CS 224N Ruminating QANet

Rafael Rafailov (rafailov@stanford.edu)

Objectives

The goal of this project is to improve the architecture of Machine Comprehension models by expanding attention mechanisms with a deeper recurrent inference chain. I have two major contributions:

- Extend the Ruminating Reader Block ([1]) to a recurrent chain and combine the model with the Encoder Block of QANet ([2]).
- Propose and entirely new convolutional Ruminating Block, with dual attention layers.

Model 1

The first model applies the original Ruminating Block with some modifications. Blocks are then stacked together in a sequential manner. It consists of three main units:

- **Attention Summarization Layer** The attention Summarization Layer deploys an LSTM model on the output of the Attention Layer to return hidden states

$$\mathbf{s}_i^t = [\mathbf{s}_{i,fwd}^t; \mathbf{s}_{i,rev}^t] \in \mathbb{R}^{4H}$$

and summary state

$$\mathbf{s}^t = [\mathbf{s}_{len\_c,fwd}^t; \mathbf{s}_{1,rev}^t] \in \mathbb{R}^{4H}$$

.

The superscript t indicates that this is stack-t of the ruminating chain.

- **Context Summarization Layer** This layer is a gating mechanism, which aims to compute query-aware context vectors:

$$\mathbf{z}_i^t = \tanh(W_{C_z}^1 \mathbf{s}_i^t + W_{C_z}^2 \mathbf{c}_i + b_{C_z})$$

$$\mathbf{f}_i^t = \sigma(W_{C_f}^1 \mathbf{s}_i^t + W_{C_f}^2 \mathbf{c}_i + b_{C_f})$$

$$\mathbf{c}_i^t = \mathbf{f}_i^t \circ \mathbf{c}_i + (1 - \mathbf{f}_i^t) \circ \mathbf{z}_i^t$$

Here $\mathbf{c}_i$ are the context word encodings.

- **Query Summarization Layer** Similar to the Context Summarization Layer, but using the summary state vector.

$$\mathbf{z}_i^t = \tanh(W_{Q_z}^1 \mathbf{s}^t + W_{Q_z}^2 \mathbf{q}_i + b_{Q_z})$$

$$\mathbf{f}_i^t = \sigma(W_{Q_f}^1 \mathbf{s}^t + W_{Q_f}^2 \mathbf{q}_i + b_{Q_f})$$

$$\mathbf{q}_i^t = \mathbf{f}_i^t \circ \mathbf{q}_i + (1 - \mathbf{f}_i^t) \circ \mathbf{z}_i^t$$

Here $\mathbf{q}_i$ are the query word encodings.

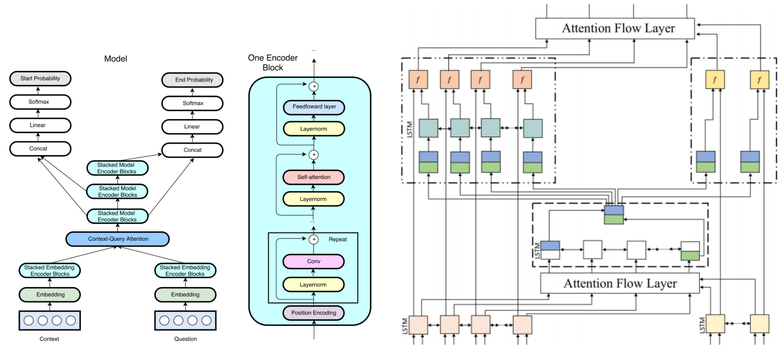

Figure 1: **A**: Original QANet structure, with an Encoder Block (figure adapted from [1]), **B**: Original Ruminating Block (figure adapted from [2])

Model 2

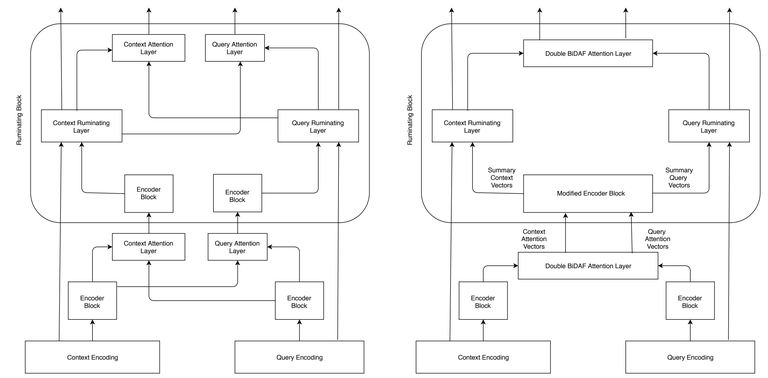

Figure 2: **A**: Model 2, as originally trained, **B**: Alternative Model 2 with Attention Layer weight sharing and Modified Encoder Block

Model 2 is an original structure for the Ruminating Block.

- **Attention Layer** We add an additional Attention Module, that swaps the Query-to-Context and Context-to-Query steps in the original to compute query-attention vectors.
- **Summarization Layer** The summary LSTM is replaced by an Encoder Block, which operates on the output of both attention layers and returns context-summaries $\mathbf{s}_c^t$ and query-summaries $\mathbf{s}_q^t$.
- **Ruminating Layers** Both Ruminating Layers operate like the Context Summarization Layer from Model 1, using $\mathbf{s}_c^t$ and $\mathbf{s}_q^t$ respectively.
- **Modeling Layer** The modeling layer takes as input the context-attention vectors from the last itteration of the chain.

Results

Summary of model results:

Trained Models Result Summary				
Model	Development Set		Test Set	
	EM	F1	EM	F1
Baseline	55.00	57.00	-	-
Baseline + Ruminating Chain v1	57.33	60.92	-	-
QANet + Ruminating Block v1	63.59	67.02	60.03	63.70
Model 1 + Multi-head Attention	63.12	66.56	-	-
QANet + Ruminating Chain v2	**63.89**	**67.52**	**61.05**	**64.63**

Major observations:

- Adding the Ruminating Chain with the architecture from Model 1 to the Baseline model increased performance on the dev set significantly.
- Adding multiple attention heads (or quadratic similarity) does not improve performance.
- Our proposed Ruminating Chain adds 1 point in EM and F1 scores over the original version.

Conclusion and next steps

This project proposed an extension of the existing Ruminating Block into a chain and combined it with the general set-up of QANet, which improved performance over each model. We also proposed a an entirely new original model for the Ruminating Block and showed that it outperformed all previous models. There are multiple avenues for further development:

- Explore the inference capacity of the ruminating chain: varying depth and attention complexity.
- Train the alternative form of Model 2.
- Perform more extensive hyper-parameter search (i.e. learning rate annealing).

References

[1] Yichen Gong and Samuel R. Bowman.
Ruminating reader: Reasoning with gated multi-hop attention.
ArXiv preprint, 2017.
http://arxiv.org/1704.07415.

[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 preprint, 2018.
https://arxiv.org/abs/1608.07905.