ref}
- DuReader uses BLEU-4

Pred	the	the	the	cat		
Ref	the	cat	sits	on	a	mat

Precision = 4/4 = 1
BLEU-1 = 2/4 = .5

Experiment Results

All models are fine-tuned on bert-base-chinese, with parameters: learning rate 3e-5, max sequence length 384, document stride 128, training epochs 2

Fine-tuning Process	BLEU Score
10000 training examples; effective batch size 4	7.44
10000 training examples; effective batch size 8	7.41
10000 training examples; effective batch size 12	7.37
20000 training examples; effective batch size 4	7.31
20000 training examples; effective batch size 12	7.37

Analysis of Results

- Relatively poor evaluation results:
  - Small training set size: 20000 (~20 hour training time on GTX1080)
  - 10% of DuReader dataset (Search + Zhidao)
- Increasing batch size does not help, likely due to the fewer update steps on a small training set
- BERT does not seem to perform very well with small amounts of fine-tuning data on this task

Future Work

- Analyze feature extraction/tokenization code and improve its efficiency to allow for faster training on large datasets
- Larger training set and different hyperparameters
- Improve target generation process for training

Thank You!

Northwestern | McCORMICK SCHOOL OF ENGINEERING