Hierarchical / Dual Attention

Xiao i - Chen Lu
School of Computer and Software Engineering - ICA

Outline

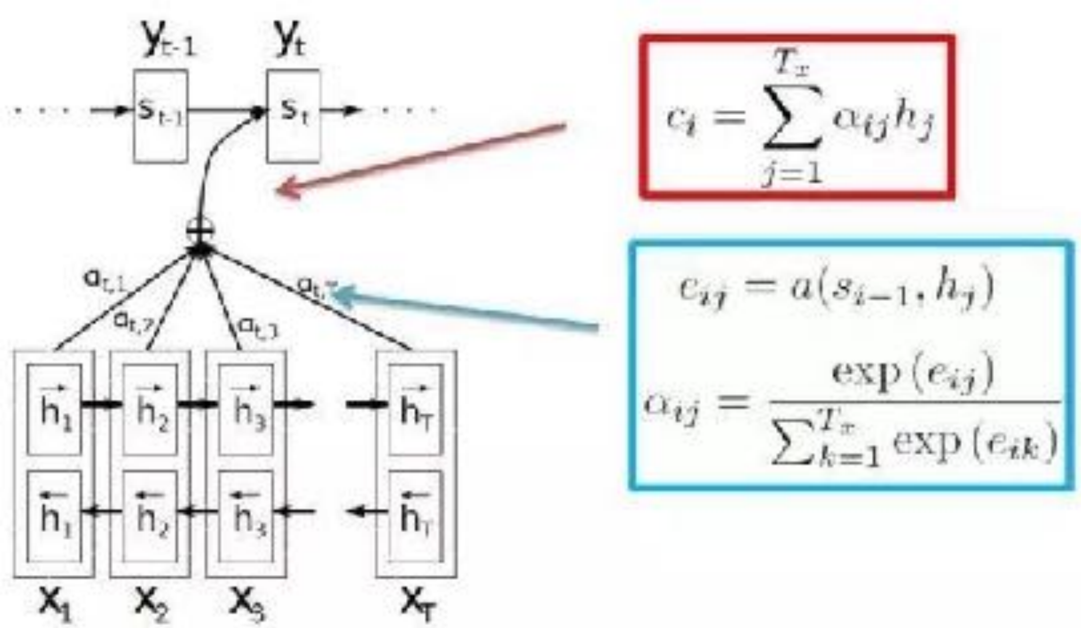
- Introduction
- [**KDD17**] A Context-aware Attention Network for Interactive Question Answering
- [**SIGIR17**] Leveraging Contextual Sentence Relations for Extractive Summarization Using a Neural Attention Model
- [**RecSys17**] Interpretable Convolutional Neural Networks with Dual Local and Global Attention for Review Rating Prediction
- Conclusion

Outline

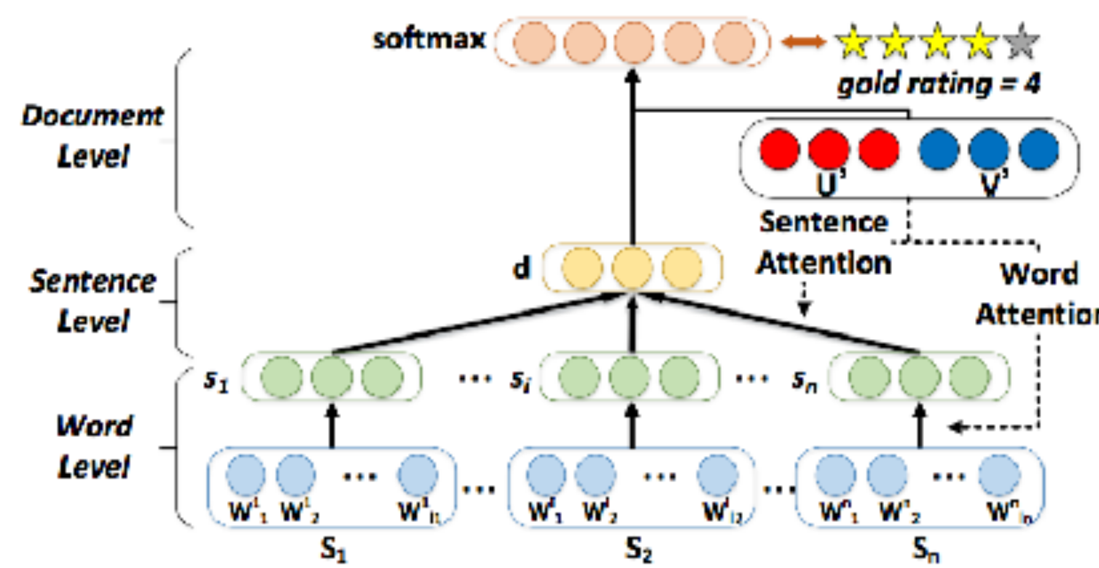
- Introduction
- [KDD17] A Context-aware Attention Network for Interactive **Question Answering**
- [SIGIR17] Leveraging Contextual Sentence Relations for **Extractive Summarization** Using a Neural Attention Model
- [RecSys17] Interpretable Convolutional Neural Networks with Dual Local and Global Attention for **Review Rating Prediction**
- Conclusion

Introduction

- Attention Model

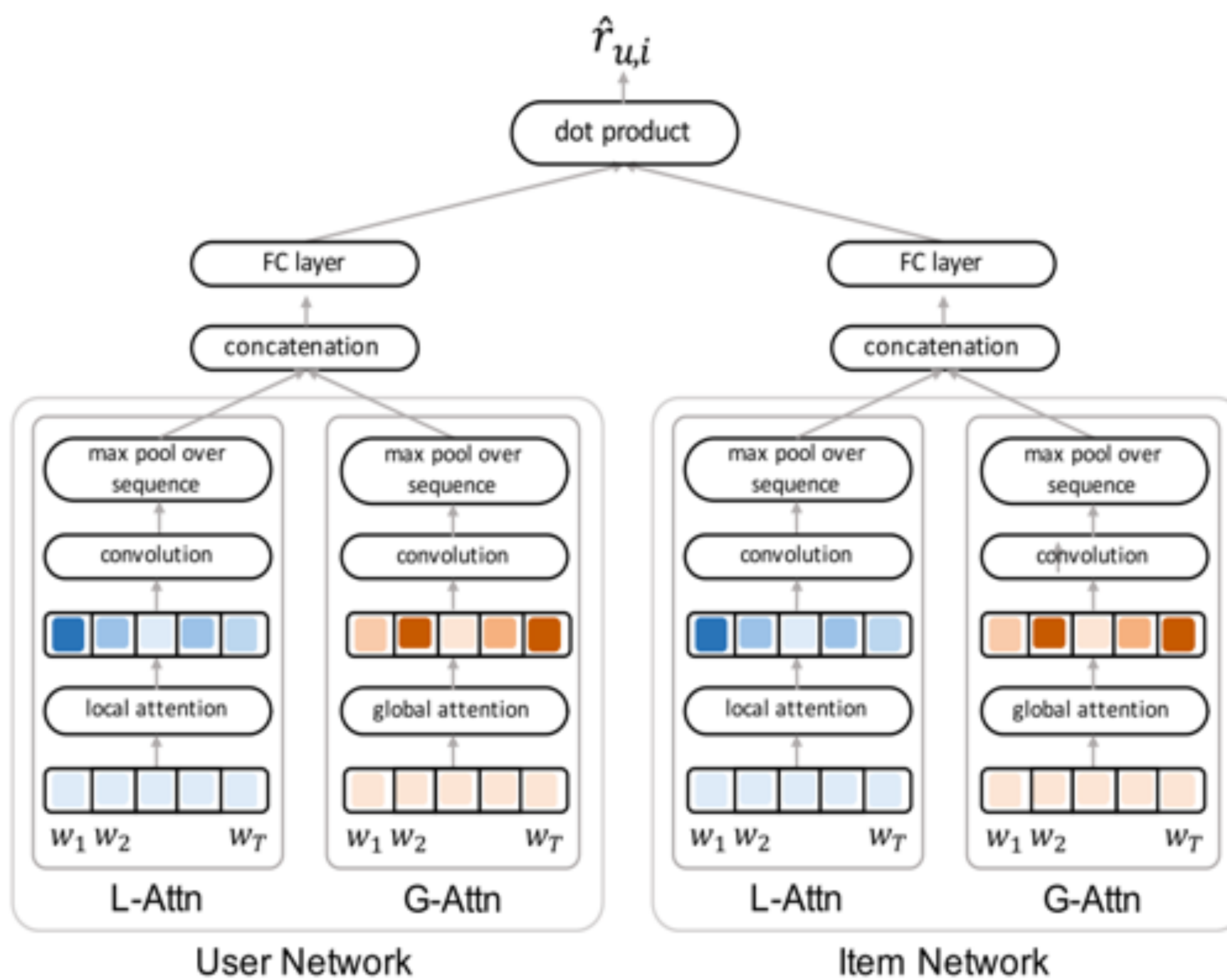


- Hierarchical Attention



Introduction

- Dual Attention



A Context-aware Attention Network for Interactive Question Answering*

Huayu Li[1], Martin Renqiang Min[2], Yong Ge[3], Asim Kadav[2]
[1]Department of Computer Science, UNC Charlotte
[2]Machine Learning Group, NEC Laboratories America
[3]Management Information Systems, University of Arizona
hli38@uncc.edu,{renqiang,asim}@nec-labs.com,yongge@email.arizona.edu.

Task

- QA: predicting answers from **statements** and **questions**.
- An **encoder-decoder** framework
- A **sequence-to-sequence** model
- EX.

The office is north of the kitchen.
The garden is south of the kitchen.
Q: What is north of the kitchen?
A: Office

Limitation of Related Work

- Fail to model **context-dependent** meaning of words.
- Fail to address **unknown states** under which systems do not have enough information to answer given questions.
- EX.

The office is north of the kitchen. The garden is south of the kitchen. Q: What is north of the kitchen? A: Office	The master bedroom is east of the garden. The guest bedroom is east of the office. Q: What is the bedroom east of? A: Unknown

Which bedroom the user refers to?

Context-aware Attention Network

- Learning Rep. for Sentences:

  - Context-dependent **word-level attention** for more accurate statement representations.

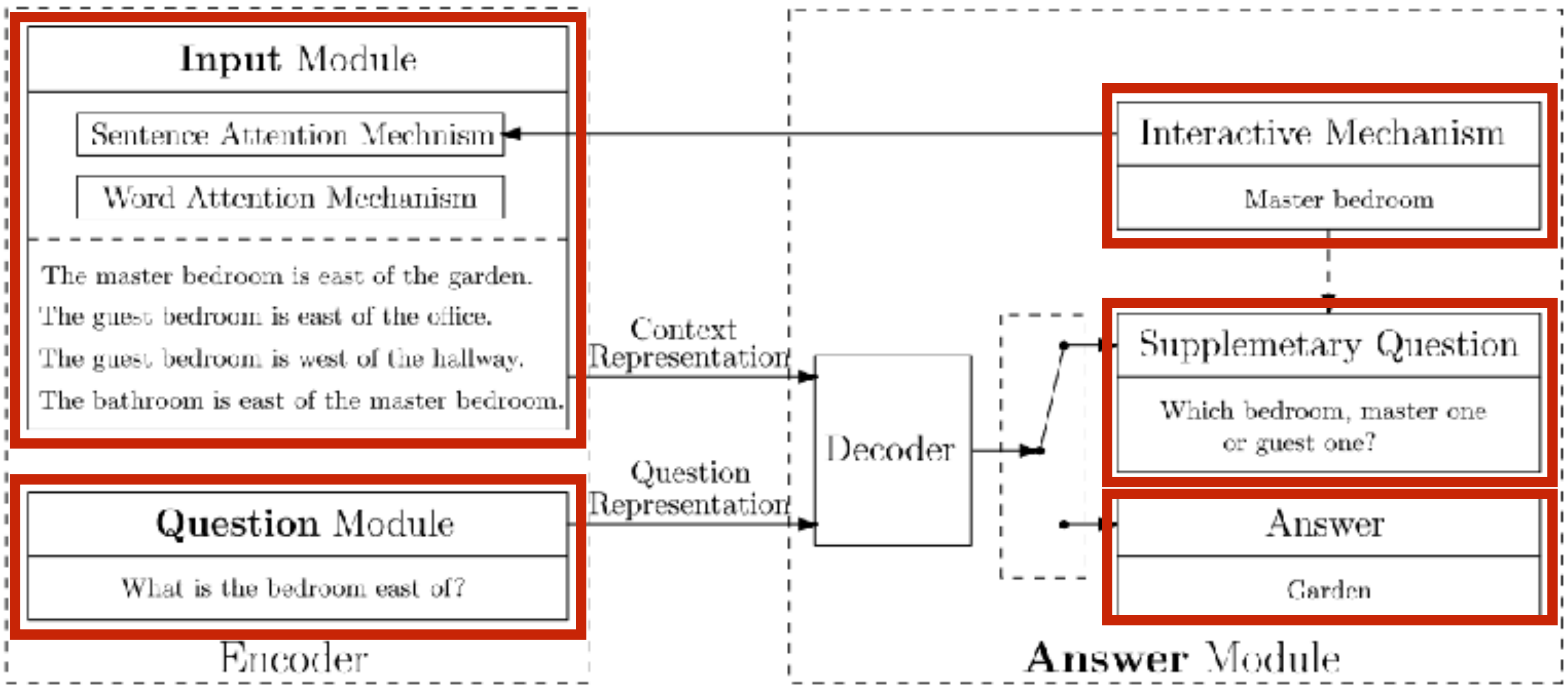
  - Question-guided **sentence-level attention** for context modeling.

- Interactive Question Answering:

  - A mechanism to **interact with user** to comprehensively understand a given question.

Context-aware Attention Network

- EX.



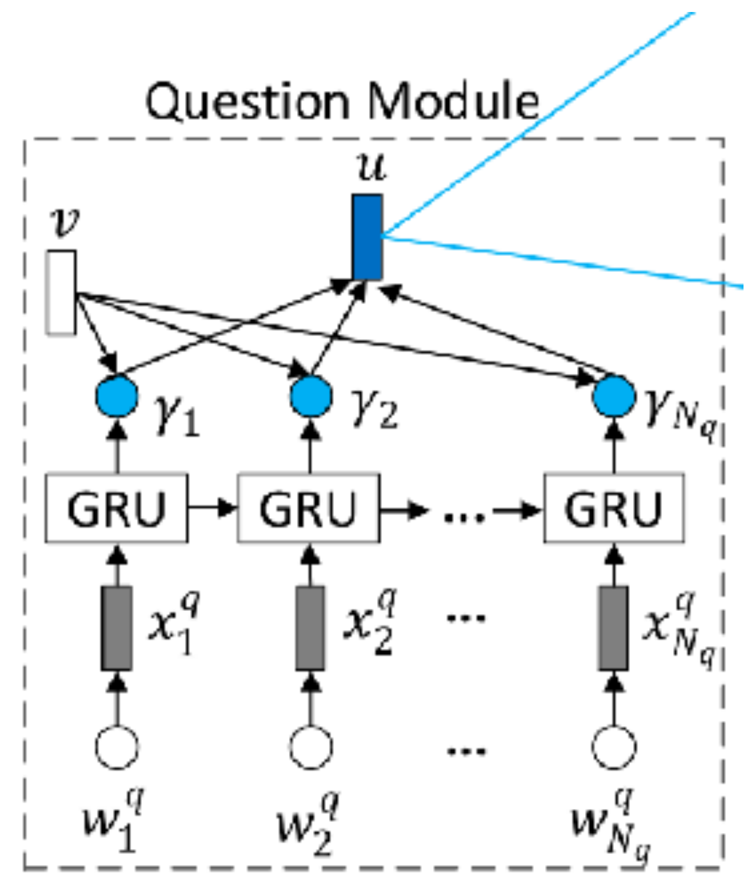
Context-aware Attention Network

- Question Module:

$$g_j^q = GRU_w(g_{j-1}^q, x_j^q)$$

$$\gamma_j = softmax(\mathbf{v}^T \mathbf{g}_j^q)$$

$$\mathbf{u} = \mathbf{W}_{ch} \sum_{j=1}^{N_q} \gamma_j \mathbf{g}_j^q + \mathbf{b}_c^{(q)}$$



Context-aware Attention Network

- Input Module:

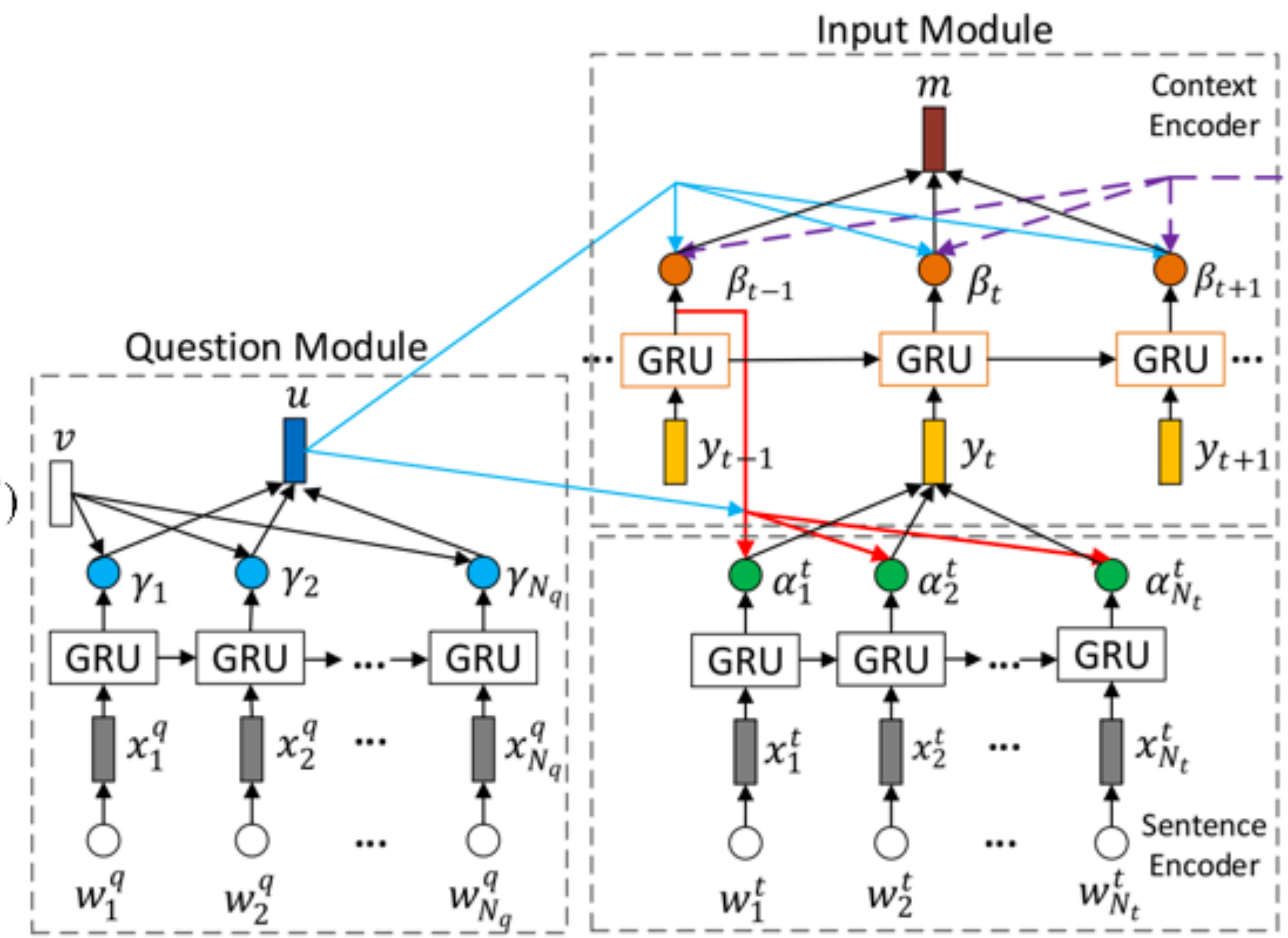
- Sentence Encoder

$$\mathbf{h}_i^t = GRU_w(\mathbf{h}_{i-1}^t, \mathbf{x}_i^t)$$

$$\mathbf{e}_i^t = \sigma(\mathbf{W}_{ee} tanh(\mathbf{W}_{es}\mathbf{s}_{t-1} + \mathbf{W}_{eh}\mathbf{h}_i^t + \mathbf{b}_e^{(1)}) + \mathbf{b}_e^{(2)})$$

$$\alpha_i^t = softmax(\mathbf{u}^T \mathbf{e}_i^t)$$

$$\mathbf{y}_t = \sum_{i=1}^{N_t} \alpha_i^t \mathbf{h}_i^t$$



Context-aware Attention Network

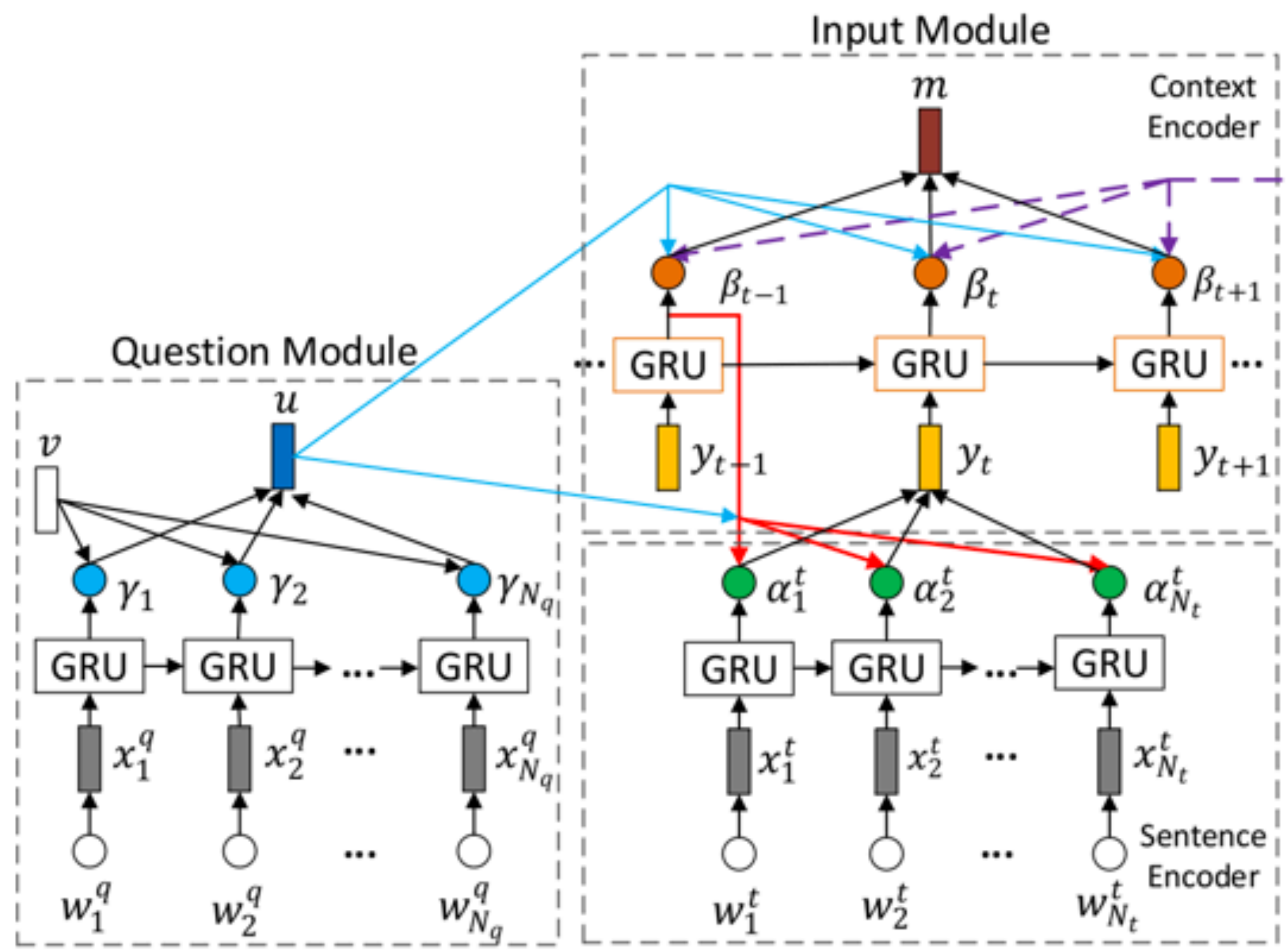
- Input Module:

  - Context Encoder

$$\mathbf{s}_t = GRU_s(\mathbf{s}_{t-1}, \mathbf{y}_t)$$

$$\beta_t = softmax(\mathbf{u}^T \mathbf{s}_t)$$

$$\mathbf{m} = \sum_{t=1}^{N} \beta_t \mathbf{s}_t$$



Context-aware Attention Network

- Answer Module: two output cases

  - Generating an **answer** after receiving the context and question information.

  - Generating a **supplementary question** and then uses the user's feedback to predict an answer.

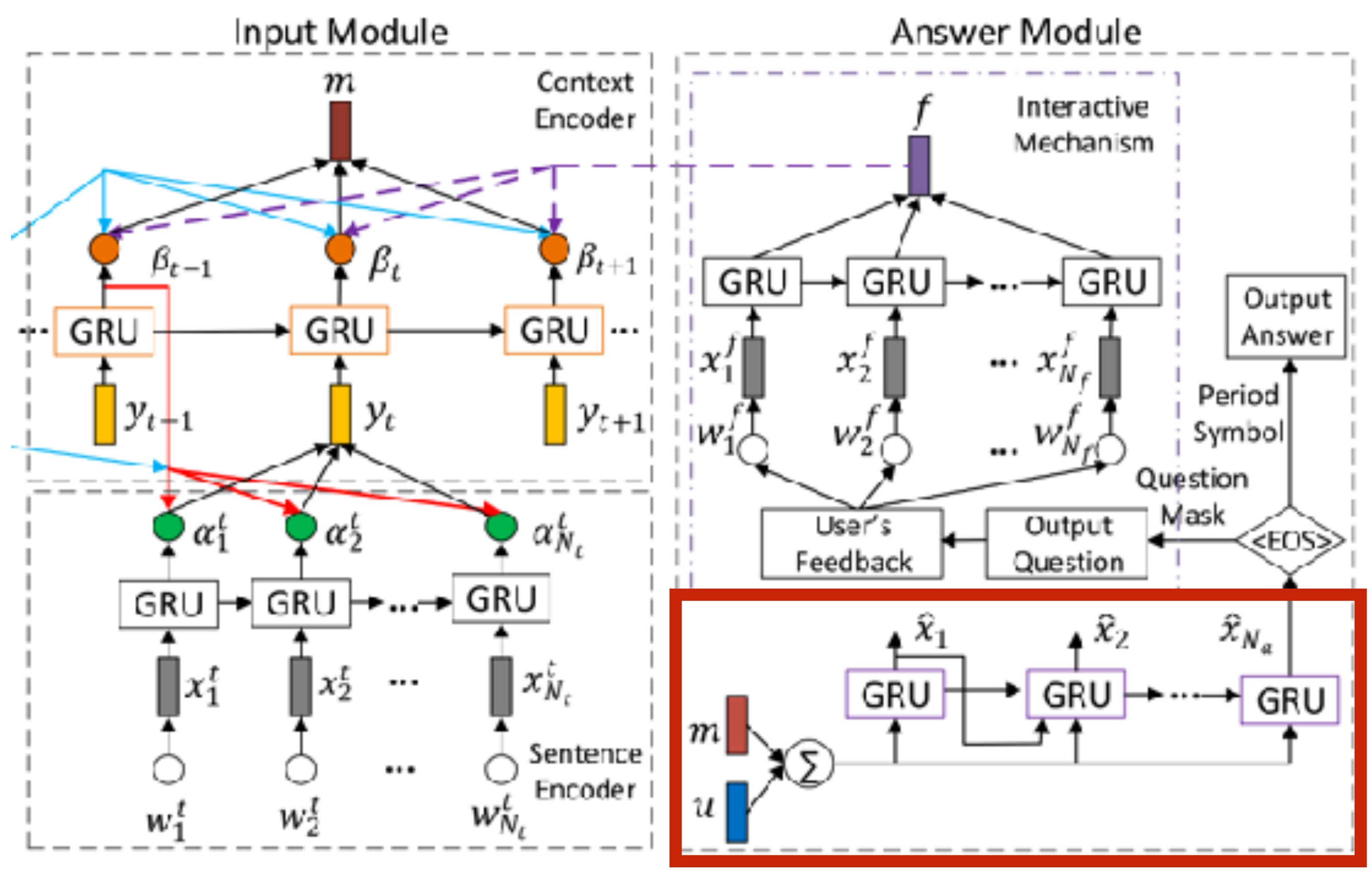
Context-aware Attention Network

- Answer Module:

  - Answer Generation

$$\hat{\mathbf{x}}_k \overset{\mathbf{W}_w}{=} softmax(\mathbf{W}_{od}\mathbf{z}_k + \mathbf{b}_o)$$

$$\mathbf{z}_k = GRU_d(\mathbf{z}_{k-1}, [\mathbf{m}+\mathbf{u}; \hat{\mathbf{x}}_{k-1}]).$$

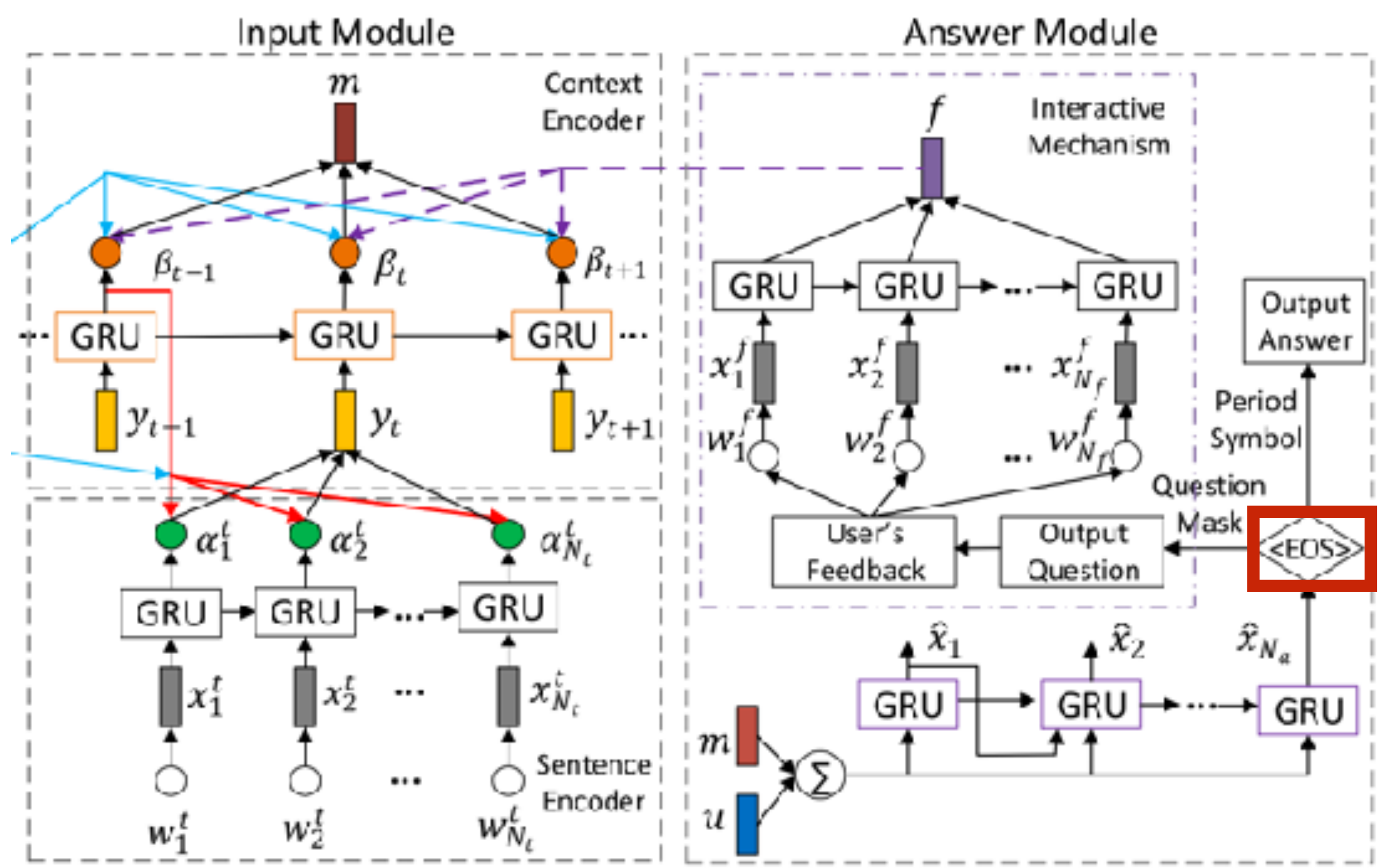


Context-aware Attention Network

- Answer Module:

  - Output Choices

"*The Sentence generated by the decoder ends with a special symbol, either a* ***question mask*** *or a* ***period symbol****.* "



Context-aware Attention Network

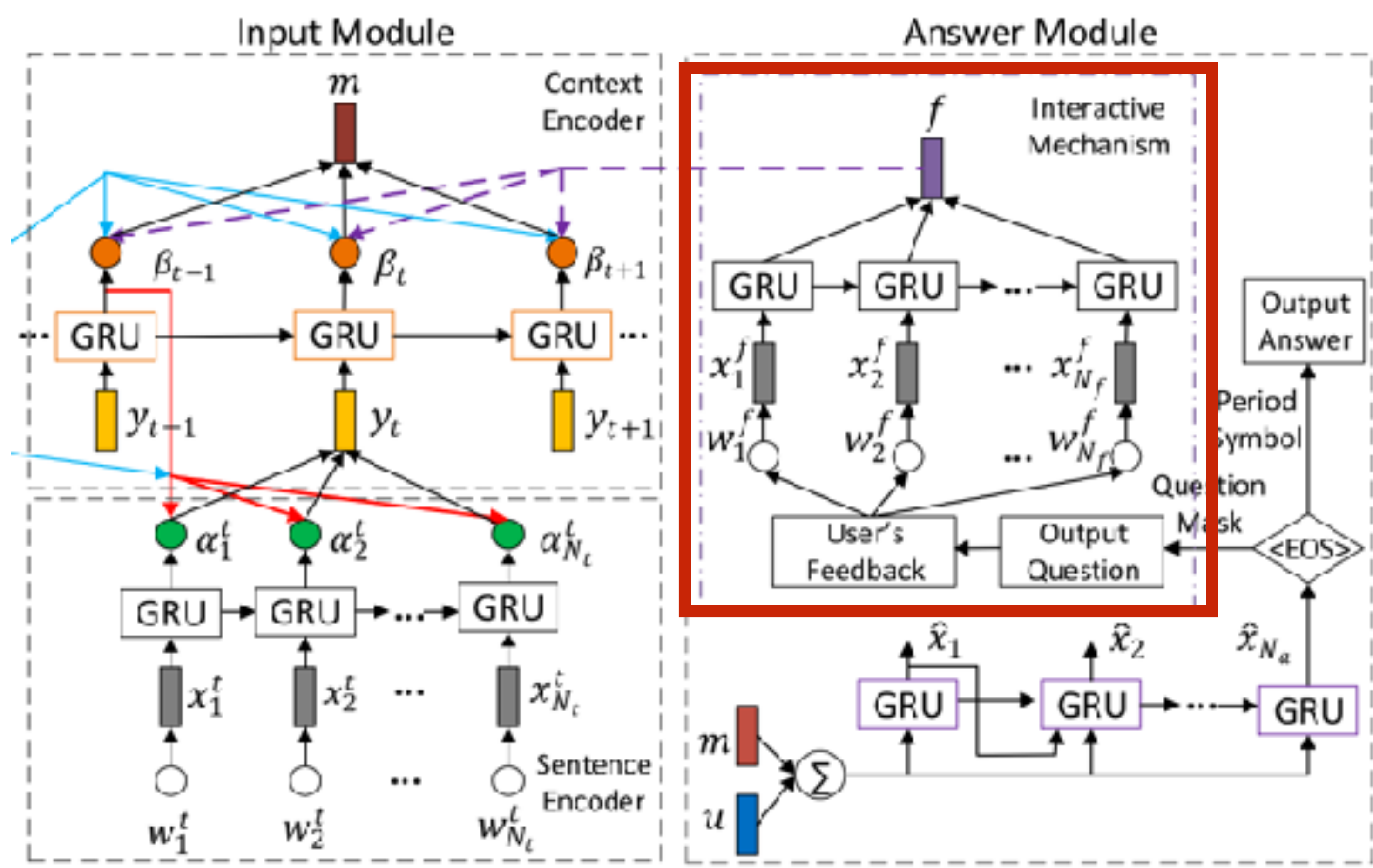
- Answer Module:

  - Interactive Mechanism

    (1) Generate a supplementary question;

    (2) User provide a feedback;

    (3) The feedback is used for answer prediction;



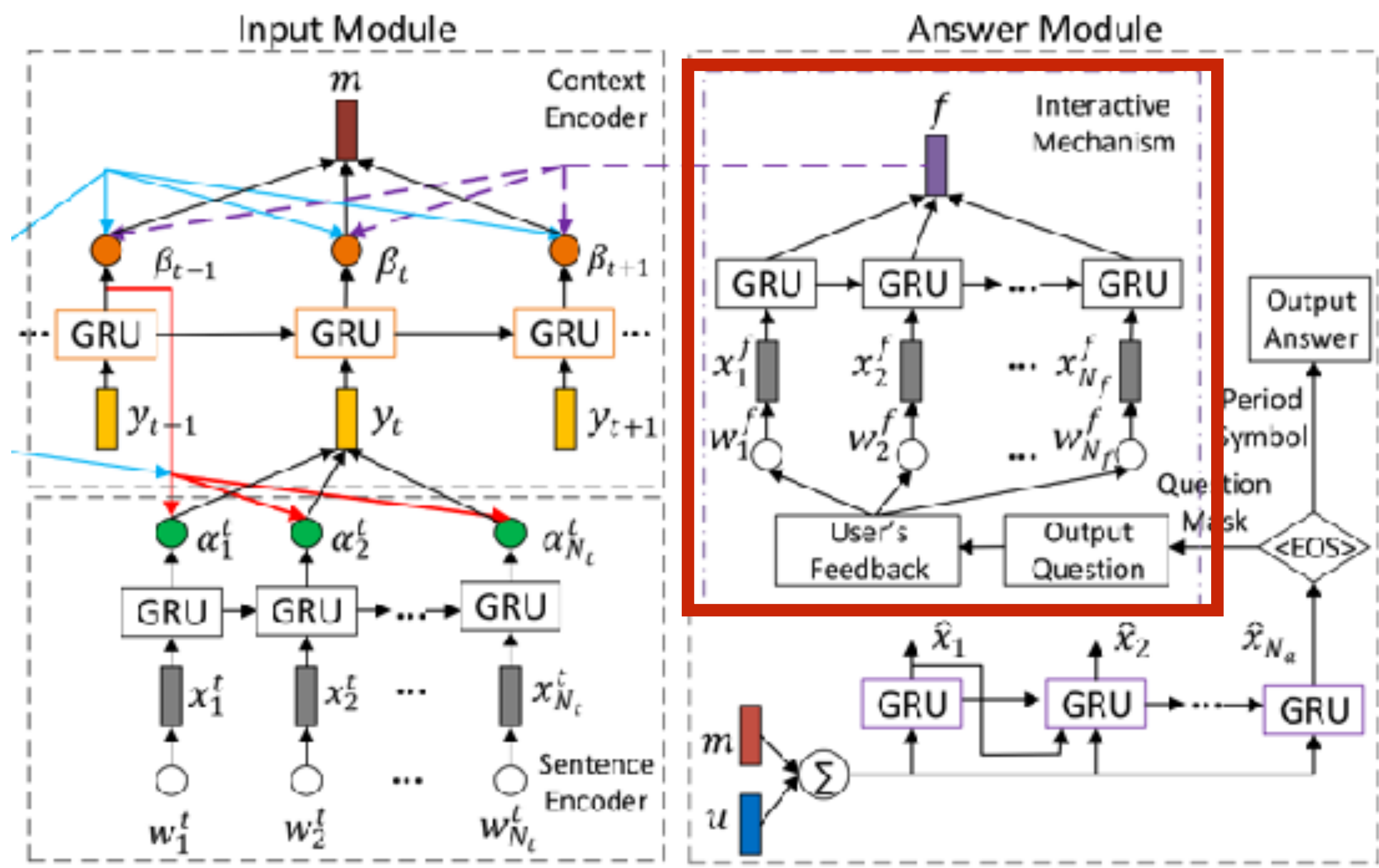
Context-aware Attention Network

- Answer Module:

  - Interactive Mechanism

$$\mathbf{g}_d^f = GRU_w(\mathbf{g}_{d-1}^f, \mathbf{x}_d^f)$$

$$\mathbf{f} = \frac{1}{N_f}\sum_{d=1}^{N_f}\mathbf{g}_d^f$$



Context-aware Attention Network

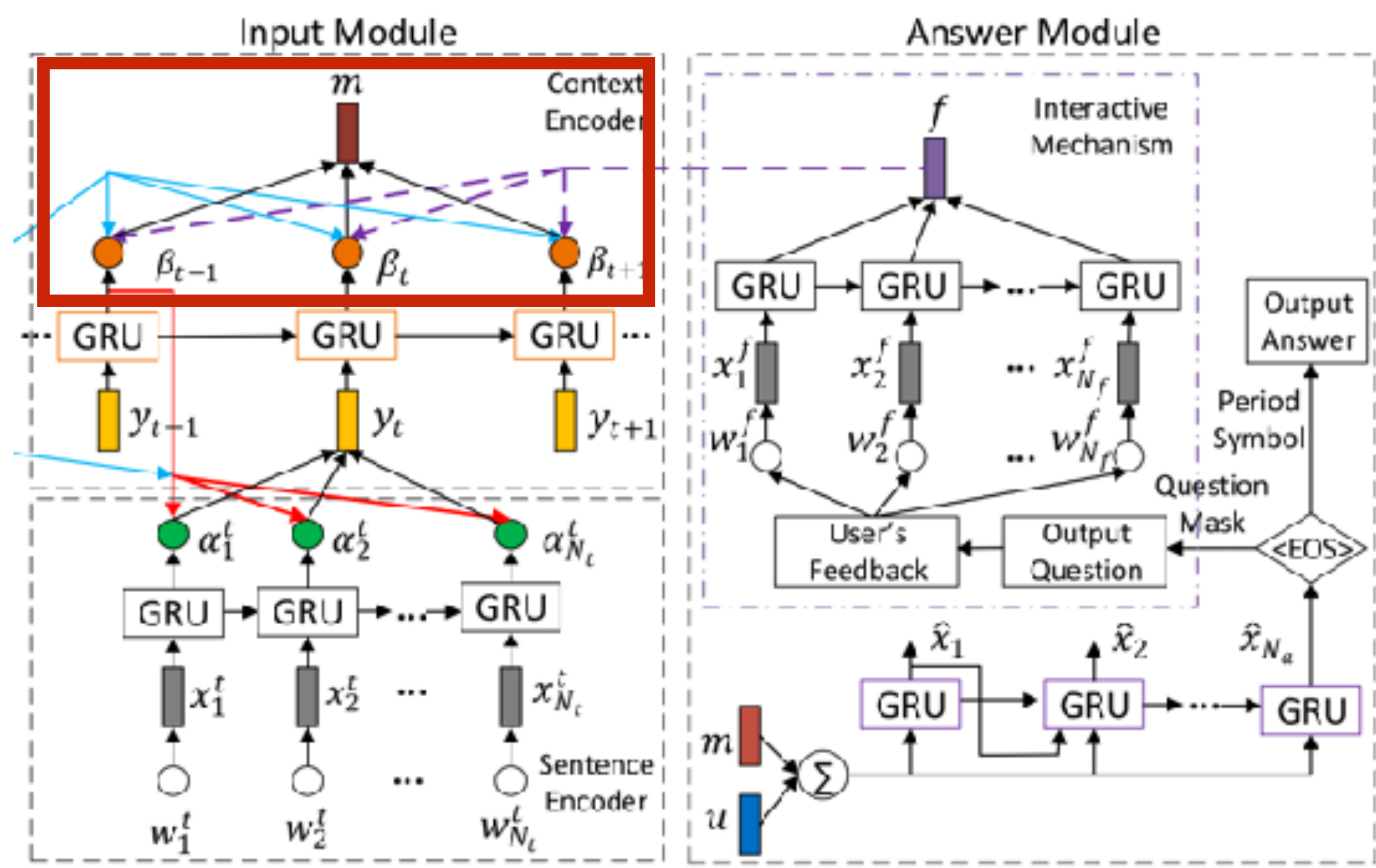
- Answer Module:

  - Interactive Mechanism

$$\mathbf{r} = tanh(\mathbf{W}_{rf}\mathbf{f} + \mathbf{b}_r^{(f)})$$

$$\cancel{\beta_t = softmax(\mathbf{u}^T \mathbf{s}_t)}$$

$$\beta_t = softmax(\mathbf{u}^T \mathbf{s}_t + \mathbf{r}^T \mathbf{s}_t)$$

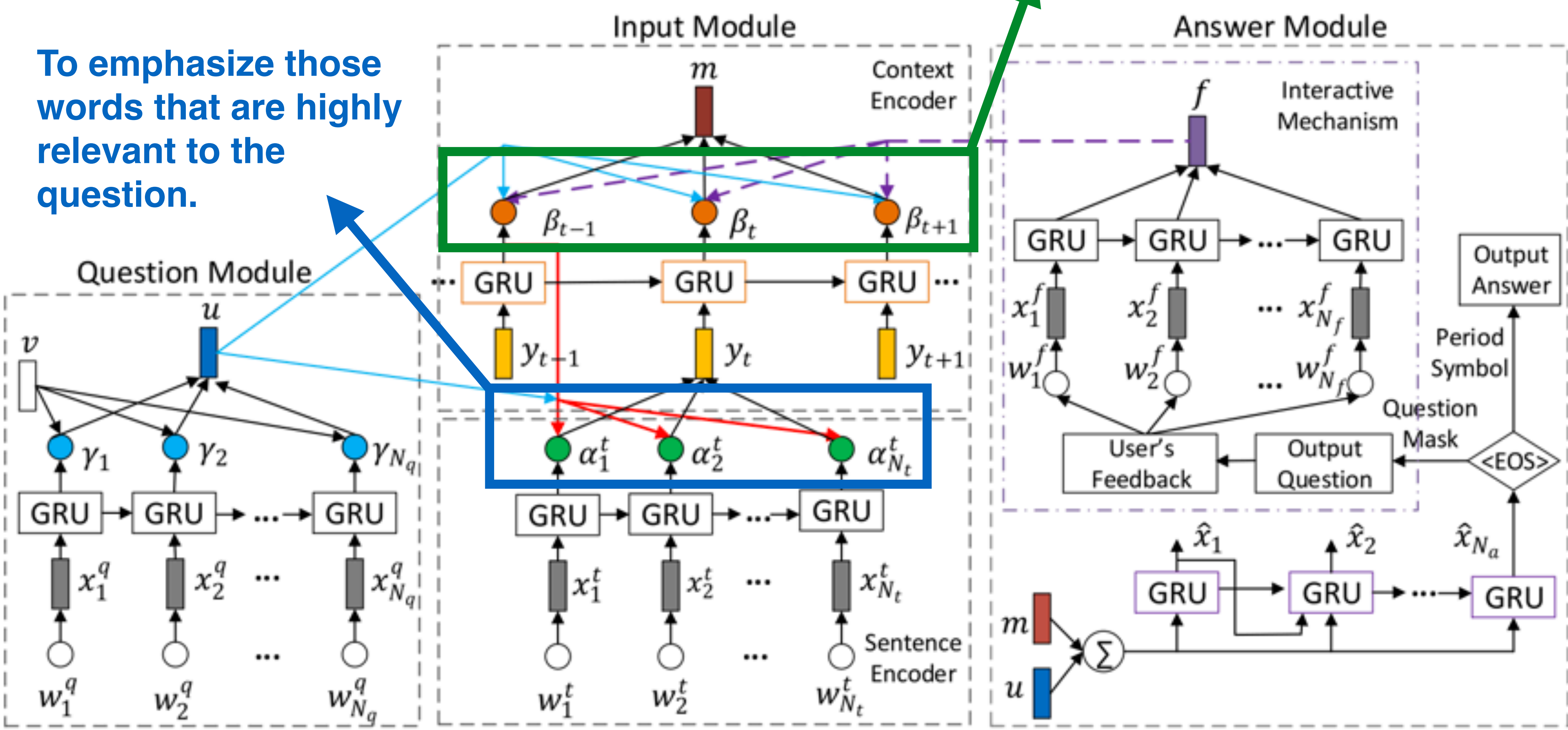


Context-aware Attention Network

- Overall:

A sentence level attention mechanism is enabled to emphasize those sentences that are highly relevant to the question.

To emphasize those words that are highly relevant to the question.



Leveraging Contextual Sentence Relations for Extractive Summarization Using a Neural Attention Model

Pengjie Ren
jay.ren@outlook.com
Shandong University
Jinan, China

Zhumin Chen
chenzhumin@sdu.edu.cn
Shandong University
Jinan, China

Zhaochun Ren
renzhaochun@jd.com
Data Science Lab, JD.com
Beijing, China

Furu Wei
fuwei@microsoft.com
Microsoft Research Asia
Beijing, China

Jun Ma
majun@sdu.edu.cn
Shandong University
Jinan, China

Maarten de Rijke
derijke@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Task

- ES: Aim to generate **a short text summary** for document by **selecting salient sentences** in the document.
- **Sentence scoring**: measure the importance of sentences.
- **Sentence selection**: consider both the importance and redundancy.

Limitation of Related Work

Dataset	Approach	ROUGE-1	ROUGE-2
DUC 2001	*t-SR*	34.82	7.76
	PriorSum	35.98	7.89
	Upper bound	40.82	14.76
DUC 2002	*t-SR*	37.33	8.98
	PriorSum	36.63	8.97
	Upper bound	43.78	15.97
DUC 2004	*t-SR*	37.74	9.60
	PriorSum	38.91	10.07
	Upper bound	41.75	13.73

missing semantic information.

missing contextual relations.

Argue that: sentence importance also depends on contextual relations.

Contextual Relation-based Summarization

- Sentence Scoring:

$$f(S_t \mid \theta) \sim \text{ROUGE-2}(S_t \mid S_{ref})$$

- Sentence Selection:

$$\Psi^* = \arg\max_{\Psi \subseteq D} \sum_{S_t \in \Psi} f(S_t \mid \theta)$$

$$\text{such that} \sum_{S_t \in \Psi} |S_t| \leq l \text{ and } r(\Psi) \text{ hold}$$

Contextual Relation-based Summarization

- Sentence Scoring:

  - Estimate the ability of ***St*** to summarize its **preceding** context:

$$f_{pc}(\mathrm{v}(S_t), \mathrm{v}_{pc}(S_t)) = \cos(\mathrm{v}(S_t), \mathrm{v}_{pc}(S_t))$$

  - Estimate the ability of ***St*** to summarize its **following** context:

$$f_{fc}(\mathrm{v}(S_t), \mathrm{v}_{fc}(S_t)) = \cos(\mathrm{v}(S_t), \mathrm{v}_{fc}(S_t)),$$

Contextual Relation-based Summarization

- Sentence Scoring:

  - **CRSum** + **Surface Features**:

$$f(S_t \mid \theta) = \mathrm{MLP}\left(\begin{bmatrix} f_{pc}(\mathrm{v}(S_t), \mathrm{v}_{pc}(S_t)) \\ f_{fc}(\mathrm{v}(S_t), \mathrm{v}_{fc}(S_t)) \\ \mathrm{v}(S_t) \\ f_{len}(S_t) \\ f_{pos}(S_t) \\ f_{tf}(S_t) \\ f_{df}(S_t) \end{bmatrix}\right)$$

Contextual Relation-based Summarization

- Sentence Scoring:

  - Sentence modeling: $\mathbf{v}(S_t)$

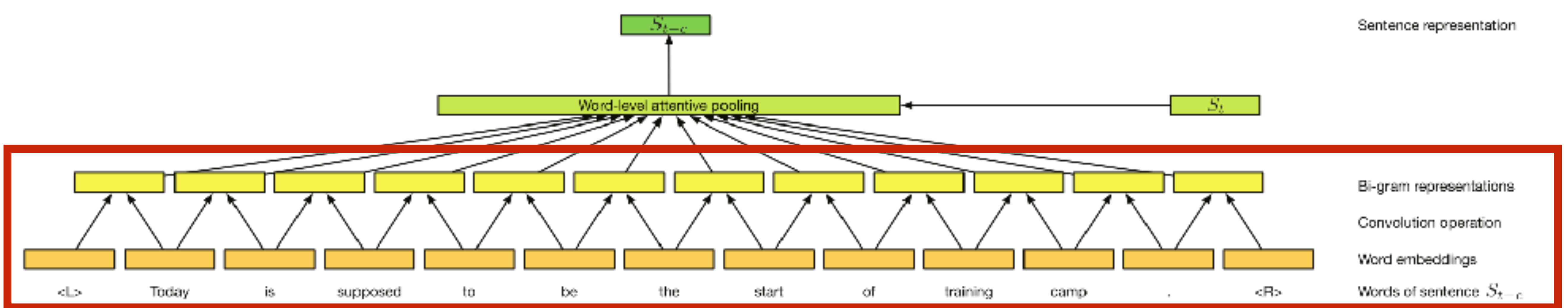


Figure 3: Attentive Pooling Bi-gram Convolutional Neural Network (AP-Bi-CNN) for sentence modeling.

Contextual Relation-based Summarization

- Sentence Scoring:

  - Sentence modeling: $\mathbf{v}(S_t)$

$$\mathrm{bi}(i, i+1) = \begin{bmatrix} \mathrm{v}_i \\ \mathrm{v}_{i+1} \end{bmatrix}$$

$$\mathrm{v}_{bi}(i, i+1) = f(\mathrm{W}_c^T \cdot \mathrm{bi}(i, i+1) + b)$$

$$\mathrm{v}(S_t) = \max_{\mathrm{v}_{bi}(i,i+1) \in V_{bi}(S_t)} \mathrm{v}_{bi}(i, i+1)$$

Contextual Relation-based Summarization

- Sentence Scoring:

    - Sentence modeling: $\mathrm{v}(S_{t-c})$

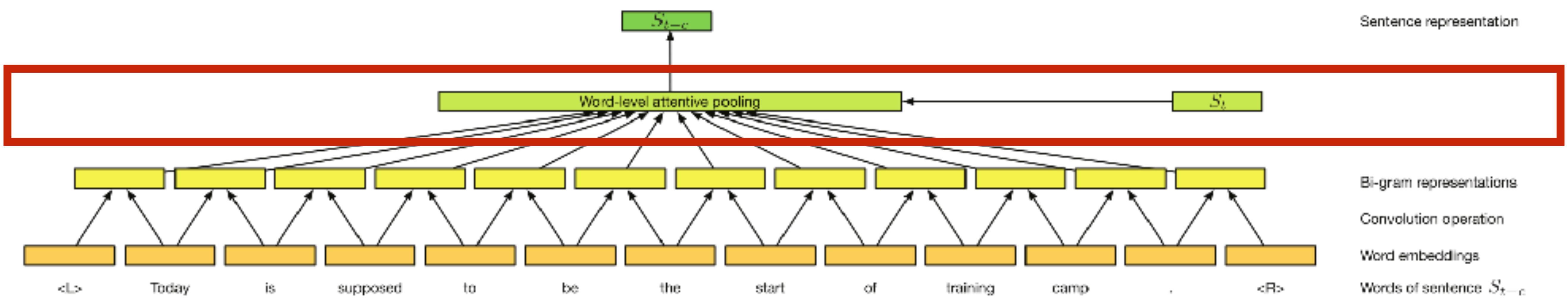


Figure 3: Attentive Pooling Bi-gram Convolutional Neural Network (AP-Bi-CNN) for sentence modeling.

Contextual Relation-based Summarization

- Sentence Scoring:

  - Sentence modeling:

$$v(S_{t-c}) = \max_{v_{bi}(i,i+1) \in V_{bi}(S_{t-c})} w_{bi}(i, i+1) \cdot v_{bi}(i, i+1)$$

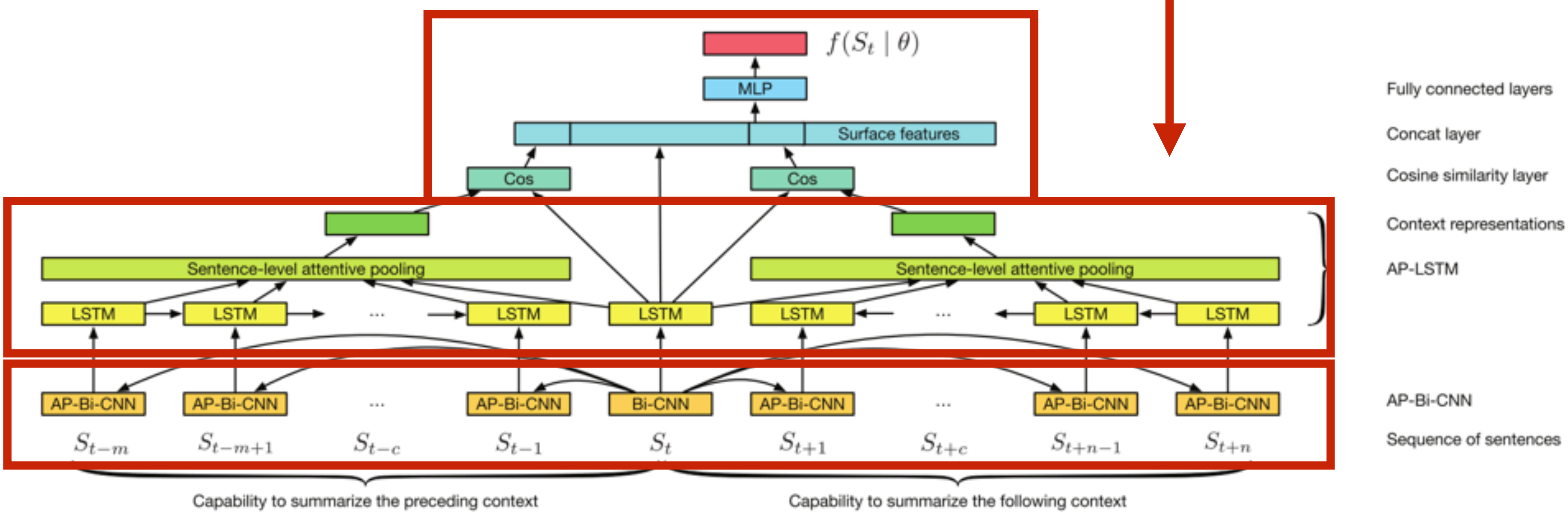
$$\begin{bmatrix} w_{bi}(0,1) \\ \vdots \\ w_{bi}(i, i+1) \\ \vdots \\ w_{bi}(|S_{t-c}|, |S_{t-c}+1|) \end{bmatrix} = \mathrm{softmax}\left(\begin{bmatrix} \cos(v_{bi}(0,1), v(S_t)) \\ \vdots \\ \cos(v_{bi}(i, i+1), v(S_t)) \\ \vdots \\ \cos(v_{bi}(|S_{t-c}|, |S_{t-c}+1|), v(S_t)) \end{bmatrix}\right)$$

Contextual Relation-based Summarization

- Sentence Scoring:

  - Context modeling:

Use sentence relations to learning the pooling weights for Attention module.



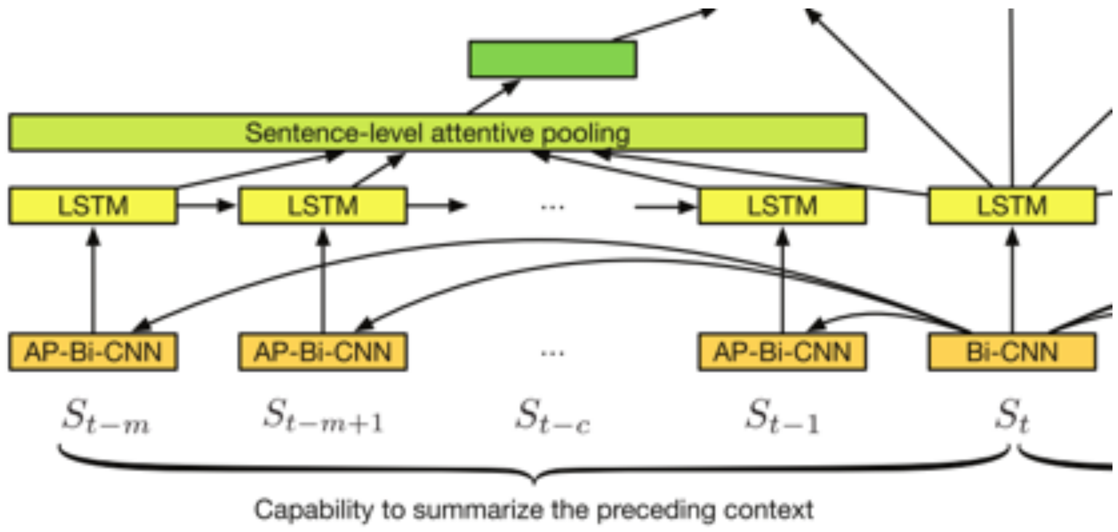
Contextual Relation-based Summarization

- Sentence Scoring:

  - Context modeling:

$$v_{pc}(S_t) = \max_{h_{t-i} \in V_{pc}} w_{t-i} \cdot h_{t-i}$$

$$\begin{bmatrix} w_{t-m} \\ \vdots \\ w_{t-i} \\ \vdots \\ w_{t-1} \end{bmatrix} = \text{softmax}\left(\begin{bmatrix} \cos(h_{t-m}, h_t) \\ \vdots \\ \cos(h_{t-i}, h_t) \\ \vdots \\ \cos(h_{t-1}, h_t) \end{bmatrix}\right)$$



Contextual Relation-based Summarization

- Sentence Selection:

  - *"We use **Greedy** as the sentence selection algorithm."*

  - In each step, a new sentence *St* is added to the summary, when:

(1) It has the highest score in the remaining sentences;
(2) $\frac{\textit{bi-gram-overlap}(S_t, \Psi)}{f_{len}(S_t)} \leq 1 - \lambda$, where $\textit{bi-gram-overlap}(S_t, \Psi)$ is the count of bi-gram overlap between sentence S_t and the current summary Ψ.

Interpretable Convolutional Neural Networks with Dual Local and Global Attention for Review Rating Prediction

Sungyong Seo
University of Southern California
sungyons@usc.edu

Jing Huang
Visa Research, Visa Inc.
jinhuang@visa.com

Hao Yang
Visa Research, Visa Inc.
haoyang@visa.com

Yan Liu
University of Southern California
yanliu.cs@usc.edu

Task

- RP: **Predict the rating** of a user to a new item that has not been rated by the user.
- RRP: Review rating prediction.

Limitation of Related Work

- Cold start problem.
- Content ignorance.

  Argue that:
  Using review text is one approach to alleviate the above issues.

- Cannot be interpretable.

  Argue that:
  Attention layers give us the ability to interpret what model is doing.

Dual Attention-based Model

- Local Attention-based Module (L-Attn):

  - Learn rep. of **local information keywords** which provides us insight on a user's preferences or an item's properties.

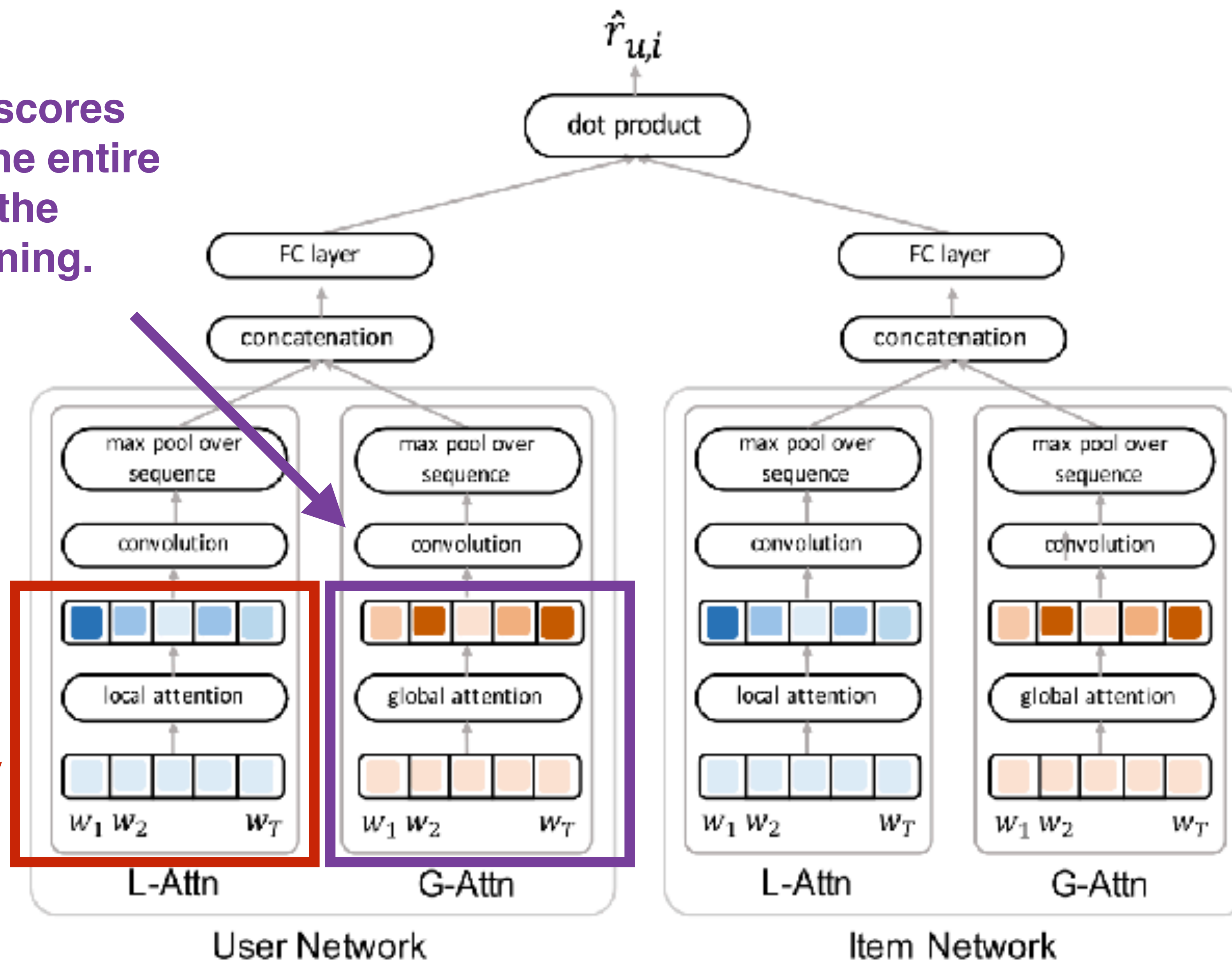
- Global Attention-based Module (G-Atten):

  - Learn rep. from the **original review word sequences** which focuses on the semantic meaning of the whole review text.

Dual Attention-based Model

The global attention scores are computed from the entire input text to capture the global semantic meaning.

The local attention selects informative keywords from a local window.



Dual Attention-based Model

- Local Attention-based Module (L-Attn):

$$\mathbf{X}_{l-att,i} = (\mathbf{x}_{i+\frac{-w+1}{2}}, \mathbf{x}_{i+\frac{-w+3}{2}}, \cdots, \mathbf{x}_i, \cdots, \mathbf{x}_{i+\frac{w-1}{2}})^\top.$$

$$\mathbf{s}(i) = g(\mathbf{X}_{l-att,i} * \mathbf{W}^1_{l-att} + b^1_{l-att}), \qquad i \in [1, T]$$

$$\hat{\mathbf{x}}^L_t = \mathbf{s}(t)\mathbf{x}_t$$

$$\mathbf{Z}_{l-att}(t,i) = g(\hat{\mathbf{x}}^L_t * \mathbf{W}^2_{l-att}(:,i) + \mathbf{b}^2_{l-att}(i)) \qquad i \in [1, n_{l-att}]$$

$$\mathbf{z}_{l-att}(i) = \textsc{Max}(\mathbf{Z}_{l-att}(:,i))$$

Dual Attention-based Model

- Global Attention-based Module (G-Attn):

$$\hat{\mathbf{X}}_{g-att,i} = (\hat{\mathbf{x}}_i^G, \hat{\mathbf{x}}_{i+1}^G, \cdots, \hat{\mathbf{x}}_{i+w_f-1}^G)^\top$$

$$\mathbf{Z}_{g-att}(i,j) = g(\hat{\mathbf{X}}_{g-att,i} * \mathbf{W}_{g-att}(:,:,j) + \mathbf{b}_{g-att}(j))$$

$$i \in [1, T - w_f + 1], \quad j \in [1, n_{g-att}]$$

$$\mathbf{z}_{g-att}(j) = \mathrm{Max}\left(\mathbf{Z}_{g-att}(:,j)\right)$$

Conclusion

- **Perform tasks more accurately**.
- **Better-learned representation**: context-dependent attention.
- **Interpretable representation**: *“Aiming to move a step further from machine learning to machine reasoning.”*

Thanks

Xiao i - Chen Lu
School of Computer and Software Engineering - ICA