QUESTION ANSWERING

QANet on SQuAD 2.0

INTRODUCTION

- We worked on the task of extractive question answering using SQuAD 2.0 which adds unanswerable questions to about half of the dataset.
- QANet is an efficient architecture built for SQuAD 1.1 that uses an Encoder Block module to replace RNNs.
- The Encoder Block relies on a Transformer-like component which uses Positional Encodings, Self Attention, and Position-wise Feedforward Networks
- We experimented with Data Augmentation through back-translation, local self-attention, and augmented loss function, and adding tag features.

PROBLEM

- Given a context paragraph C, and a question Q, provide the token span C[i:j] which answers the question or state it's unanswerable.
- There are two metrics reported for this task:
  - Exact Match (EM) measures the percentage of questions that received the *exact* correct answer span prediction
  - F1 gives a averaged measure of overlap between the predicted and ground truth answer spans. To calculate this both are treated as a bag of tokens.

DATASET

Figure 1. SQuAD 2.0 Question Breakdown

- Our course's SQuAD 2.0 dataset contains 129,941 samples.
- The dev and test set contain 6078 and 5915 samples, respectively with roughly equal % of unanswerable/answerable questions.
- The train set has roughly twice as many answerable questions as unanswerable.
- Majority of questions begin with "What."

For many years, Sudan had an Islamist regime under the leadership of Hassan al-Turabi. His National Islamic Front first gained influence when strongman General Gaafar al-Nimeiry invited members to serve in his government in 1979. Turabi built a powerful economic base with money from foreign Islamist banking systems, especially those linked with Saudi Arabia. He also recruited and built a cadre of influential loyalists by placing sympathetic students in the university and military academy while serving as minister of education.

How did Turabi build a strong economic base?

True answer: *money from foreign Islamist banking systems*

Predicted answer: *with money from foreign Islamist banking systems*

Augmented Data by Back-translating Questions

Original Question	Paraphrase
When did Beyonce become popular?	When did Beyonce start becoming popular ?
What did the officials order the Chinese media to stop broadcasting?	What did officials order Chinese news media to stop reporting ?
What can be used to shoot without manually targeting enemies?	What can be used to shoot without the need to manually target enemies ?

MODEL ARCHITECTURE

START INDEX   END INDEX

Figure 2. Model Diagram

Model Details:

- A layer dropout was added between each sublayer
- TAG Features: POS, dependency, entity
- Batch size: 32
- Hidden size: 128
- MAX context length(train): 400
- Learning rate: 1e-3
- Dropout prob: 0.1
- Optimizer: Adam
- $\beta_1 = 0.8$
- $\beta_2 = 0.999$

Hyperparameter Tuning

Figure 3. Dev F1 and NLL plots on the well-tuned baseline with char-CNN embeddings using different hyperparameters

- Dropout of 0.1 performed the best on the baseline as well as QANet
- The loss went further down after adding weight decay of 1e-2
- QANet performed better with a hidden size of 128 vs 96

Self-Attention Modifications

a. Naive local attention: Splitting original sequence into K sub-sequences

b. Offset Attention: Splitting original sequence into K sub-sequences and offsetting in the next layer

c. Offset attention + Query representation

Query ->

ABLATION

Component	F1	EM
Full Model	67.15	63.65
Aug. Loss	67.39	63.91
Tag features	65.79	62.22
Char Embed	58.1	56.27
Data Augmentation	67.72	64.37

- We observe that char embeddings significantly impact performance and that the tag features provide marginal gains
- It was not immediately clear to us how much the some of the other components helped or hurt

RESULTS

Figure 4. Dev F1 and Loss plots of improvements over baseline

Our best model achieved a F1 score of **0.68 on dev** and **0.64 on test** set. It achieved an EM score of 0.62 and 0.60 on dev and test set respectively.

In World War II, Charles de Gaulle and the Free French used the overseas colonies as bases from which they fought to liberate France. However after 1945 anti-colonial movements began to challenge the Empire. France fought and lost a bitter war in Vietnam in the 1950s. Whereas they won the war in Algeria, the French leader at the time, Charles de Gaulle, decided to grant Algeria independence anyway in 1962. Its settlers and many local supporters relocated to France. Nearly all of France's colonies gained independence by 1960, but France retained great financial and diplomatic influence. It has repeatedly sent troops to assist its former colonies in Africa in suppressing insurrections and coups d'état.

After 1945, what challenged the British empire?

True Answer: *N/A*

Predicted Answer: *N/A*

The efficiency of a Rankine cycle is usually limited by the working fluid. Without the pressure reaching supercritical levels for the working fluid, the temperature range the cycle can operate over is quite small; in steam turbines, turbine entry temperatures are typically 565 °C (the creep limit of stainless steel) and condenser temperatures are around 30 °C. This gives a theoretical Carnot efficiency of about 63% compared with an actual efficiency of 42% for a modern coal-fired power station. This low turbine entry temperature (compared with a gas turbine) is why the Rankine cycle is often used as a bottoming cycle in combined-cycle gas turbine power stations.[citation needed]

What is the turbine entry temperature of a Rankine turbine, in degrees Celsius?

True answer: *N/A*   Predicted answer: *565 °C*

ANALYSIS

Figure 5. Bidirectional attention visualization using similarity matrix between question and context.

Figure 6. Model metrics(dev) on different question types.

Figure 7. Model metrics(dev) by context length groups.

CONCLUSION

- Our QANet implementation reached an F1 of 63.841 and EM of 60.389 on the test leaderboard and improved over a well-tuned baseline.
- Model requires more reasoning capabilities to understand unanswerable questions and perform better on "why" questions
- Larger context length is not detrimental in this context(max 400 tokens), thought extreme lengths might be
- In practice we did not find this model more efficient than the BiDAF baseline

FUTURE WORK

- We want to further explore local attention on a dataset with larger context lengths
- We would like to dig into why there was a large gap between our dev and test results
- It'd be interesting to incorporate techniques such as a span reader and pointer networks and try adding an RNN before or after the encoder blocks as this might help with transformers

REFERENCES

- Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V.Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Neural Information Processing Systems (NIPS), 2017

ANUPRIT KALE anuprit@stanford.edu EDGAR VELASCO edgarv1@stanford.edu