Query-Reduction Networks for Question Answering

M. Seo, S. Min, A. Farhadi, H. Hajishirzi

University of Washington
Seoul National University
Allen Institute for Artificial Intelligence

arXiv: 1606.04582

Reviewed by : Bill Zhang
University of Virginia
https://qdata.github.io/deep2Read/

Outline

Introduction

Basic Premise and Motivation

- Want to address QA problem where multiple facts are required
- Examples of recent tasks are story-based QA and dialog tasks
- RNNs have inherently unstable long-term memory, making them unsuitable for multi-hop reasoning; can use global attention, but this doesn't account for time step of sentences
- Propose Query-Reduction Network (QRN) which reduces query to more informed queries over time
  - Query = Where is the apple? $\Rightarrow$ "Sandra picked up the apple" $\Rightarrow$ Query = Where is Sandra?
- Better encode locality information because it does not use global memory access controller

Model Diagrams
QRN

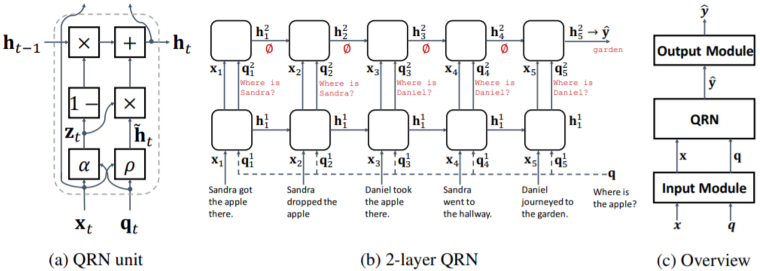
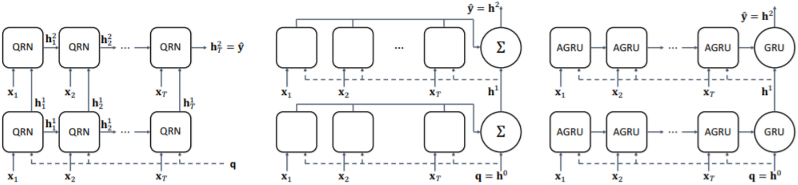


Figure 1: (1a) QRN unit, (1b) 2-layer QRN on 5-sentence story, and (1c) entire QA system (QRN and input / output modules). $\boldsymbol{x}$, $\boldsymbol{q}$, $\hat{\boldsymbol{y}}$ are the story, question and predicted answer in natural language, respectively. $\mathbf{x} = \langle \mathbf{x}_1, \ldots, \mathbf{x}_T \rangle$, $\mathbf{q}$, $\hat{\mathbf{y}}$ are their corresponding vector representations (upright font). α and ρ are update gate and reduce functions, respectively. $\hat{\mathbf{y}}$ is assigned to be $\mathbf{h}_5^2$, the local query at the last time step in the last layer. Also, red-colored text is the inferred meanings of the vectors (see 'Interpretations' of Section 5.3).

Model Diagrams

Comparison



(a) QRN (b) N2N (Sukhbaatar et al., 2015) (c) DMN+ (Xiong et al., 2016)

Figure 2: The schematics of QRN and the two state-of-the-art models, End-to-End Memory Networks (N2N) and Improved Dynamic Memory Networks (DMN+), simplified to emphasize the differences among the models. AGRU is a variant of GRU where the update gate is replaced with soft attention, proposed by Kumar et al. (2016). For QRN and DMN+, only forward direction arrows are shown.

Model

- Given context (list of T sentences $x_1 \ldots x_T$) and question q, generate answer $\hat{y}$; true answer is y
- Three stages: input module, QRN layers, output module
- Input module maps each sentence x_i and q to $\mathbb{R}^d$
- QRN layers generate predicted answer $\hat{\mathrm{y}} \in \mathbb{R}^d$ using vectors from input module
- Output module converts $\hat{\mathrm{y}}$ to natural language answer $\hat{y}$

QRN Unit

- QRN updates its hidden state (reduced query) through time and layers
- Accepts 2 inputs (local query vector $q_t \in \mathbb{R}^d$ and sentence vector $x_t \in \mathbb{R}^d$); produces 2 outputs (reduced query vector $h_t \in \mathbb{R}^d$ and x_t with no modification)
- Use update gate function $\alpha : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ and reduce function $\rho : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$
- Update gate measures how relevant sentence is to query
- Reduce function produces reduced query

QRN Unit

Function Formulas

- Update gate similar to global attention mechanism, only uses sigmoid on current memory slot, not entire memory (i.e. local sigmoid attention)

$$z_t = \alpha(\mathbf{x}_t, \mathbf{q}_t) = \sigma(\mathbf{W}^{(z)}(\mathbf{x}_t \circ \mathbf{q}_t) + b^{(z)}) \tag{1}$$

$$\tilde{\mathbf{h}}_t = \boldsymbol{\rho}(\mathbf{x}_t, \mathbf{q}_t) = \tanh(\mathbf{W}^{(\mathbf{h})}[\mathbf{x}_t; \mathbf{q}_t] + \mathbf{b}^{(\mathbf{h})}) \tag{2}$$

$$\mathbf{h}_t = z_t \tilde{\mathbf{h}}_t + (1 - z_t)\mathbf{h}_{t-1} \tag{3}$$

where z_t is the scalar update gate, $\tilde{\mathbf{h}}_t$ is the candidate reduced query, and $\mathbf{h}_t$ is the final reduced query at time step t, $\sigma(\cdot)$ is sigmoid activation, $\tanh(\cdot)$ is hyperboolic tangent activation (applied element-wise), $\mathbf{W}^{(z)} \in \mathbb{R}^{1\times d}$, $\mathbf{W}^{(\mathbf{h})} \in \mathbb{R}^{d\times 2d}$ are weight matrices, $b^{(z)} \in \mathbb{R}$, $\mathbf{b}^{(\mathbf{h})} \in \mathbb{R}^d$ are bias terms, $\circ$ is element-wise vector multiplication, and $[;]$ is vector concatenation along the row. As a base case, $\mathbf{h}_0 = \mathbf{0}$. Here we have explicitly defined α and $\boldsymbol{\rho}$, but they can be any reasonable differentiable functions.

QRN Unit

- Can stack QRN units (in earlier figure); let $q_t^{k+1} = h_t^k$
- Incorporate bi-direction since sometimes query answers depend on future sentences; use sum of both direction states $q_t^{k+1} = \overrightarrow{h}_t^k + \overleftarrow{h}_t^k$
- Take $\hat{\mathrm{y}} = h_t^K$ where K is number of QRN layers, then convert to $\hat{y}$ using output module

QRN Unit

Extensions

- Reset Gate: Function $\beta : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, 1]$ which allows nullification of candidate query
- Vector Gate: Allow update and reset gates to be vectors instead of scalars for more fine-controlled gating

$$r_t = \beta(\mathbf{x}_t, \mathbf{q}_t) = \sigma(\mathbf{W}^{(r)}(\mathbf{x}_t \circ \mathbf{q}_t) + b^{(r)}) \tag{5}$$

where $\mathbf{W}^{(r)} \in \mathbb{R}^{1 \times d}$ is a weight matrix, and $b^{(r)} \in \mathbb{R}$ is a bias term. Equation 3 is rewritten as

$$\mathbf{h}_t = z_t r_t \tilde{\mathbf{h}}_t + (1 - z_t)\mathbf{h}_{t-1}. \tag{6}$$

Parallelization

- Can decompose equation 3 ($h_t = z_t \tilde{h}_t + (1 + z_t)h_{t-1}$) into computing over only candidate reduced queries ($\tilde{h}_t$)) without worrying about previous hidden state
- More details in paper

$$\mathbf{h}_t = \sum_{i=1}^{t} \left[\prod_{j=i+1}^{t} 1 - z_j \right] z_i \tilde{\mathbf{h}}_i = \sum_{i=1}^{t} \exp \left\{ \sum_{j=i+1}^{t} \log\left(1 - z_j\right) \right\} z_i \tilde{\mathbf{h}}_i .$$

Experiments

Data and Model Details

- Tested on bAbI story-based QA, bAbI dialog, and DSTC2 (Task 6) dialog datasets
- For input module, use trainable embedding matrix $A \in \mathbb{R}^{dxV}$ to get d dimensional one-hot vector for each word in sentence or query; then get sentence or query representation using Postion Encoder (Weston et al., 2015)

Experiments

Data and Model Details

- For story-based QA output model, use V-way (V = size of vocabulary) single layer softmax layer, then pick argmax word
- For dialog output model, use fixed number single-layer softmax classifiers to sequentially output next word

- Compare with baselines and previous state-of-the-art models: LSTM, End-to-end Memory Networks (N2N), Dynamic Memory Networks (DMN+), Gated End-to-end Memory Networks (GMemN2N), and Differentiable Neural Computer (DNC)
- Also perform ablations with number of layers, reset gate, gate vectorization, and dimension of hidden vector

Experiments

Table Results

Task	1k						10k				
	Previous works				QRN		Previous works				QRN
	LSTM	N2N	DMN+†	GMemN2N	2r	3r	N2N	DMN+	GMemN2N	DNC	6r200
# Failed	20	10	16	10	7	**5**	3	1	3	2	**0**
Average error rates	51.3	15.2	33.2	12.7	**9.9**	11.3	4.2	2.8	3.7	3.8	**0.3**

Task	Plain				With Match		
	Previous works		QRN		Previous works		QRN
	N2N	GMemN2N	2r	2r100	N2N+	GMemN2N+	2r+
bAbI dialog Average error rates	13.9	14.3	**5.5**	**5.5**	6.7	5.4	**1.5**
bAbI dialog (OOV) Average error rates	30.3	27.9	**11.1**	**11.1**	11.2	10.3	**2.3**
DSTC2 dialog Average error rates	58.9	52.6	49.5	**48.9**	59.0	51.3	**49.3**

Table 1: (top) bAbI QA dataset (Weston et al., 2016): number of failed tasks and average error rates (%). † is obtained from `github.com/therne/dmn-tensorflow`. (bottom) bAbI dialog and DSTC2 dialog dataset (Bordes and Weston, 2016) average error rates (%) of QRN and previous work (LSTM, N2N, DMN+, GMemN2N, and DNC). For QRN, the first number (1, 2, 3) indicates the number of layers, 'r' means the reset gate is used, and the last number (100, 200), if exists, indicates the dimension of the hidden state, where the default value is 50. '+' indicates that 'match' (See Appendix for details) is used. The task-wise results are shown in Appendices: Table 2 (bAbI QA) and Table 3 (dialog datasets). See Section 5.3 for details.

Experiments

Ablation Analysis

- Model could not reason well when layers too low; harder to train when layers too high
- Reset gate helps results
- Vector gates hurt for 1k dataset since model overfits or converges to local minima
- Larger hidden size helps some cases

Experiments

More Observations

- Parallelization speeds up QRN on average by 6.2x
- Advantage of QRN is we can interpret intermediate queries using decoder; can track how query changes
- Can also visualize reset and update gate magnitudes; low reset gate magnitude r means candidate query from current sentence is misrepresentative, low update gate magnitude z means sentence irrelevant to query

Experiments

Magnitude Visualization

	Layer 1			Layer 2
Task 2: Two Supporting Facts	z^1	$\overrightarrow{r}^1$	$\overleftarrow{r}^1$	z^2
Sandra picked up the apple there.	0.95	0.89	0.98	0.00
Sandra dropped the apple.	0.83	0.05	0.92	0.01
Daniel grabbed the apple there.	0.88	0.93	0.98	0.00
Sandra travelled to the bathroom.	0.01	0.18	0.63	0.02
Daniel went to the hallway.	0.01	0.24	0.62	0.83
Where is the apple?	hallway			

	Layer 1			Layer 2
Task 15: Deduction	z^1	$\overrightarrow{r}^1$	$\overleftarrow{r}^1$	z^2
Mice are afraid of wolves.	0.11	0.99	0.13	0.78
Gertrude is a mouse.	0.77	0.99	0.96	0.00
Cats are afraid of sheep.	0.01	0.99	0.07	0.03
Winona is a mouse.	0.14	0.85	0.77	0.05
Sheep are afraid of wolves.	0.02	0.98	0.27	0.05
What is Gertrude afraid of?	wolf			

	Layer 1			Layer 2
Task 3: Displaying options	z^1	$\overrightarrow{r}^1$	$\overleftarrow{r}^1$	z^2
resto-paris-expen-frech-8stars?	0.00	1.00	0.96	0.91
Do you have something else?	0.41	0.99	0.00	0.00
Sure let me find another option.	1.00	0.00	0.00	0.12
resto-paris-expen-frech-5stars?	0.00	1.00	0.96	0.91
No this does not work for me.	0.00	0.00	0.14	0.00
Sure let me find an other option.	1.00	0.00	0.00	0.12
What do you think of this? resto-paris-expen-french-4stars				

	Layer 1			Layer 2
Task 6: DSTC2 dialog	z^1	$\overrightarrow{r}^1$	$\overleftarrow{r}^1$	z^2
Spanish food.	0.84	0.07	0.00	0.82
You are lookng for a spanish restaurant right?	0.98	0.02	0.49	0.75
Yes.	0.01	1.00	0.33	0.13
What part of town do you have in mind?	0.20	0.73	0.41	0.11
I don't care.	0.00	1.00	0.02	0.00
What price range would you like?	0.72	0.46	0.52	0.72
I don't care.	API CALL spanish R-location R-price			

Figure 3: (top) bAbI QA dataset (Weston et al., 2016) visualization of update and reset gates in QRN '2r' model (bottom two) bAbI dialog and DSTC2 dialog dataset (Bordes and Weston, 2016) visualization of update and reset gates in QRN '2r' model. Note that the stories can have as many as 800+ sentences; we only show part of them here. More visualizations are shown in Figure 4 (bAbI QA) and Figure 5 (dialog datasets).

Conclusion

- Introduced QRNs for QA and dialog tasks which require multi-hop reasoning
- Showed state-of-the-art performance for story-based QA and dialog tasks
- QRN effectively handles time dependency and long-term dependency problems present in attention mechanisms and RNNs
- QRNs can be parallelized and address RNN's vanishing gradient problem

References

- https://arxiv.org/pdf/1606.04582.pdf