IID SQuAD QANet Project

Daniel Guillen, Claire Mai

Stanford University - Department of Computer Science

{eliasd,cmai21}@stanford.edu

Problem

- NLP task: question-and-answering problem
- Input:
  - Context paragraph with n words $C = \{c_1, \ldots, c_n\}$
  - Input question with m words $Q = \{q_1, \ldots, q_m\}$
- Output: Span prediction (if is no answer available in the context paragraph, the model abstains from answering)
- **Context**: "The 8- and 10-county definitions are not used for the greater Southern California Megaregion, one of the 11 megaregions of the United States. The megaregion's area is more expansive, extending east into Las Vegas, Nevada, and south across the Mexican border into Tijuana."
- **Question**: "What is the name of the region that is not defined by the eight or 10 county definitions?"
- **Answer**: "Southern California Megaregion"

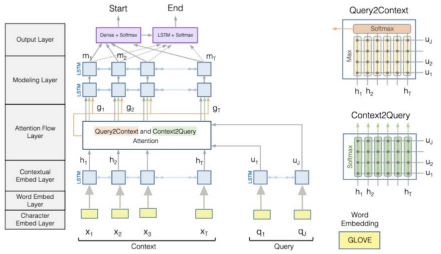
Data

Data Source: SQuAD 2.0 (Stanford Question Answering Dataset)

- Sourced from Wikipedia articles
- Contains 150,000 questions: 100,000 answerable questions and 50,000 adversarial unanswerable questions
- Split into 3 separate datasets: train, dev, and test.
- Train set has 129,941 examples
- Dev set has 6078 examples
- Test set has 5915 examples

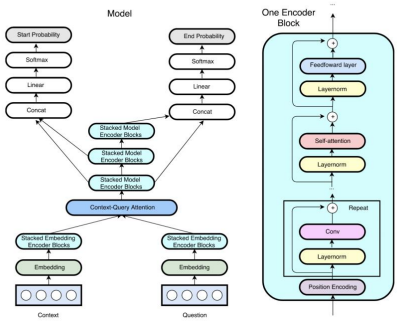
Methods: BiDAF

BiDAF (baseline)



Methods: QANet

QANet



Regularization:

1. L2 weight decay, $\lambda = 3x10^{-7}$
2. Dropout rate of 0.05 and 0.1 is applied to the character embeddings and word embeddings respectively
3. Layer dropout is applied to each encoder block
   - Each sublayer l has a survival probability of $p_l = 1 - \frac{l}{L}(1 - p_L)$ where L is the last layer and $p_L = 0.9$ [4].

Loss Function: negative log likelihood

$$L(\theta) = -\frac{1}{N}\sum[log(p^1_{y^1_i}) + log(p^2_{y^2_i})] \quad (1)$$

- y^1_i is the ground truth starting position of the ith example
- y^2_i is the ground truth ending position of the ith example
- θ contains all the trainable variables
- Optimizer: Adam with $\beta_1 = 0.8$ and $\beta_2 = 0.999$
- Inference: choose (start, end) that maximizes $p^2_{start}p^2_{end}$ such that $start \leq end$

Experiments

1. All models trained for 2.5 million iterations
2. Note QANet-final used a warmup scheme learning rate where first 1000 steps LR exponentially increases from 0 to 0.001

Model	Optimizer	LR	F1-score(dev)	F1-score(test)	EM-score(dev)	EM-score(test)
Baseline	Adadelta	0.5	58	N/A	55	N/A
QANet-v1	Adam	0.001	50.294	N/A	50.294	N/A
QANet-v2	Adam	0.001	52.193	N/A	52.193	N/A
QANet-final	Adam	0.001^*	64.5	62.570	61.334	58.969

Analysis

- QANet-v1 did worse than the baseline because we had not yet implemented the encoder block layers
- QANet-v2 mis-implemented the separable depthwise convolutions and forgot a ReLU activation and masked the attention maps incorrectly
- self-attention is a key component the QANet model
  - allows every word in the input attend to every other word
  - placed greater emphasis on non word elements which lead to worse performance.

		Ground Truth		
		Answer	No Answer	Total
Prediction	Answer	2410	1203	3613
	No Answer	438	1900	2338
	Total	2848	3103	5951

Model	TPR	TNR	FPR	FNR
QANet-final	84.62%	61.23%	38.77%	15.38%

Conclusion

- Achieved good results on the SQuAD 2.0 dataset (dev F1-score of 64.5 and test F1-score of 62.57)
- Dev set has more unanswerable questions than answerable ones, but model predicted an answer more often than no-answer as indicated by the TPR being higher than the TNR
- Limitations: Only tested our model on the SQuAD 2.0 dataset
- Don't know if QANet model will produce promising results on other Q&A datasets or perform well on other tasks

Future Work

- Incorporate ensemble methods by combining multiple 'weaker' learners to build a stronger learn for the task
  - Could potentially average start/end probabilities that each model outputs
  - Could also implore majority voting whether the most popular start and end tokens amongst the models will be the answer for the predicted span

References

References

[1] Pranav Rajpurkar et al. "SQuAD: 100, 000+ Questions for Machine Comprehension of Text". In: *CoRR* abs/1606.05250 (2016). arXiv: 1606.05250. URL: http://arxiv.org/abs/1606.05250.

[2] Minjoon Seo et al. *Bidirectional Attention Flow for Machine Comprehension*. 2018. arXiv: 1611.01603 [cs.CL].

[3] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. *Highway Networks*. 2015. arXiv: 1505.00387 [cs.LG].

[4] Adams Wei Yu et al. "QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension". In: *CoRR* abs/1804.09541 (2018). arXiv: 1804.09541. URL: http: