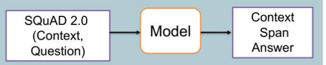
You Just Want Attention

CS224N Final Project | Winter 2022

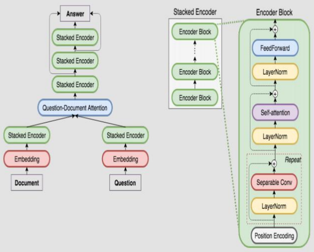
Pranav Jain Swastika Dutta

Objective



- Improve base **BiDAF** code with character encoding.
- Implement **QANet** architecture and experiment with different embeddings and hyperparameters.
- **Ensemble** results from high performing QANet and BiDAF models.

Methods

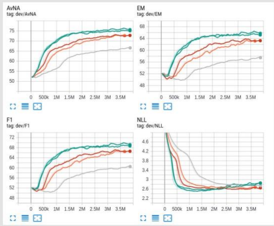


QANet Architecture

Experiments

- Character Embeddings in BiDAF and QANet.
- Glove vs fastText embeddings.
- Conditional Output Layer: End of answer is a dependent on the start
- Variants of QANet hyperparameters: Encoder layers, attentions heads, hidden size, regularization, and others.
- Data augmentation, Active Learning: Simplify 'hard' questions
- Ensemble Model: Majority Voting. Majority confidence-weighted voting to dissolve ties.

Observations



Effect of hidden size and learning rate dominates other factors like regularization, number of attention heads.

Results

Evaluation of BiDAF model:

Table 1: Comparison of GloVe vs fastText on SQuAD dev set

Model	AvNA	Overall EM	Overall F1	EM *	F1 * [1]
BiDAF with GloVe	66.5	56.2	59.2	59.1	57.5
BiDAF with fastText	68	57.8	61.3	59.9	60.3

Table 2: Analysis of character embeddings in BiDAF

Model	AvNA	Overall EM	Overall F1
BiDAF w character embedding	68.1	58.1	61.4
BiDAF w/o character embedding	66.5	56.3	59.5

Table 3: Analysis of conditonal output layer in BiDAF

Model	AvNA	Overall EM	Overall F1
BiDAF w conditional output	68.1	58.1	61.4
BiDAF w/o conditional output	68.7	58.3	61.5

Evaluation of QANet Model

Table 4: Results of different QANet models

Model	AvNA	Overall EM	Overall F1
QANet with 128 hidden size, batch size 64	75.2	65.18	69.88
QANet with 128 hidden size, batch size 64, L2	**75.53**	**65.78**	**70.13**
QANet with 128 hidden size, batch size 32	75.13	64.9	69.11
Larger QANet with 128 hidden size, batch size 32	74.5	64.5	68.99
QANet with 128 hidden size, batch size 32, 4 attention heads	74.2	64.3	69.33
QANet with 64 hidden size, batch size 64	72.3	64.3	68.99
QANet with 64 hidden size, batch size 32	72.9	64.1	68.79

Evaluation of Ensemble Method

Table 5: Analysis of Ensemble Models

Model	AvNA	Overall EM	Overall F1
Vanilla Majority Voting	76.13	67.13	71.11
Weighted Majority Voting	76.55	68.54	71.31

Analysis

Evaluation Metrics:
- Overall EM, F1, AvNA scores
- EM, F1 for answerable questions
- FP, FN AvNA scores

Table 6: Analysis of question types in dev set

Question type	What	How	Where	Why	Who	When	Others
Examples starting with question type	907	68	37	4	75	106	4881
Examples with question type	3726	590	257	91	688	427	289

Table 7: Analysis of answer positions in dev set

Answer position	No Answer	First line of context	Later lines of context
Number of examples	3168	918	1992

Conclusion

- We develop an efficient QANet model with single model with 70.13/ 67.31 dev set F1/EM score
- Ensembling improves dev set F1/EM score to 71.63/68.85 (Rank 5) and test set F1/EM score to 69.42/66.46 (Rank 6).
- Potentially improve score by data augmentation, context-based embeddings
- **Bonus: We are planning to use our findings and model to start a non-profit company to help kids' self learn English.**

Acknowledgements

Thank you Ethan Chi and Elaine Sui for guiding us in the project!

References

- QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension, Adams et. al.
- Machine Learning Using Match-LSTM and Answer Pointer, Wang et. al.