

# 阅读理解中的数值推理

—陈豪

- ▶ Drop Dataset
- ▶ Paper: NumNet
- ▶ Paper: QDGAT

## ► Drop Dataset

- DROP 数据集由 AI2 实验室2019年提出
- 涉及到数学运算问题的处理能力的数据集
- 数据集从维基百科中提取段落、通过众包生成问答对
- 训练集中约有77k个问题，开发集中约有9.5k个问题

## Drop:Question analysis

| Reasoning                               | Passage (some parts shortened)  | Question   | Answer       | BiDAF          |
|---|---|--|--------------|----------------|
| Subtraction<br>(28.8%)                  | That year, his <b>Untitled (1981)</b> , a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was <b>sold by Robert Lehrman for \$16.3 million, well above its \$12 million high estimate.</b>   | How many more dollars was the <b>Untitled (1981)</b> painting sold for than the 12 million dollar estimation?      | 4300000      | \$16.3 million |
| Comparison<br>(18.2%)                   | In <b>1517, the seventeen-year-old King sailed to Castile.</b> There, his Flemish court .... <b>In May 1518, Charles traveled to Barcelona in Aragon.</b>   | Where did Charles travel to first, Castile or Barcelona?   | Castile      | Aragon         |
| Selection<br>(19.4%)                    | In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, <b>Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack</b> to tell the story of the events that led up to the battle.  | Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?           | Don Mueller  | Baker          |
| Addition<br>(11.7%)                     | Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, located in a pink zone near Šibenik, and captured the village at 4:45 p.m. on <b>2 March 1992.</b> The JNA formed a battlegroup to counterattack the <b>next day.</b>  | What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?             | 3 March 1992 | 2 March 1992   |
| Count<br>(16.5%)<br>and Sort<br>(11.7%) | Denver would retake the lead with kicker <b>Matt Prater nailing a 43-yard field goal</b> , yet Carolina answered as kicker <b>John Kasay ties the game with a 39-yard field goal.</b> ... Carolina closed out the half with <b>Kasay nailing a 44-yard field goal.</b> ... In the fourth quarter, Carolina sealed the win with <b>Kasay's 42-yard field goal.</b> | Which kicker kicked the most field goals?  | John Kasay   | Matt Prater    |
| Coreference<br>Resolution<br>(3.7%)     | <b>James Douglas</b> was the second son of Sir George Douglas of Pittendreich, and Elizabeth Douglas, daughter David Douglas of Pittendreich. Before <b>1543 he married Elizabeth</b> , daughter of James Douglas, 3rd Earl of Morton. <b>In 1553 James Douglas succeeded to the title and estates of his father-in-law.</b>                                      | How many years after he married Elizabeth did James Douglas succeed to the title and estates of his father-in-law? | 10           | 1553           |



(a) For span type answers



(b) For number type answers

Figure 1: Distribution of the most popular question prefixes for two different subsets of the training data.



## ► Drop:Question analysis

| Reasoning                               | Passage (some parts shortened)  | Question   | Answer       | BiDAF          |
|---|---|--|--------------|----------------|
| Subtraction<br>(28.8%)                  | That year, his <b>Untitled (1981)</b> , a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was <b>sold by Robert Lehrman for \$16.3 million, well above its \$12 million high estimate.</b>   | How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?             | 4300000      | \$16.3 million |
| Comparison<br>(18.2%)                   | In <b>1517, the seventeen-year-old King sailed to Castile.</b> There, his Flemish court . . . . <b>In May 1518, Charles traveled to Barcelona in Aragon.</b>  | Where did Charles travel to first, Castile or Barcelona?   | Castile      | Aragon         |
| Selection<br>(19.4%)                    | In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, <b>Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack</b> to tell the story of the events that led up to the battle.  | Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?           | Don Mueller  | Baker          |
| Addition<br>(11.7%)                     | Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, located in a pink zone near Šibenik, and captured the village at 4:45 p.m. on <b>2 March 1992.</b> The JNA formed a battlegroup to counterattack the <b>next day.</b>  | What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?             | 3 March 1992 | 2 March 1992   |
| Count<br>(16.5%)<br>and Sort<br>(11.7%) | Denver would retake the lead with kicker <b>Matt Prater nailing a 43-yard field goal</b> , yet Carolina answered as kicker <b>John Kasay ties the game with a 39-yard field goal.</b> . . . Carolina closed out the half with <b>Kasay nailing a 44-yard field goal.</b> . . . In the fourth quarter, Carolina sealed the win with <b>Kasay's 42-yard field goal.</b> | Which kicker kicked the most field goals?  | John Kasay   | Matt Prater    |
| Coreference<br>Resolution<br>(3.7%)     | <b>James Douglas</b> was the second son of Sir George Douglas of Pittendreich, and Elizabeth Douglas, daughter David Douglas of Pittendreich. Before <b>1543 he married Elizabeth</b> , daughter of James Douglas, 3rd Earl of Morton. <b>In 1553 James Douglas succeeded to the title and estates of his father-in-law.</b>  | How many years after he married Elizabeth did James Douglas succeed to the title and estates of his father-in-law? | 10           | 1553           |

► Drop:Answer analysis

| Answer Type    | Percent | Example             |
|----------------|---------|---------------------|
| NUMBER         | 66.1    | 12                  |
| PERSON         | 12.2    | Jerry Porter        |
| OTHER          | 9.4     | males               |
| OTHER ENTITIES | 7.3     | Seahawks            |
| VERB PHRASE    | 3.5     | Tom arrived at Acre |
| DATE           | 1.5     | 3 March 1992        |

Table 3: Distribution of answer types in training set, according to an automatic named entity recognition.

- 回答一个问题平均需要考虑2.18个span
- span之间的平均距离是26个单词
- 20%的样本需要至少3个跨度
- 大部分的答案是数值和专有名词

# EMNLP 2019

## **NumNet: Machine Reading Comprehension with Numerical Reasoning**

**Qiu Ran<sup>1\*</sup>, Yankai Lin<sup>1\*</sup>, Peng Li<sup>1</sup>, Jie Zhou<sup>1</sup>, Zhiyuan Liu<sup>2</sup>**

<sup>1</sup>Pattern Recognition Center, WeChat AI, Tencent Inc, China

<sup>2</sup>Department of Computer Science and Technology, Tsinghua University, Beijing, China

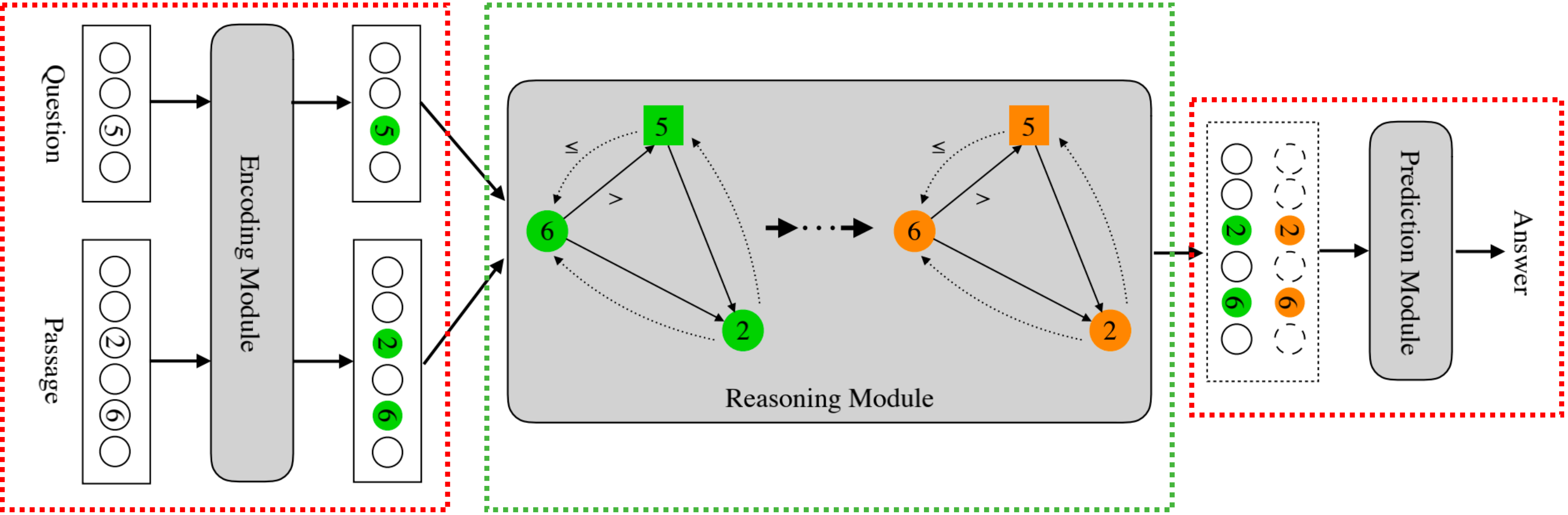
Institute for Artificial Intelligence, Tsinghua University, Beijing, China

State Key Lab on Intelligent Technology and Systems, Tsinghua University, Beijing, China

{soulcaptran, yankailin, patrickpli, withtomzhou}@tencent.com

liuzy@tsinghua.edu.cn

► NumNet Model:



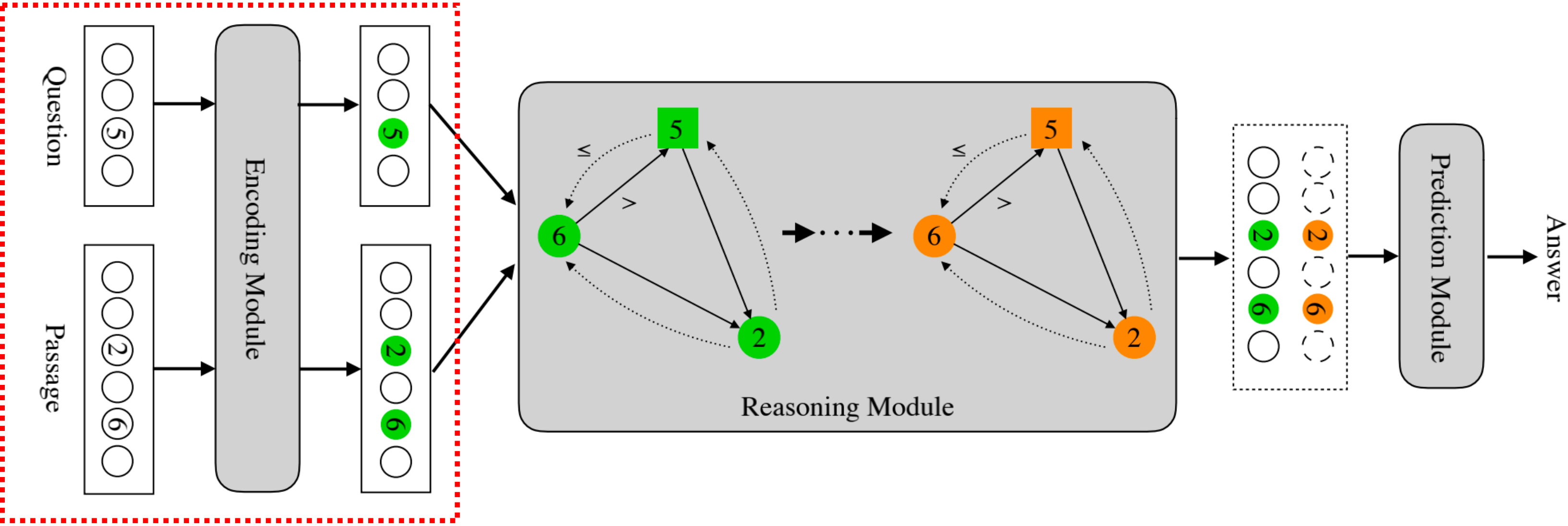


► Example:

| Question  | Passage  | Answer  |
|---|--|---------|
| What is the second longest field goal made?                 | ... The Seahawks immediately trailed on a scoring rally by the Raiders with kicker <i>Sebastian Janikowski nailing a 31-yard field goal</i> ... Then in the third quarter <i>Janikowski made a 36-yard field goal</i> . Then <i>he made a 22-yard field goal</i> in the fourth quarter to put the Raiders up 16-0 ... The Seahawks would make their only score of the game with kicker <i>Olindo Mare hitting a 47-yard field goal</i> . However, they continued to trail as <i>Janikowski made a 49-yard field goal</i> , followed by RB Michael Bush making a 4-yard TD run. | 47-yard |
| How many age groups made up more than 7% of the population? | Of Saratoga Countys population in 2010, <i>6.3%</i> were between ages of 5 and 9 years, <i>6.7%</i> between 10 and 14 years, 6.5% between 15 and 19 years, <i>5.5%</i> between 20 and 24 years, <i>5.5%</i> between 25 and 29 years, <i>5.8%</i> between 30 and 34 years, <i>6.6%</i> between 35 and 39 years, <i>7.9%</i> between 40 and 44 years, <i>8.5%</i> between 45 and 49 years, <i>8.0%</i> between 50 and 54 years, <i>7.0%</i> between 55 and 59 years, <i>6.4%</i> between 60 and 64 years, and <i>13.7%</i> of age 65 years and over ...                          | 5       |

Table : Example questions from the DROP dataset which require numerical comparison

► NumNet Model:



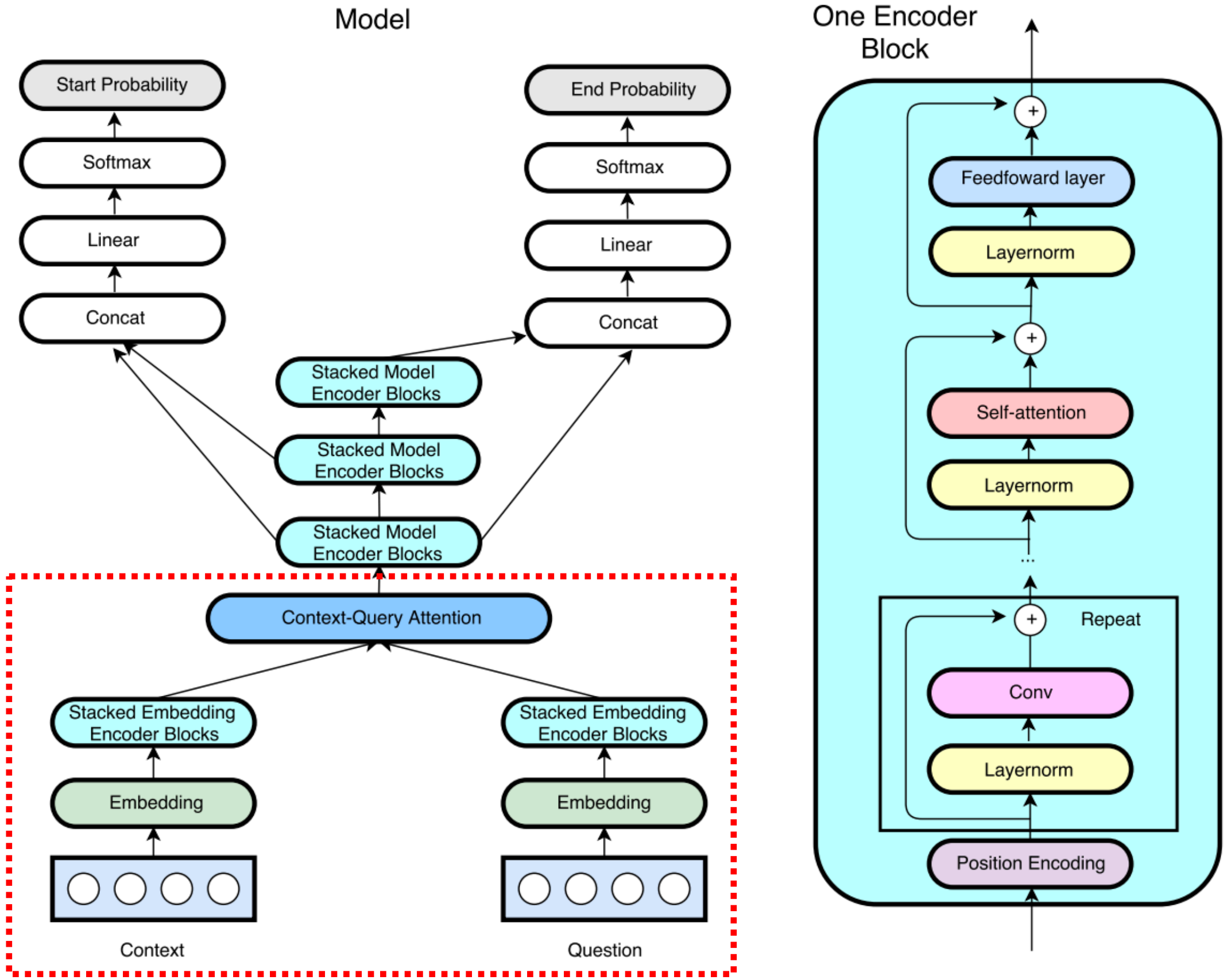
# ► Encoding Module:

$$Q = \text{QANet-Emb-Enc}(Q), \quad (1)$$

$$P = \text{QANet-Emb-Enc}(P), \quad (2)$$

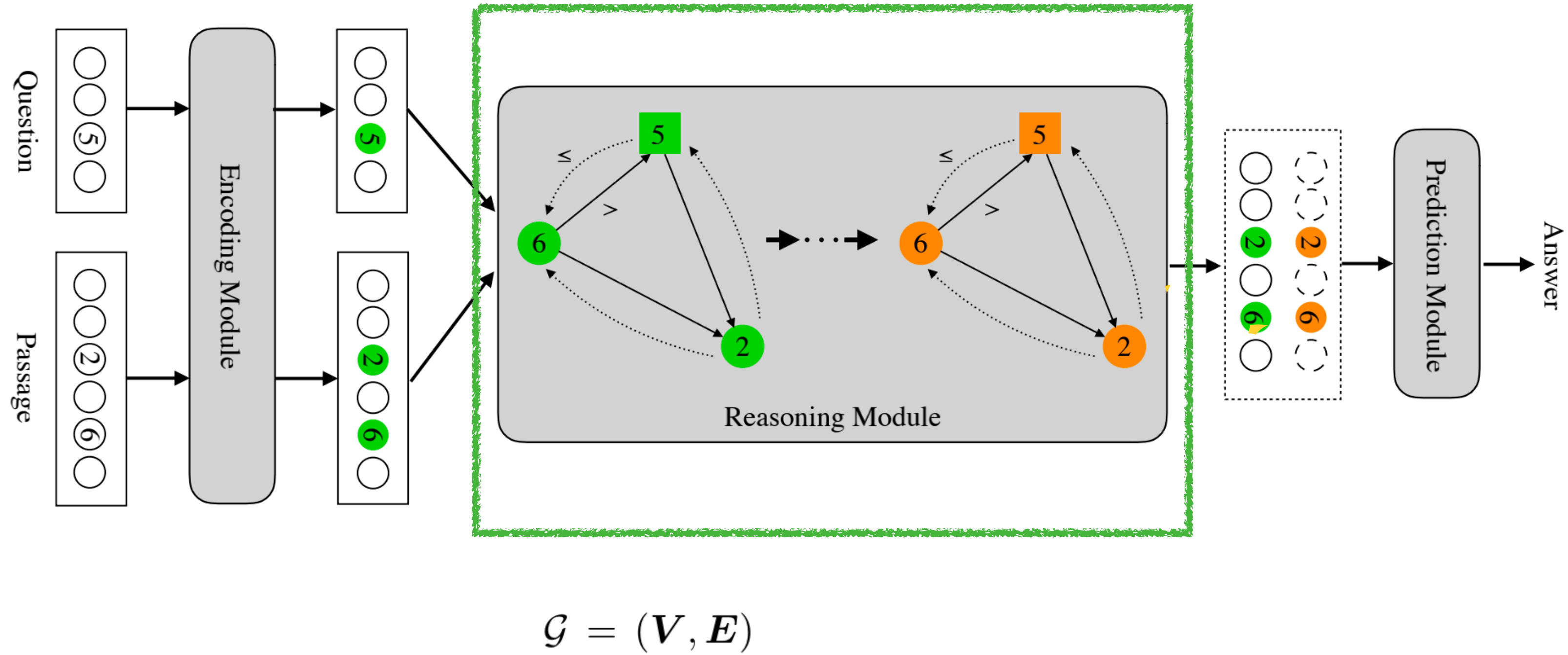
$$\bar{Q} = \text{QANet-Att}(P, Q), \quad (3)$$

$$\bar{P} = \text{QANet-Att}(Q, P), \quad (4)$$



QANet: Combining local convolution with global self-attention for reading comprehension.(ACLR2018)

► Reasoning Module:





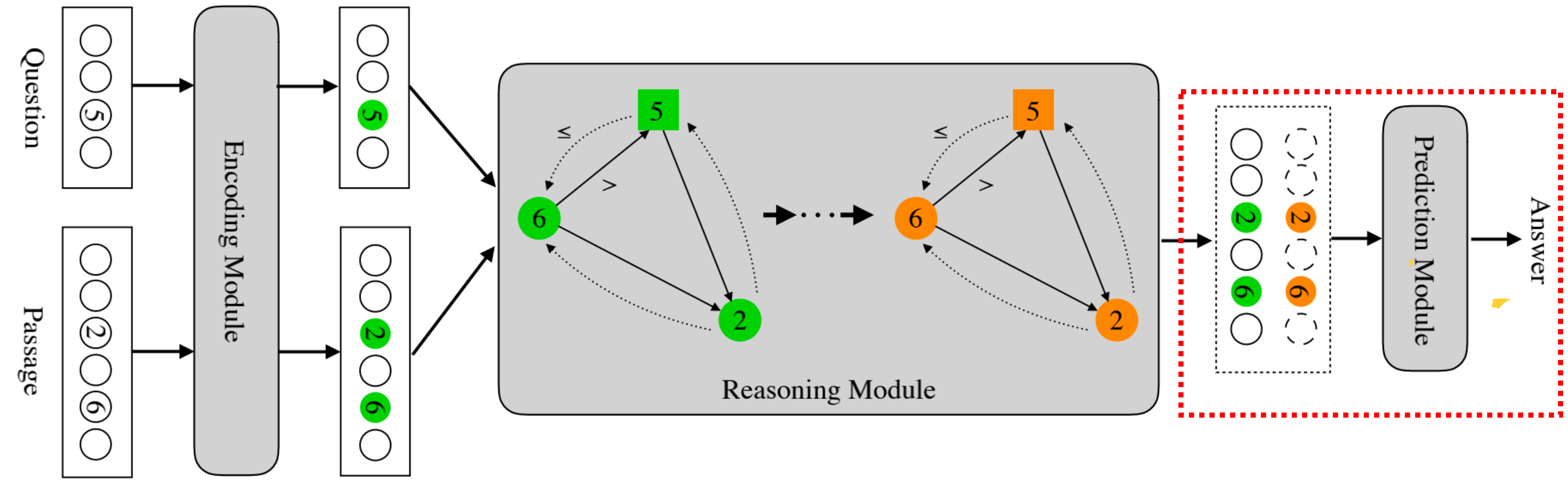
## ► Numerical Reasoning:

Node Relatedness Measure:  $\alpha_i = \text{sigmoid}(\mathbf{W}_v \mathbf{v}[i] + b_v), \quad (10)$

Message Propagation:  $\tilde{\mathbf{v}}'_i = \frac{1}{|\mathcal{N}_i|} \left( \sum_{j \in \mathcal{N}_i} \alpha_j \mathbf{W}^{\mathbf{r}_{ji}} \mathbf{v}[j] \right), \quad (11)$

Node Representation Update:  $\mathbf{v}'_i = \text{ReLU}(\mathbf{W}_f \mathbf{v}_i + \tilde{\mathbf{v}}'_i + \mathbf{b}_f), \quad (12)$

# ▶ Prediction Module:



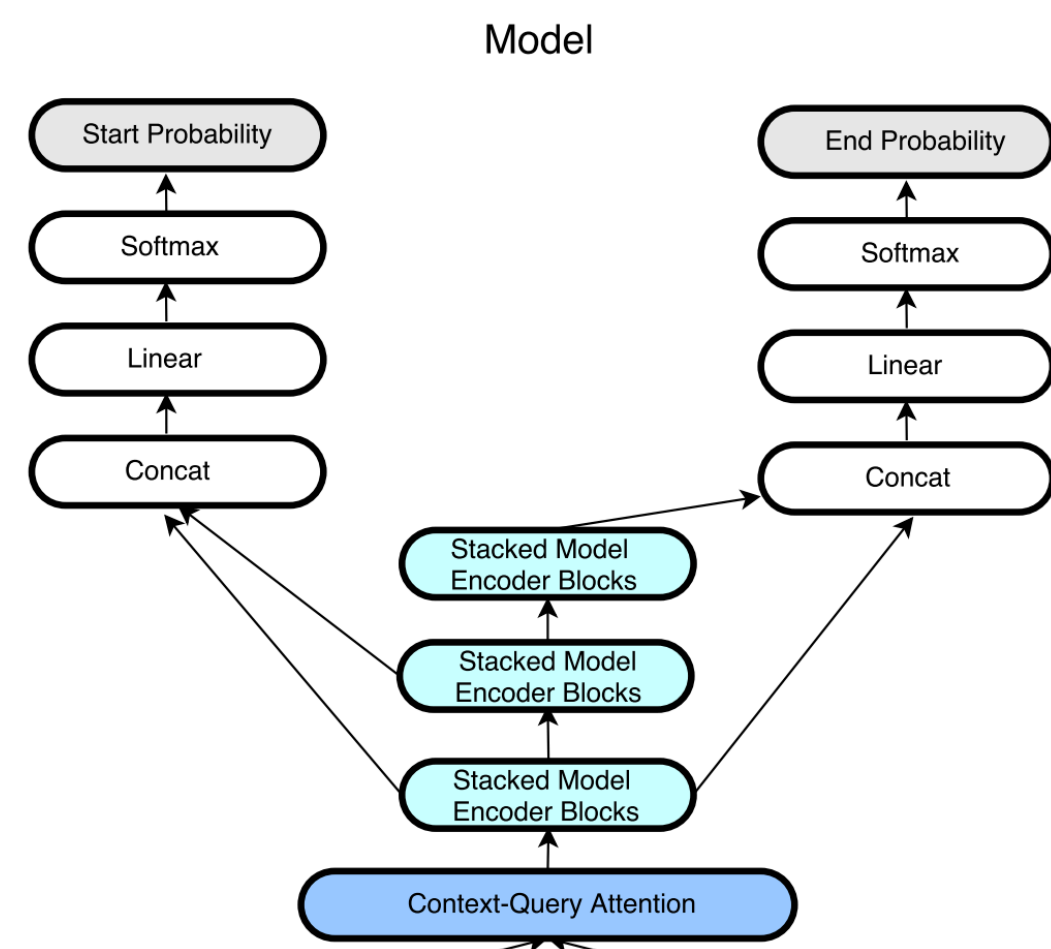
- Passage span
- Question span
- Count
- Arithmetic expression

## ► Prediction Module:

- Passage span

$$\mathbf{p}^{\text{p-start}} = \text{softmax}(\text{FFN}([\mathbf{M}_0; \mathbf{M}_1])), \quad (1)$$

$$\mathbf{p}^{\text{p-end}} = \text{softmax}(\text{FFN}([\mathbf{M}_0; \mathbf{M}_2])) \quad (2)$$



- Question span

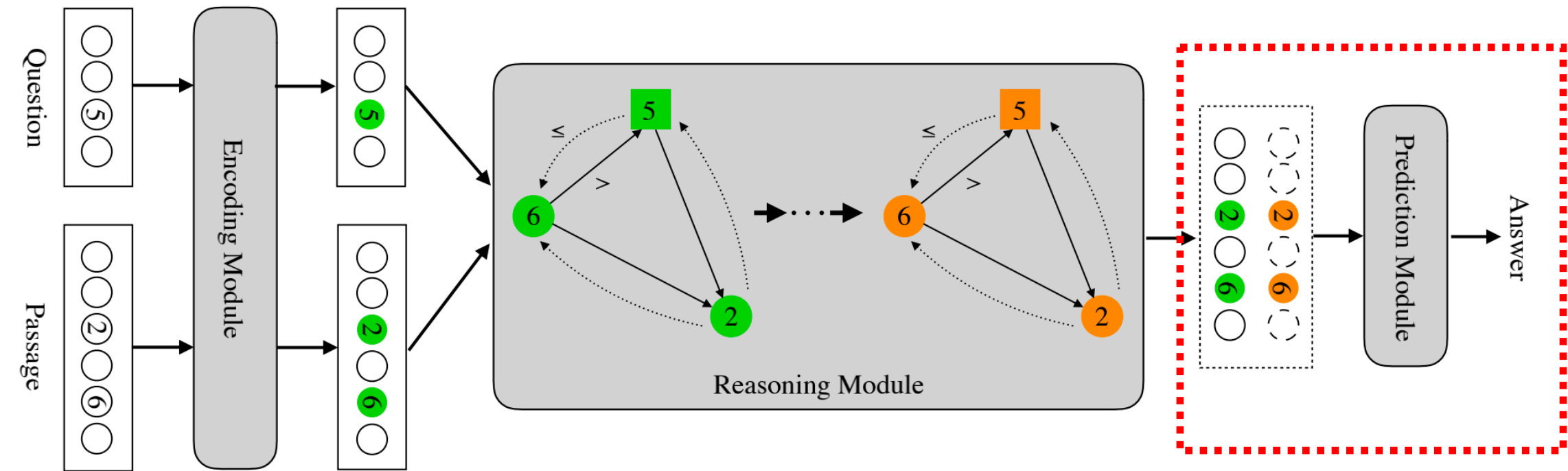
$$\boldsymbol{\alpha}^P = \text{softmax}(\mathbf{W}^P \bar{\mathbf{P}}), \quad (3)$$

$$\mathbf{h}^P = \boldsymbol{\alpha}^P \bar{\mathbf{P}} \quad (4)$$

$$\mathbf{p}^{\text{q-start}} = \text{softmax}(\text{FFN}([\mathbf{Q}; \mathbf{e}^{|Q|} \otimes \mathbf{h}^P])), \quad (5)$$

$$\mathbf{p}^{\text{q-end}} = \text{softmax}(\text{FFN}([\mathbf{Q}; \mathbf{e}^{|Q|} \otimes \mathbf{h}^P])) \quad (6)$$

# ▶ Prediction Module:



● Count

$$\mathbf{p}^{\text{count}} = \text{softmax}(\text{FFN}(\mathbf{h}^P)) \tag{7}$$

● Arithmetic expression

$$\mathbf{p}_i^{\text{sign}} = \text{softmax}(\text{FFN}(\mathbf{h}_i^N)) \tag{8}$$



▸ Experiments :

| Method                   | Dev          |              | Test         |              |
|--------------------------|--------------|--------------|--------------|--------------|
|                          | EM           | F1           | EM           | F1           |
| <b>Semantic Parsing</b>  |              |              |              |              |
| Syn Dep                  | 9.38         | 11.64        | 8.51         | 10.84        |
| OpenIE                   | 8.80         | 11.31        | 8.53         | 10.77        |
| SRL                      | 9.28         | 11.72        | 8.98         | 11.45        |
| <b>Traditional MRC</b>   |              |              |              |              |
| BiDAF                    | 26.06        | 28.85        | 24.75        | 27.49        |
| QANet                    | 27.50        | 30.44        | 25.50        | 28.36        |
| BERT                     | 30.10        | 33.36        | 29.45        | 32.70        |
| <b>Numerical MRC</b>     |              |              |              |              |
| NAQANet                  | 46.20        | 49.24        | 44.07        | 47.01        |
| NAQANet+                 | 61.47        | 64.85        | 60.82        | 64.29        |
| <b>NumNet</b>            | <b>64.92</b> | <b>68.31</b> | <b>64.56</b> | <b>67.97</b> |
| <b>Human Performance</b> | -            | -            | 94.09        | 96.42        |

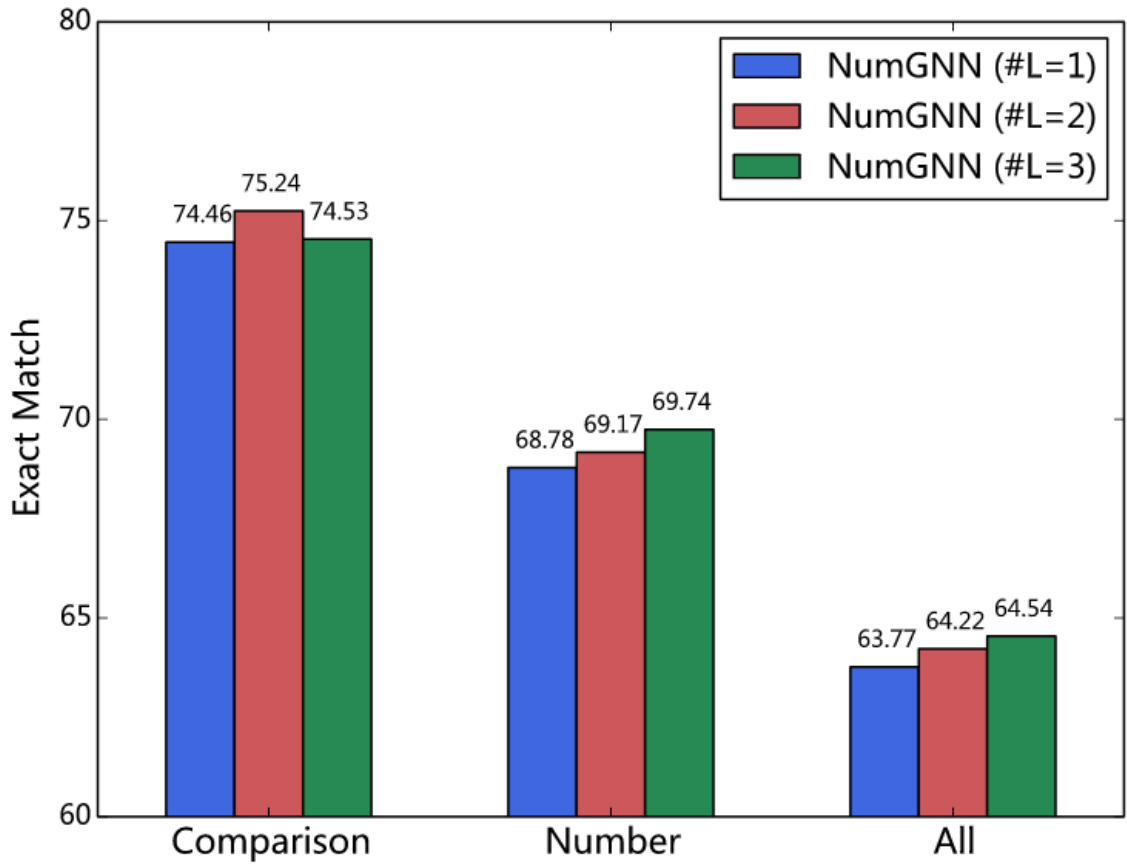





Figure 2: Effect of GNN layer numbers (# L).

# DROP Leaderboard

| Rank  | Submission   | Created  | F1  | Exact Match  |
|--|--|---|--|---|
| 1  | <b>QDGAT - ALBERT</b><br><i>AntGroup KG &amp; NLP</i>                      | 09/08/2020  | 0.9010   | 0.8704  |
| 2  | <b>Numeric Transformer - Albert</b><br><i>OneConnect GammaLab NYC</i>      | 03/17/2020  | 0.8911   | 0.8582  |
| 3  | <b>QDGAT Ensemble</b><br><i>AntGroup KG &amp; NLP</i>                      | 12/16/2019  | 0.8838   | 0.8546  |
| 4  | <b>sna_albert+ Ensemble</b><br><i>OneConnect GammaLab</i>                  | 12/03/2019  | 0.8795   | 0.8494  |
| 5  | <b>QDGAT - RoBERTa</b><br><i>AntGroup KG &amp; NLP</i>                     | 06/01/2020  | 0.8779   | 0.8474  |
| 6  | <b>Numeric Transformer - RoBERTa</b><br><i>OneConnect GammaLab NYC</i>     | 03/03/2020  | 0.8759   | 0.8429  |
| 7  | <b>QDGAT</b><br><i>AntGroup KG &amp; NLP</i>                               | 12/04/2019  | 0.8725   | 0.8412  |
| 8  | <b>NumNet+ v2 Ensemble</b><br><i>WeChat AI (Ronqin Yang, Qiu R...</i>      | 10/14/2019  | 0.8616   | 0.8314  |
| 9  | <b>na_albert+</b><br><i>PingAn Gammalab</i>                                | 11/17/2019  | 0.8534   | 0.8196  |
| 10   | <b>NumNet+ v2</b><br><i>WeChat AI (Rongqin Yang, Qiu ...</i>               | 10/11/2019  | 0.8484   | 0.8152  |
| 11   | <b>TASE - RoBERTa</b><br><i>Elad Segal*, Avia Efrat*, Mor...</i>           | 01/10/2020  | 0.8362   | 0.8042  |
| 12   | <b>ALBERT-Calculator</b><br><i>Daniel Andor, Luheng He, Kent...</i>        | 10/17/2019  | 0.8356   | 0.7985  |
| 13   | <b>NumNet+</b><br><i>WeChat AI (Rongqin Yang, Qiu ...</i>                  | 09/26/2019  | 0.8299   | 0.7936  |
| 14   | <b>BERT-Calculator Ensemble</b><br><i>Daniel Andor. Luheng He. Kent...</i> | 09/04/2019  | 0.8178   | 0.7814  |

# EMNLP 2020

## Question Directed Graph Attention Network for Numerical Reasoning over Text

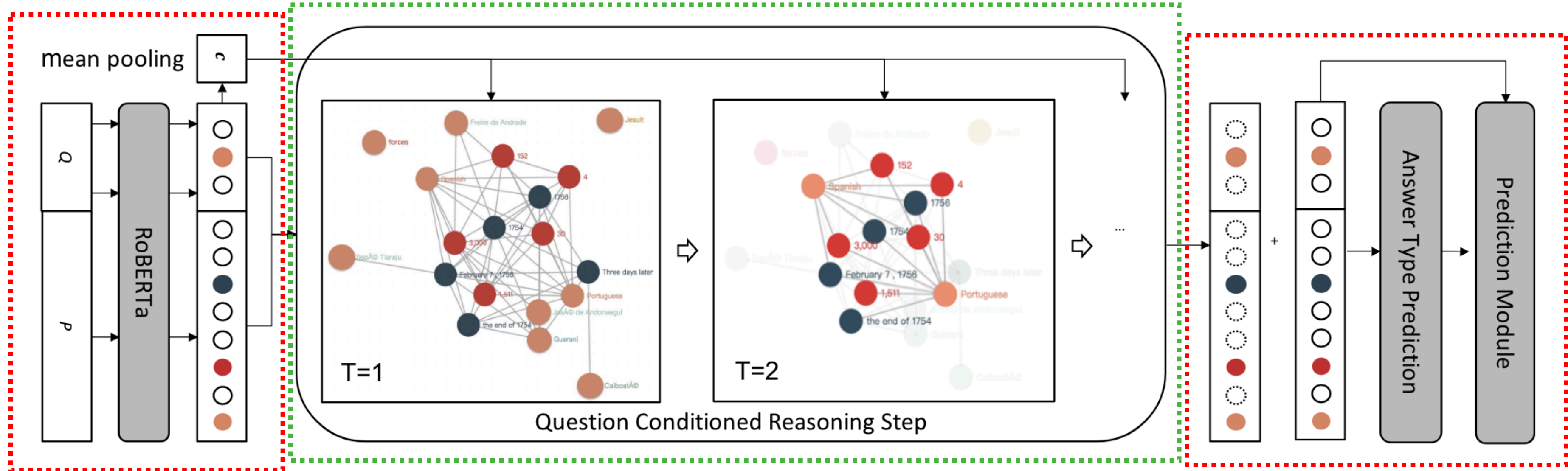
**Kunlong Chen<sup>†</sup>   Weidi Xu<sup>†</sup>   Xingyi Cheng<sup>\*†</sup>**  
**Zou Xiaochuan<sup>†</sup>   Yuyu Zhang<sup>§</sup>   Le Song<sup>†§</sup>**  
**Taifeng Wang<sup>†</sup>   Yuan Qi<sup>†</sup>   Wei Chu<sup>†</sup>**

<sup>†</sup> Ant Group

{kunlong.ckl, weidi.xwd, fanyin.cxy, xiaochuan.zxc,  
le.song, taifeng.wang, weichu.cw, yuan.qi}@antgroup.com

<sup>§</sup> College of Computing Georgia Institute of Technology  
{yuyu}@gatech.edu

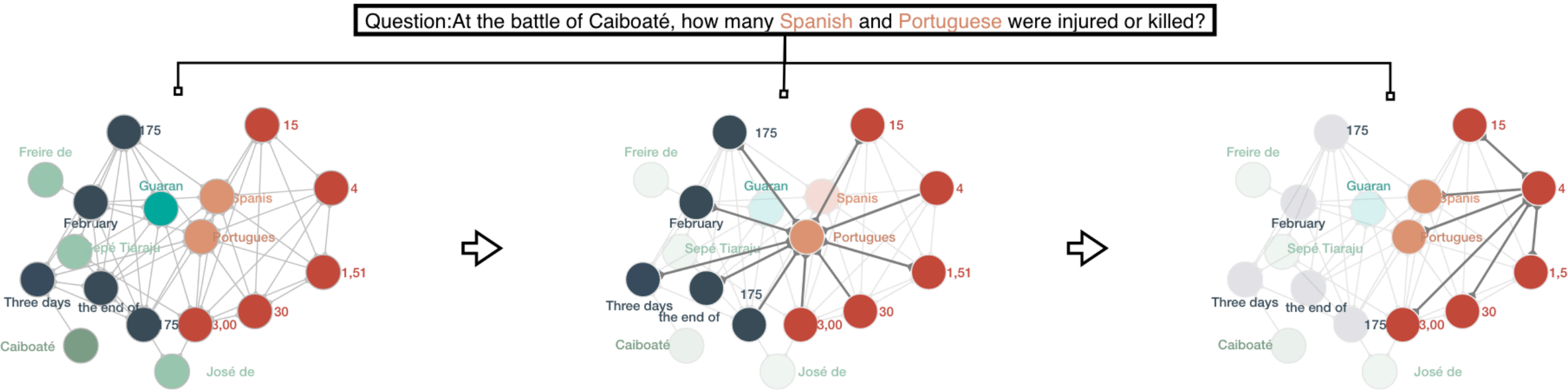
► QDGAT Model:



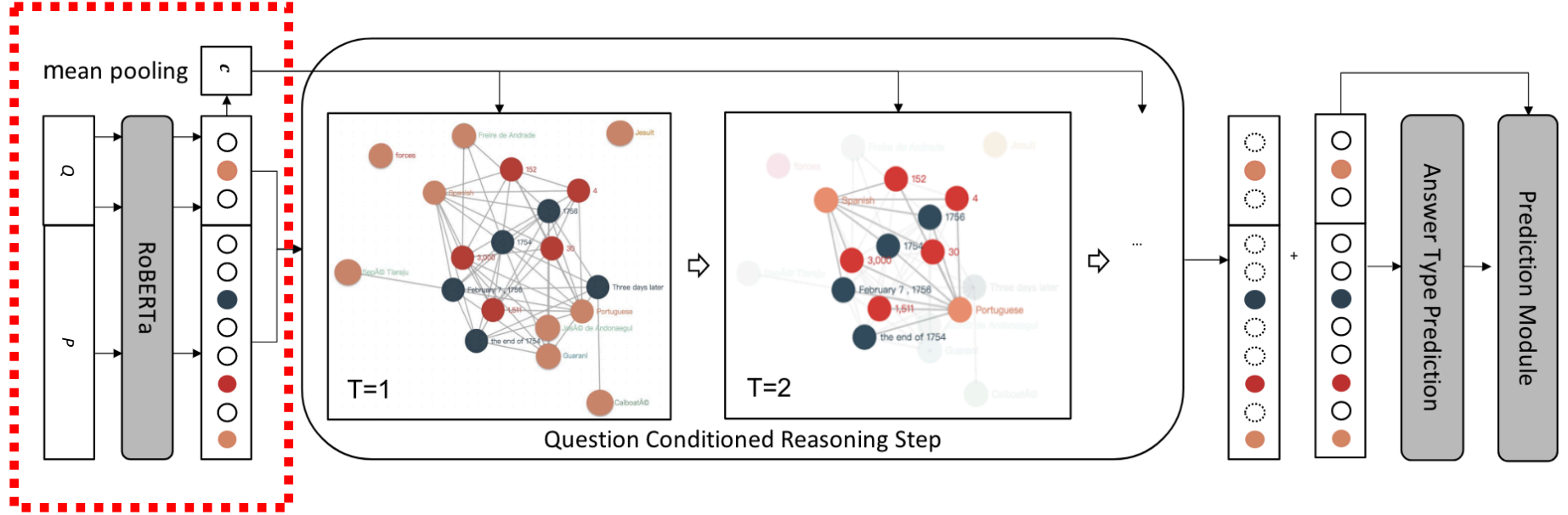


► Example:

| Question   | Passage   | Answer |
|--|---|--------|
| At the battle of CaiboatÃl how many Spanish and Portuguese were injured or killed? | ... In 1754 Spanish and Portuguese military forces were dispatched to force the Guarani to leave the area ... Hostilities resumed in 1756 when an army of 3,000 Spanish, Portuguese, and native auxiliary soldiers under JosÃl de Andonaegui and Freire de Andrade was sent to subdue the Guarani rebels. On February 7, 1756 the leader of the Guarani rebels, SepÃl Tiaraju, was killed in a skirmish with Spanish and Portuguese troops. ... 1,511 Guarani were killed and 152 taken prisoner, while 4 Spanish and Portuguese were killed and about 30 were wounded... | 34     |
| In which quarter did Stephen Gostkowski kick his shortest field goal of the game?  | The Cardinals' east coast struggles continued in the second quarter as quarterback Matt Cassel completed a 15-yard touchdown pass to running back Kevin Faulk and an 11-yard touchdown pass to wide receiver Wes Welker, followed by kicker Stephen Gostkowski's 38-yard field goal. In the third quarter, Arizona's deficit continued to climb as Cassel completed a 76-yard touchdown pass to wide receiver Randy Moss, followed by Gostkowski's 35- and 24-yard field goal. In the fourth quarter, New England concluded its domination with Gostkowski's 30-yard      | third  |



► Representation extractor:



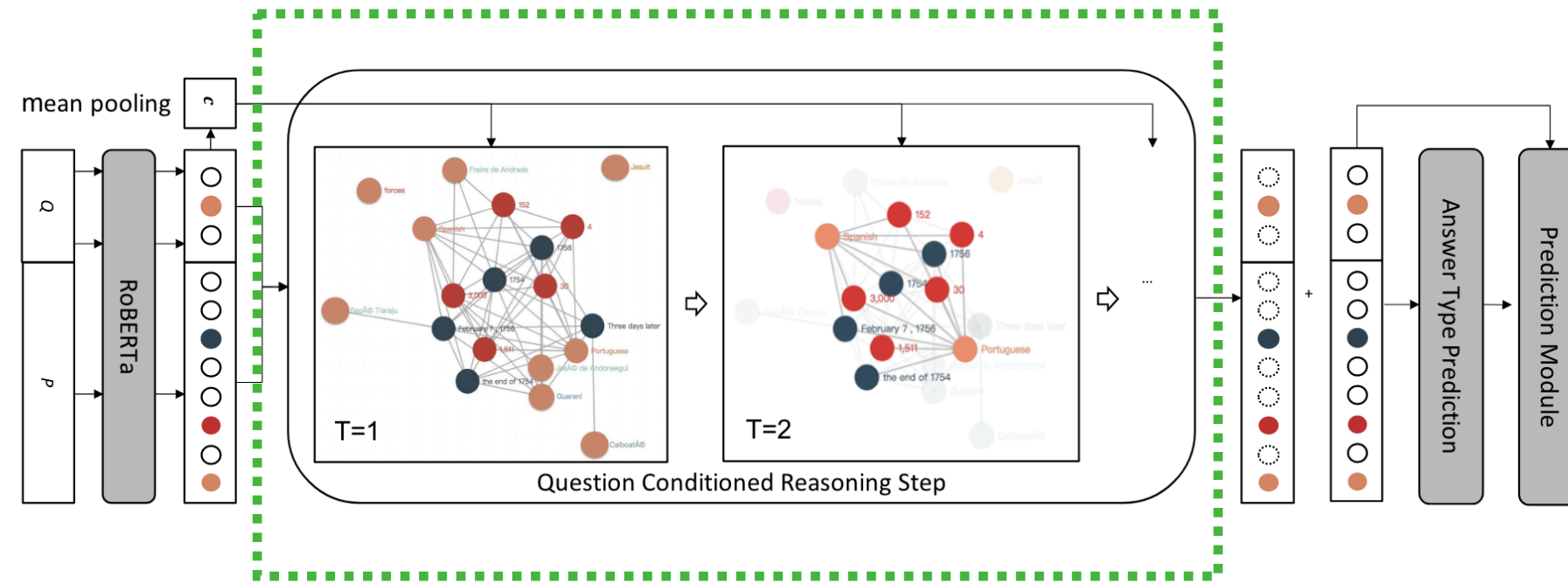
$$\hat{Q}, \hat{P} = \text{RoBERTa}(Q, P)$$

$$\mathcal{G} = (\mathbf{V}, \mathbf{E})$$

$$\mathbf{V} = \{\mathbf{N}, \mathbf{T}\},$$

{ NUMBER,PERCENT,MONEY,TIME,DATE,DURATION,ORDINAL,YARD,ENTITY }

## ► Reasoning module:



$$\mathbf{M}^Q = \mathbf{W}^M \hat{\mathbf{Q}}, \quad (2)$$

$$\mathbf{M}^P = \mathbf{W}^M \hat{\mathbf{P}}, \quad (3)$$

$$\mathbf{c} = \mathbf{W}^c \text{MEAN}(\hat{\mathbf{Q}}), \quad (4)$$

$$\mathbf{U} = \text{QDGAT}(\mathcal{G}; \mathbf{M}^P, \mathbf{M}^Q, \mathbf{c}), \quad (5)$$

$$\mathbf{m}^t = \mathbf{W}_{dc}^t g(\mathbf{W}_{fc} \mathbf{c}), \quad (6)$$

$$\mathbf{x}_q^t = \mathbf{W}_{qv}[\mathbf{v}^t : \mathbf{v}^0] \odot \mathbf{W}_{qc} \mathbf{m}^t, \quad (7)$$

$$\mathbf{x}_k^t = \mathbf{W}_{kv}[\mathbf{v}^t : \mathbf{v}^0] \odot \mathbf{W}_{kc} \mathbf{m}^t, \quad (8)$$

$$\mathbf{x}_v^t = \mathbf{W}_{vv}[\mathbf{v}^t : \mathbf{v}^0] \odot \mathbf{W}_{vc} \mathbf{m}^t, \quad (9)$$

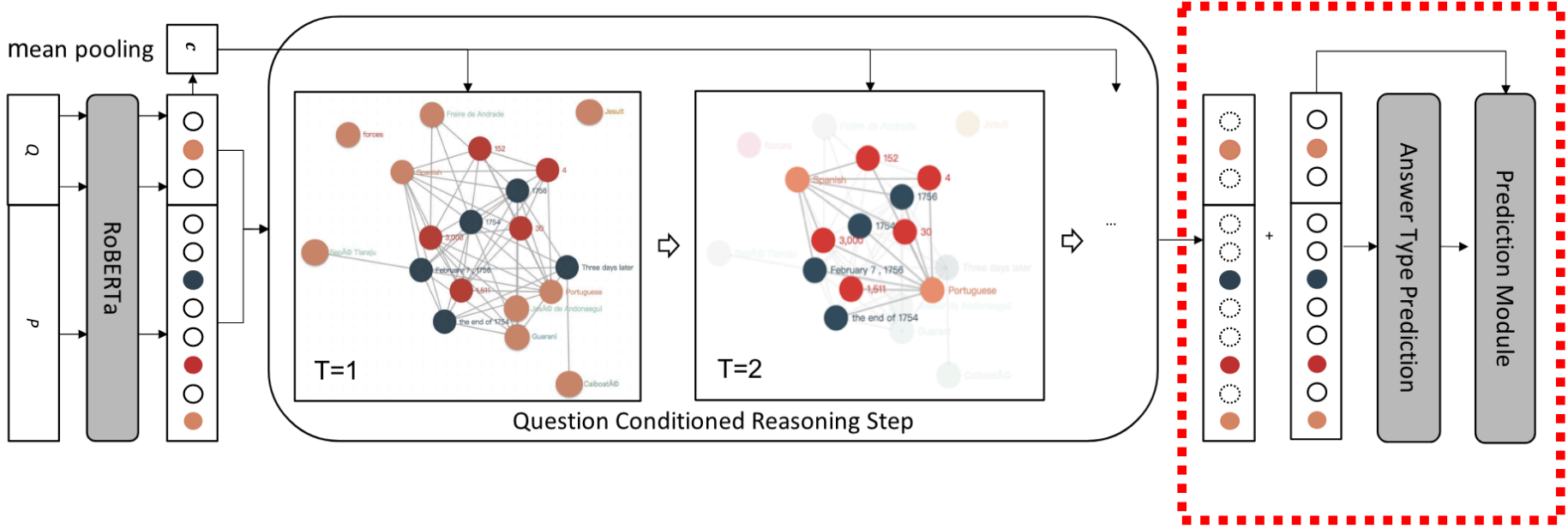
$$a_{i,j}^t = f\left(\sum_{r \in \mathcal{R}_{i,j}} \mathbf{W}_a^r[\mathbf{x}_{q,i}^t : \mathbf{x}_{k,j}^t]\right), \quad (10)$$

$$\alpha_{i,j}^t = \frac{\exp(a_{i,j}^t)}{\sum_{j' \in \mathcal{N}_i} \exp(a_{i,j'}^t)}, \quad (11)$$

$$\hat{\mathbf{x}}_i^t = \sum_{j \in \mathcal{N}_i} \alpha_{i,j} \mathbf{x}_{v,j}, \quad (12)$$

$$\mathbf{v}_i^{t+1} = \mathbf{W}_u[\mathbf{v}_i^t; \hat{\mathbf{x}}_i^t], \quad (13)$$

► QDGAT Model:



- Span
- Count
- Arithmetic expression



►Example and Results :

| Method                        | Dev                |                    | Test               |                    |
|-------------------------------|--------------------|--------------------|--------------------|--------------------|
|                               | EM                 | F1                 | EM                 | F1                 |
| Syn Dep                       | 9.38               | 11.64              | 8.51               | 10.84              |
| OpenIE                        | 8.80               | 11.31              | 8.53               | 10.77              |
| SRL                           | 9.28               | 11.72              | 8.98               | 11.45              |
| BiDAF                         | 26.06              | 28.85              | 24.75              | 27.49              |
| QANet                         | 27.50              | 30.44              | 25.50              | 28.36              |
| BERT                          | 30.10              | 33.36              | 29.45              | 32.70              |
| NAQANet                       | 46.20              | 49.24              | 44.07              | 47.01              |
| ALBERT-Calculator             | 80.22              | 83.98              | 79.85              | 83.56              |
| NumNet                        | 64.92              | 68.31              | 64.56              | 67.97              |
| NumNet+ (RoBERTa)             | 81.07 <sup>†</sup> | 84.42 <sup>†</sup> | 81.52 <sup>†</sup> | 84.84 <sup>†</sup> |
| NumNet+ (ensemble)            | 82.63 <sup>†</sup> | 85.59 <sup>†</sup> | 83.14 <sup>†</sup> | 86.16 <sup>†</sup> |
| QDGAT (RoBERTa)               | <b>82.74</b>       | <b>85.85</b>       | <b>83.23</b>       | <b>86.38</b>       |
| QDGAT <sub>p</sub> (RoBERTa)  | <b>84.07</b>       | <b>87.05</b>       | <b>84.53</b>       | <b>87.57</b>       |
| QDGAT <sub>p</sub> (ensemble) | <b>85.31</b>       | <b>88.10</b>       | <b>85.46</b>       | <b>88.38</b>       |
| Human                         |                    |                    | 94.09              | 96.42              |

Table 4: Decomposed performance on different answer types in the development set of DROP. Better results are in bold.

| Method  | Number       |              | Date         |              | Span         |              |
|---------|--------------|--------------|--------------|--------------|--------------|--------------|
|         | EM           | F1           | EM           | F1           | EM           | F1           |
| NumNet+ | 82.89        | 83.13        | 56.67        | 63.91        | 82.00        | 86.84        |
| QDGAT   | <b>86.00</b> | <b>86.23</b> | <b>60.27</b> | <b>67.48</b> | <b>84.05</b> | <b>88.53</b> |

## ► References

- Ran Q , Lin Y , Li P , et al. NumNet: Machine Reading Comprehension with Numerical Reasoning[J]. 2019.
- Dua D , Wang Y , Dasigi P , et al. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs[J]. 2019.
- Yu A W , Dohan D , Luong M T , et al. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension[J]. 2018.
- Chen K, Xu W, Cheng X, et al. Question Directed Graph Attention Network for Numerical Reasoning over Text[J]. 2020.

THANKS

— 2020.11.12