QANet for Reading Comprehension

Reetika Agrawal (reetika), Rohit Kulkarni (rohitnk), Ravi Rajagopal (rravi77)

SQuAD**2.0**

Stanford Question Answering Dataset

Problem

- The goal is to **develop a Question-Answering Model**, which takes a Question and Paragraph as Inputs, and attempts to answer the question as correctly as possible - providing a measure for how well the model can understand "text".
- The **baseline** Model is based on BiDirectional Attention Flow **(BiDAF) Architecture**.
- We implemented **QANet Architecture**, which uses Convolution and Self-Attention to replace the Sequential Recurrent Networks from the baseline Model.

Data

- Stanford Question Answering Dataset (SQuAD) v2.0
- Around 150K Questions.
- More than half the questions can't be answered using the paragraph.
- Data Split into: ~ 90.6% Train, 4.3% Dev, 4.2% Test.

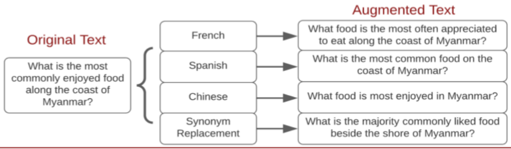
Methods

Character Embedding: Word-level embeddings do not address morphemes, misspelled or out-of-vocabulary words. We add character-level embedding to enhance input representation.

Token Features : Factual Q&As benefit from input features such as Part-of-Speech, Named-Entity Recognition, and Frequency.

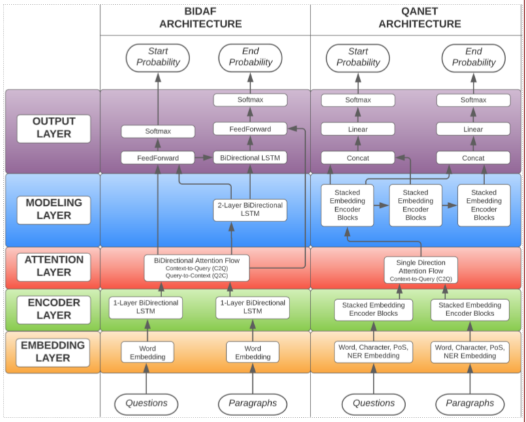
Data Augmentation: Techniques used were:

- Back-Translation using different Languages (French, Chinese, Spanish, Hindi) - to rephrase the text and introduce diversity.
- Synonym Replacement - to introduce new vocabulary into the text.
- Basic sanity checks were added to validate the augmented text: such as detecting changes in the Named Entities.



Approach

Implemented QANet, a transformer-like model which has higher speed and accuracy over BiDAF.



Output Layer was further enhanced with conditioning the p_{end} token on p_{start}. Two Methods were tested:

$$MethodA: X_{p1} = p_{start} \cdot ([M_0, M_1])$$
$$X_{p2} = [W_1.[M_0, M_2]X_{p1}]$$
$$p_{end} = softmax(W_2.X_{p2})$$
$$MethodB: X_{p2} = F.ReLU(W_2.[M_0, M_2])$$
$$p_{end} = softmax(W_3.[X_{p1}X_{p2}])$$

Ensembling was utilized to combine multiple "weaker" models to build a "stronger" model with better accuracy. Two techniques were used - average probability and majority voting.

References

- [1] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. arXiv:1611.01603, 2016.
- [2] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi,and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. arXiv:1804.09541, 2018.

Results/Analysis

	Description	F1	EM	AvNA
Single models				
BiDAF	baseline	61.29	57.99	68.07
BiDAF	char_emb	63.34	60.07	70.04
BiDAF	char_emb, 3token_feat	66.09	62.70	72.26
QANet	5Conv, 1head, 96d_model, 64char_emb	66.02	62.51	72.93
QANet*	5Conv, 8head, 128d_model, 200char_emb	68.51	64.93	74.84
QANet	7Conv, 8head, 128d_model, 200char_emb	67.98	64.21	75.00
QANet**	QANet*, 3token_feat	**69.44**	**65.89**	**75.77**
QANet	QANet**, output layer changed	68.37	64.86	74.71
QANet	QANet**, question augmented	66.99	64.38	72.81
QANet	QANet**, paragraph augmented	67.97	64.39	74.79
Ensemble models				
QANet ensemble	average prediction	71.73	68.73	76.66
QANet ensemble	majority voting	71.4	68.55	76.17
QANet+BIDAF ensemble	majority voting	**72.2**	**69.7**	**76.89**

- Basic QANet model outperformed BiDAF achieving **68.51/64.93** F1/EM score. Complexity was gradually added, in order to evaluate the importance of each element on the performance.
  - No benefit was seen in increasing the Model Encoder Stack from 5 to 7.
  - A big improvement **(+2.5 F1 Score)** was seen by increasing **Attention heads** from 1 to 8, **Hidden size** from 96 to 128.
- **Character Embedding** and **Token features** were the most important enhancements on the architecture giving **+1.1 F1 score gain** each.
- Data Augmentation was effective in diversifying the input data-set for both Questions and Paragraphs.
- Ensembling gave a better prediction than any stand-alone model. **Average probability** performed better than majority voting.

Conclusion

- **QANet out-performed BiDAF.**
- All the architectural changes and fine-tuning of parameters ended up with the highest scores of:
  - **69.44/65.89** F1/EM score on the **dev set with single-model**
  - **72.2/69.7** F1/EM on the **dev set with ensemble**
  - **69.73/67.22** F1/EM score on the **hidden test set**
- The most-common mistake is answering un-answerable question. Adding a separate head for no-answer may help.