Machine Reading Comprehension

ONE SPAN-EXTRACT + TWO MULTI-CHOICE

Document Modeling with Graph Attention Networks for Multi-grained Machine Reading Comprehension(ACL2020)

Google Natural Questions

The NQ corpus contains questions from real users, and it requires QA systems to read and comprehend an entire Wikipedia article that may or may not contain the answer to the question. The inclusion of real user questions, and the requirement that solutions should read an entire page to find the answer, cause NQ to be a more realistic and challenging task than prior QA datasets.

Example

Question: where is the bowling hall of fame located

Wikipedia page: International Bowling Hall of Fame

Long answer: The World Bowling Writers (WBW) International Bowling Hall of Fame was established in 1993 and is located in the International Bowling Museum and Hall of Fame , on the International Bowling Campus in Arlington , Texas .

Short answer: Arlington , Texas

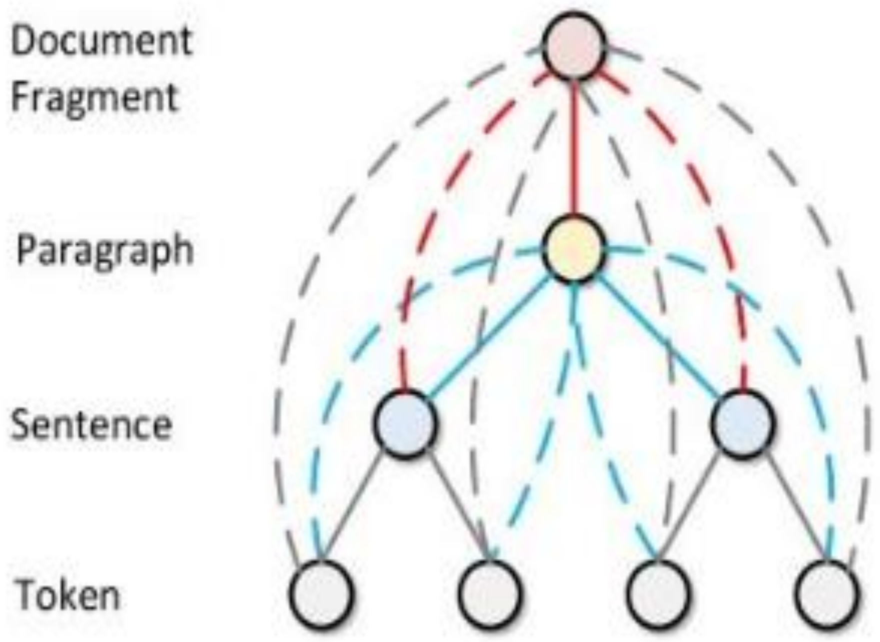
Document Modeling with Graph Attention Networks for Multi-grained Machine Reading Comprehension(ACL2020)

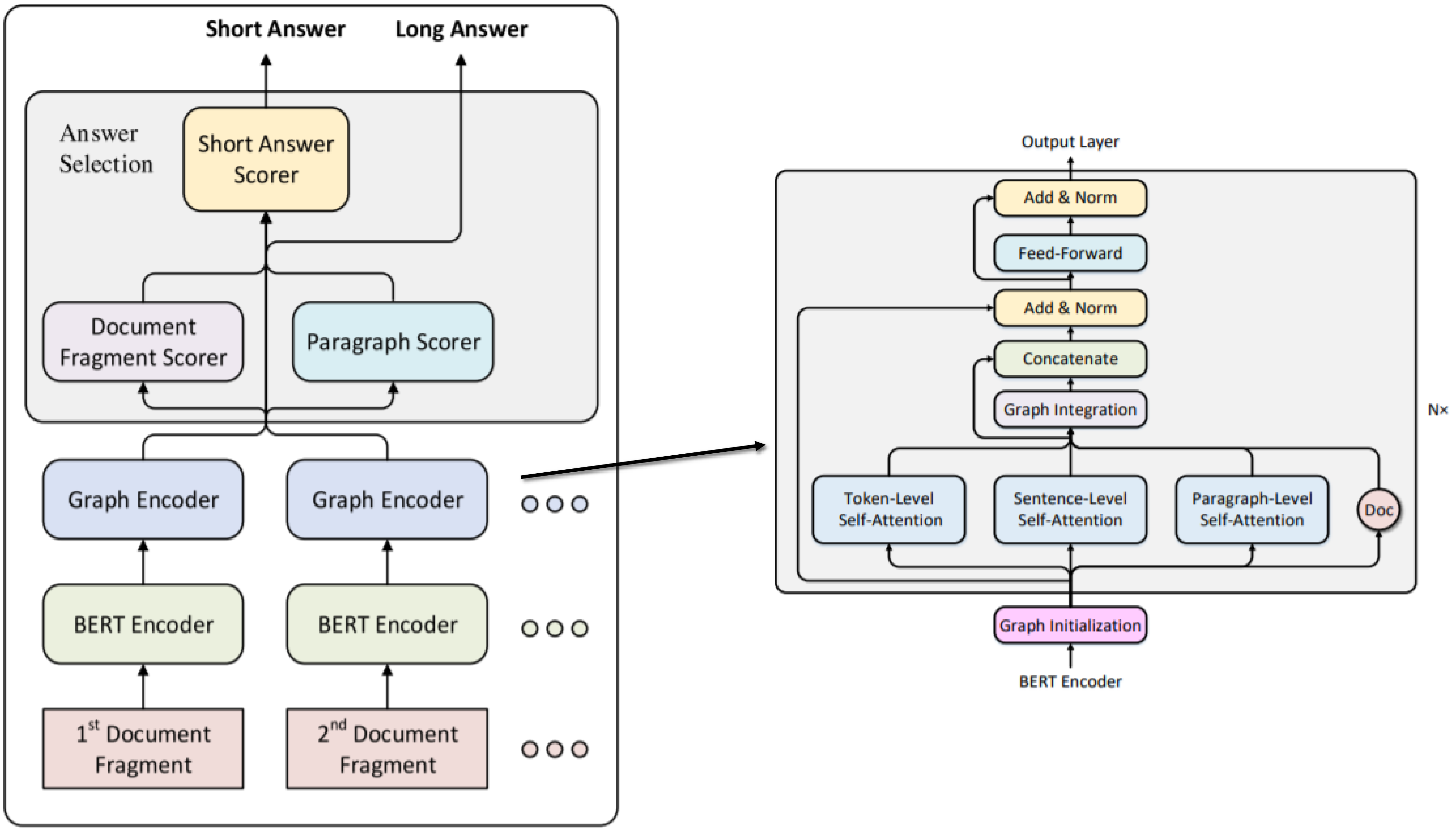
1、Dataset contain a source document, a annations, and some long answer or short answer candidates.


```
{
  "example_id": 3902,
  "document_url": "http://wikipedia.org/en/strings"
  "question_text": "what is a string",
  "document_text": "<P> A string is a list of characters in order . </P>",
  "annotations": [{
    "long_answer": { "start_token": 0, "end_token": 12 },
    "short_answers": [{ "start_token": 5, "end_token": 8 }],
    "yes_no_answer": "NONE",
  }],
  "long_answer_candidates": [
    {"start_token": 0, "end_token": 12, "top_level": True}
  ]
}
```

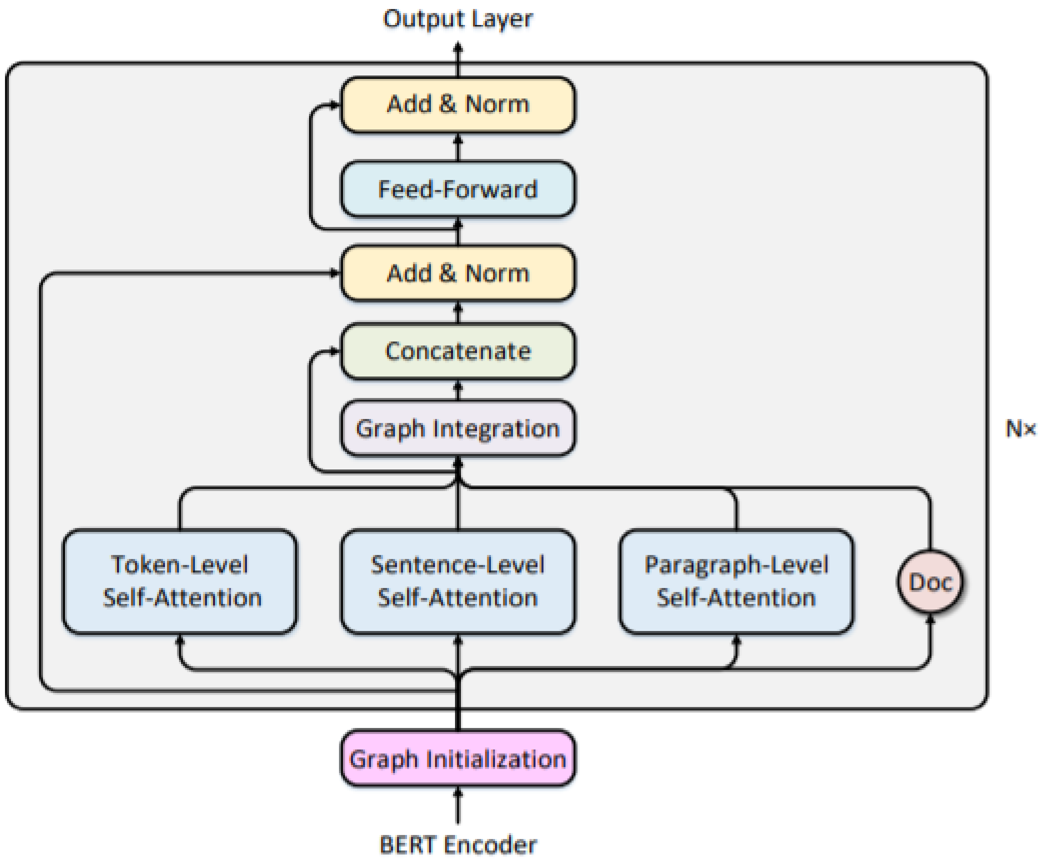
Document Modeling with Graph Attention Networks for Multi-grained Machine Reading Comprehension(ACL2020)

2、Node construct
Long document is chunked into multi small span.
The bounds of sentences is constructed using Spacy.
The bounds of paragraphs is constructed using Long candidate end.
Only one document is constructed in one sample.




Short Answer
Long Answer
Answer Selection
Short Answer Scorer
Document Fragment Scorer
Paragraph Scorer
Graph Encoder
Graph Encoder
BERT Encoder
BERT Encoder
1st Document Fragment
2nd Document Fragment
Output Layer
Add & Norm
Feed-Forward
Add & Norm
Concatenate
Graph Integration
Token-Level Self-Attention
Sentence-Level Self-Attention
Paragraph-Level Self-Attention
Doc
N×
Graph Initialization
BERT Encoder

Document Modeling with Graph Attention Networks for Multi-grained Machine Reading Comprehension(ACL2020)

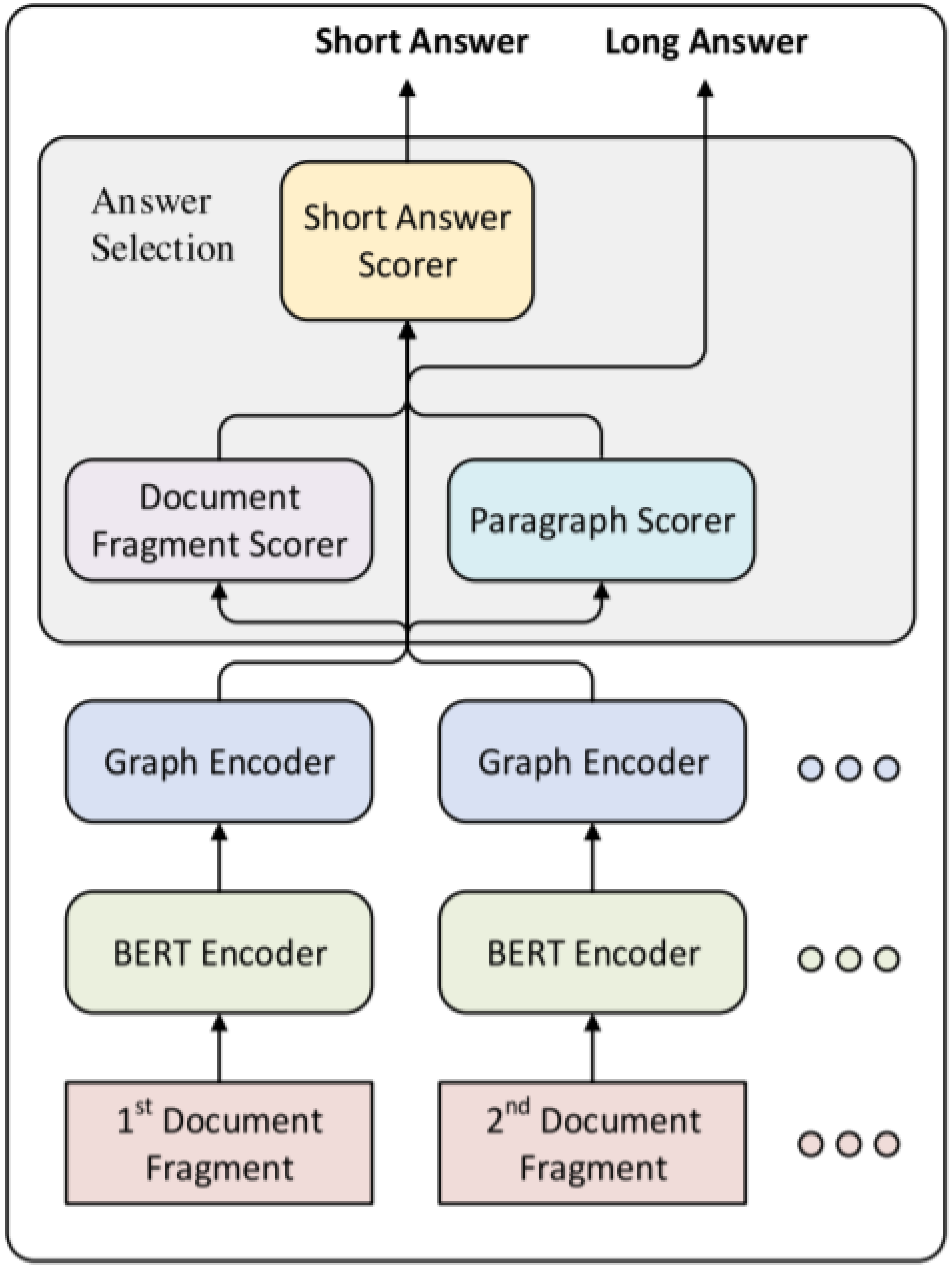


Relational Embedding

$$e_{ij} = \frac{\left(\boldsymbol{h}_i \mathbf{W}^{\mathrm{Q}}\right)\left(\boldsymbol{h}_j \mathbf{W}^{\mathrm{K}}\right)^{\mathrm{T}} + \boldsymbol{h}_i \mathbf{W}^{\mathrm{Q}}\left(\boldsymbol{a}_{ij}^{\mathrm{K}}\right)^{\mathrm{T}}}{\sqrt{d_z}}$$

$$\boldsymbol{z}_i = \sum_{j \in \mathcal{N}_i} \alpha_{ij}\left(\boldsymbol{h}_j \mathbf{W}^{\mathrm{V}} + \boldsymbol{a}_{ij}^{\mathrm{V}}\right).$$

$$\alpha_{ij} = \mathrm{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})}.$$



Short Answer:
From Token level Node score
Long Answer:
From Paragraph level Node score
Document Node:
Give Yes and No.

Model	LA. F1	SA. F1
BERT-base+Model-III	**68.9**	**51.9**
-Graph module	63.9	51.0
-Long answer prediction	65.1	51.4
-Short answer prediction	68.2	-
-Relational embedding	68.8	51.7
-Graph integration layer	68.3	51.1
-Self-attention layer	68.4	51.2

DCMN+: Dual Co-Matching Network for Multi-choice Reading Comprehension (AAAI 2020)

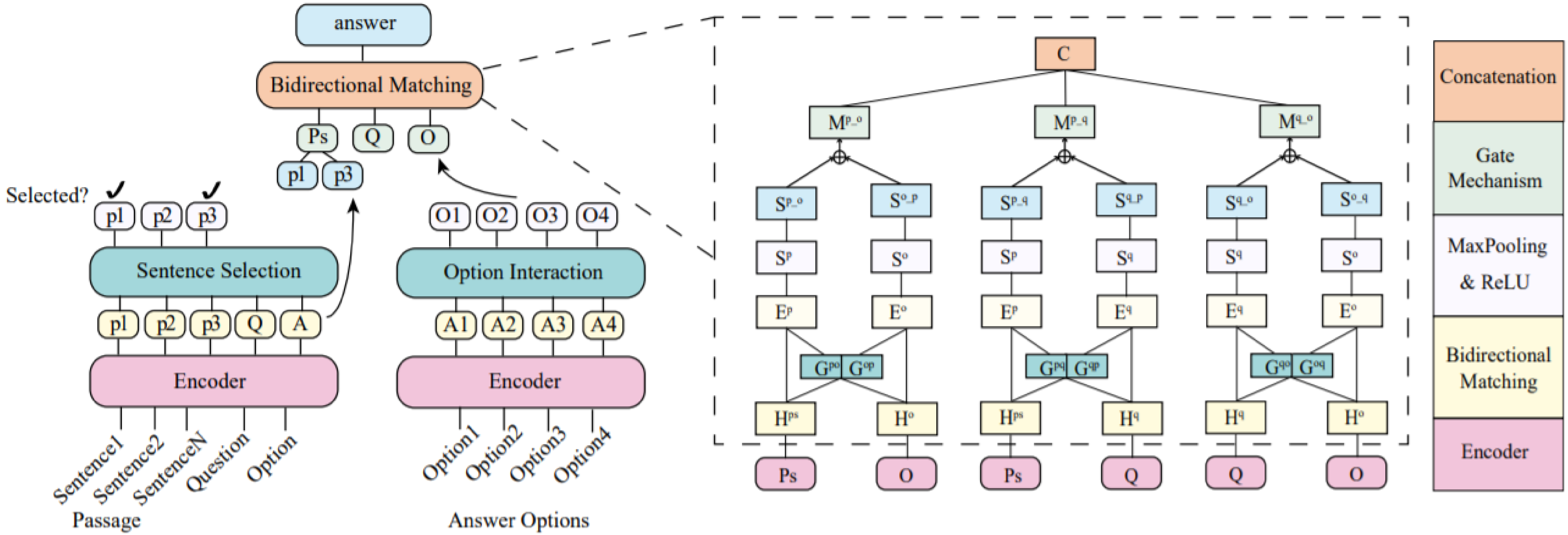
RACE (Lai et al. 2017): RACE consists of two subsets: RACE-M and RACE-H respectively corresponding to middle school and high school difficulty levels, which is recognized as one of the largest and most difficult datasets in multi-choice reading comprehension.

Passage: *Runners in a relay race pass a stick in one direction. However, merchants passed silk, gold, fruit, and glass along the Silk Road in more than one direction. They earned their living by traveling the famous Silk Road. ...* ***The Silk Road was made up of many routes, not one smooth path.*** *They passed through what are now 18 countries. The routes crossed mountains and deserts and had many dangers of hot sun, deep snow and even battles...*

Question: *The Silk Road became less important because _ .*

A. *it was made up of different routes*
B. *silk trading became less popular*
C. ***sea travel provided easier routes***
D. *people needed fewer foreign goods*

DCMN+: Dual Co-Matching Network for Multi-choice Reading Comprehension (AAAI 2020)



Passage Sentence Selection

Cosine score:

$$\mathbf{D}^{pa} = Cosine(\mathbf{H}^{a}, \mathbf{H}^{p_i}) \in R^{|A| \times |p_i|}$$

$$\mathbf{D}^{pq} = Cosine(\mathbf{H}^{q}, \mathbf{H}^{p_i}) \in R^{|Q| \times |p_i|}$$

$$\bar{\mathbf{D}}^{pa} = MaxPooling(\mathbf{D}^{pa}) \in R^{|A|}$$

$$\bar{\mathbf{D}}^{pq} = MaxPooling(\mathbf{D}^{pq}) \in R^{|Q|}$$

$$score = \frac{\sum_{k=1}^{|A|} \bar{\mathbf{D}}_k^{pa}}{|A|} + \frac{\sum_{k=1}^{|Q|} \bar{\mathbf{D}}_k^{pq}}{|Q|}$$

Bilinear score:

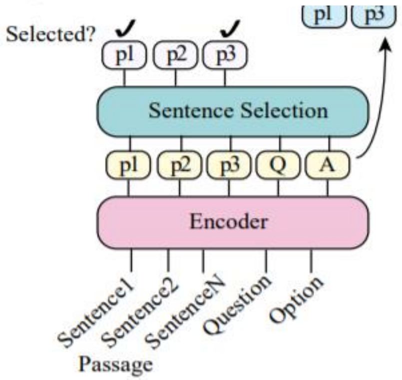
$$\alpha = SoftMax(\mathbf{H}^{q} W_1) \in R^{|Q| \times l}$$

$$\mathbf{q} = \alpha^T \mathbf{H}^{q} \in R^{l}$$

$$\bar{\mathbf{P}}_j = \mathbf{H}_j^{p_i} W_2 \mathbf{q} \in R^{l}, j \in [1, |p_i|]$$

$$\hat{\mathbf{P}}^{pq} = Max(\bar{\mathbf{P}}_1 \bar{\mathbf{P}}_2, ..., \bar{\mathbf{P}}_{|p_i|}) \in R^{l}$$

$$score = W_3^T \hat{\mathbf{P}}^{pq} + W_4^T \hat{\mathbf{P}}^{pa}$$



Answer Option Interaction

$$\mathbf{G} = SoftMax(\mathbf{H}^{a_i} W_5 \mathbf{H}^{a_j T}) \in R^{|A_i| \times |A_j|}$$

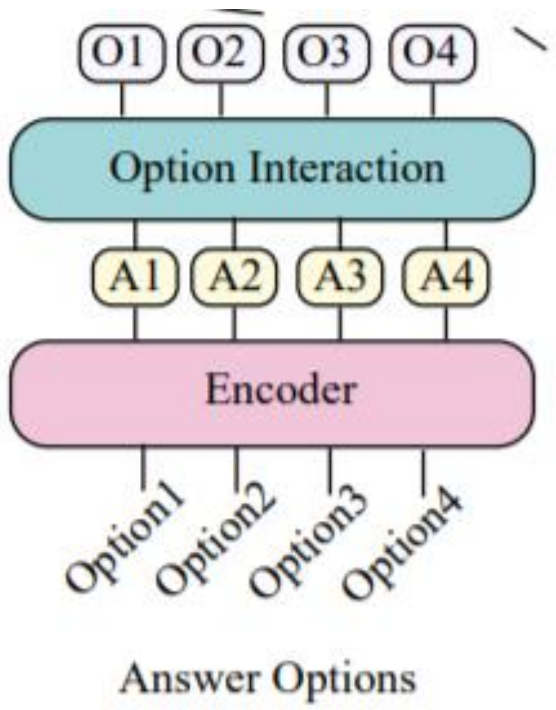
$$\mathbf{H}^{a_{i,j}} = ReLU(\mathbf{G}\mathbf{H}^{a_j}) \in R^{|A_i| \times l}$$

$$\hat{\mathbf{H}}^{a_i} = [\{\mathbf{H}^{a_{i,j}}\}_{j \neq i}] \in R^{|A_i| \times (m-1)l}$$

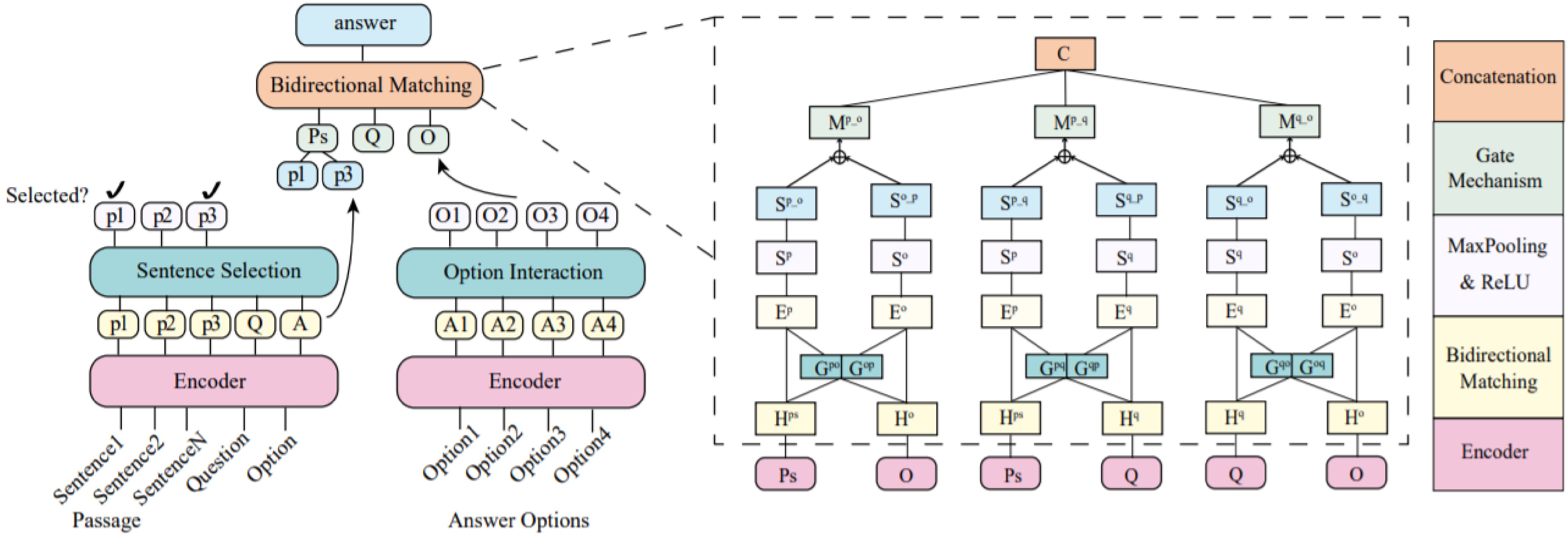
$$\bar{\mathbf{H}}^{a_i} = \hat{\mathbf{H}}^{a_i} W_6 \in R^{|A_i| \times l}$$

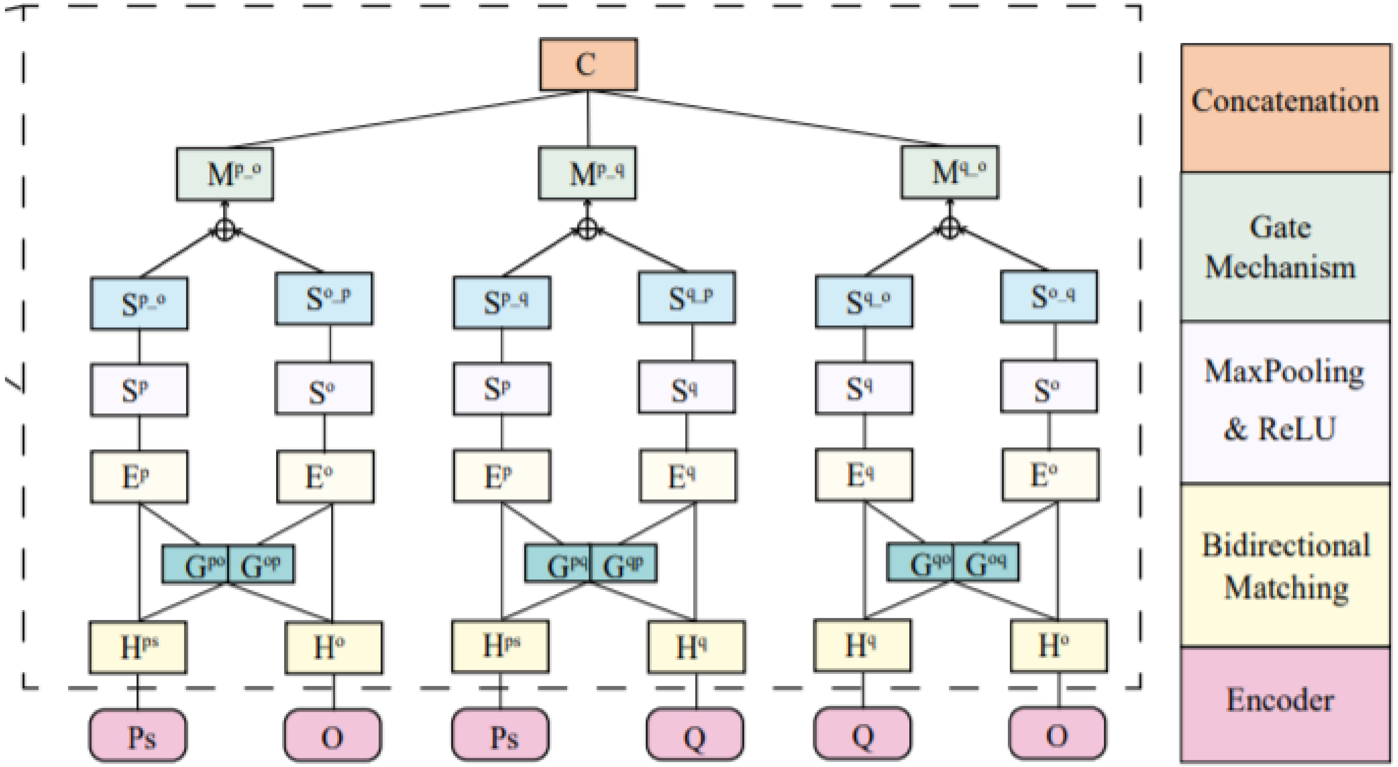
$$g = \sigma(\bar{\mathbf{H}}^{a_i} W_7 + \mathbf{H}^{a_i} W_8 + b)$$

$$\mathbf{H}^{o_i} = g * \mathbf{H}^{a_i} + (1 - g) * \bar{\mathbf{H}}^{a_i}$$



DCMN+: Dual Co-Matching Network for Multi-choice Reading Comprehension (AAAI 2020)





$$\mathbf{G}^{qo} = SoftMax(\mathbf{H}^q W_9 \mathbf{H}^{o^T})$$

$$\mathbf{G}^{oq} = SoftMax(\mathbf{H}^o W_{10} \mathbf{H}^{q^T})$$

$$\mathbf{E}^q = \mathbf{G}^{qo}\mathbf{H}^o, \mathbf{E}^o = \mathbf{G}^{oq}\mathbf{H}^q$$

$$\mathbf{S}^q = ReLU(\mathbf{E}^q W_{11})$$

$$\mathbf{S}^o = ReLU(\mathbf{E}^o W_{12})$$

$$\mathbf{S}^{q\text{-}o} = MaxPooling(\mathbf{S}^q)$$

$$\mathbf{S}^{o\text{-}q} = MaxPooling(\mathbf{S}^o)$$

$$g = \sigma(\mathbf{S}^{q\text{-}o} W_{13} + \mathbf{S}^{o\text{-}q} W_{14} + b)$$

$$\mathbf{M}^{q\text{-}o} = g * \mathbf{S}^{o\text{-}q} + (1 - g) * \mathbf{S}^{o\text{-}q}$$

DCMN+: Dual Co-Matching Network for Multi-choice Reading Comprehension (AAAI 2020)

Objective Function

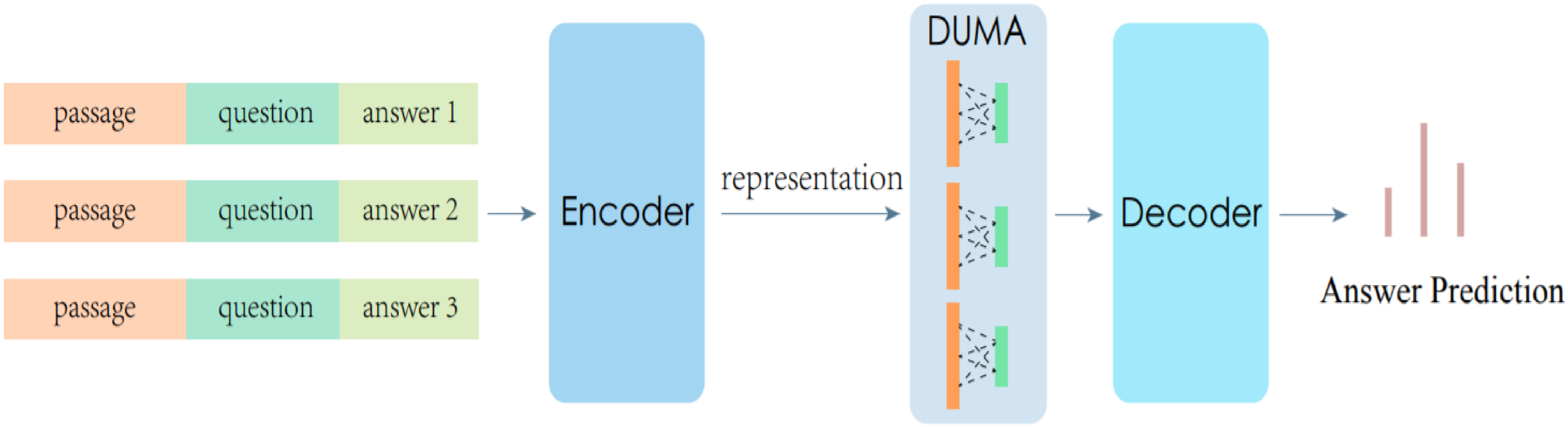
$$\mathbf{C} = [\mathbf{M}^{p\text{-}q}; \mathbf{M}^{p\text{-}o}; \mathbf{M}^{q\text{-}o}]$$

$$L(A_k|P,Q) = -log\frac{\exp(V^T\mathbf{C}_k)}{\sum_{j=1}^{m}\exp(V^T\mathbf{C}_j)}$$

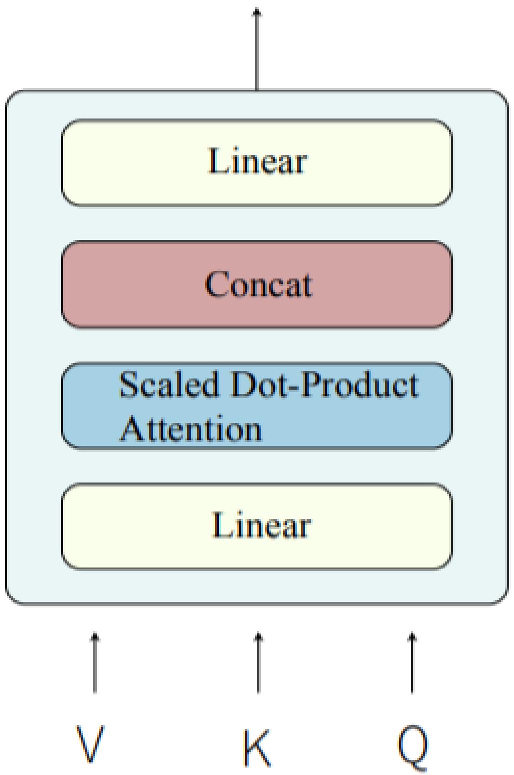
	$BERT_{base}$	$BERT_{large}$	$XLNet_{large}$
base encoder	64.6	71.8	80.1
+ DCMN	66.0 (+1.4)	73.8 (+2.0)	81.5 (+1.4)
+ DCMN + P_*SS*	66.6 (+2.0)	74.6 (+2.8)	82.1 (+2.0)
+ DCMN + P_*OI*	66.8 (+2.2)	74.4 (+2.6)	82.2 (+2.1)
+ DCMN + ALL (DCMN+)	**67.4 (+2.8)**	**75.4 (+3.6)**	**82.6 (+2.5)**

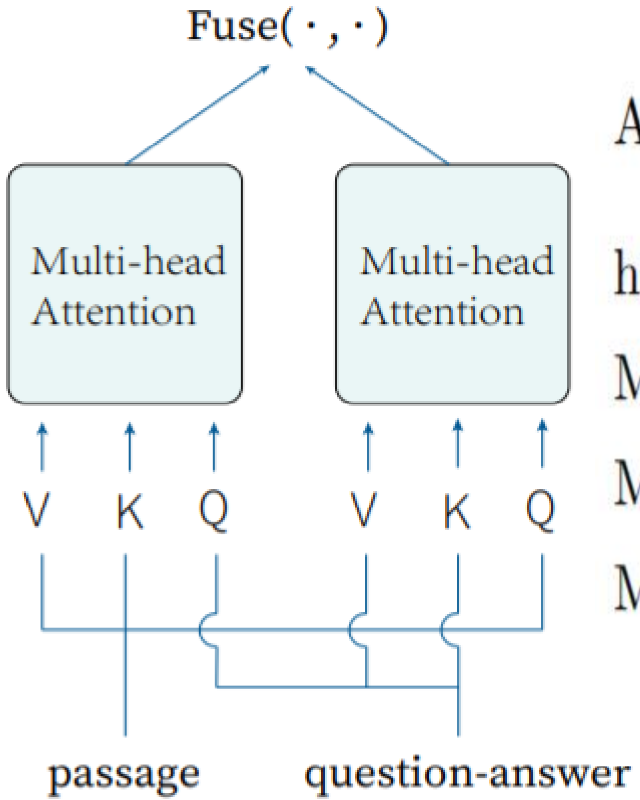
Model	RACE-M/H	RACE
HAF (Zhu et al. 2018)	45.0/46.4	46.0
MRU (Tay, Tuan, and Hui 2018)	57.7/47.4	50.4
HCM (Wang et al. 2018b)	55.8/48.2	50.4
MMN (Tang, Cai, and Zhuo 2019)	61.1/52.2	54.7
GPT (Radford 2018)	62.9/57.4	59.0
RSM (Sun et al. 2019)	69.2/61.5	63.8
OCN (Ran et al. 2019)	76.7/69.6	71.7
XLNet (Yang et al. 2019)	85.5/80.2	81.8
$BERT_{base}$*	71.1/62.3	65.0
$BERT_{large}$*	76.6/70.1	72.0
$XLNet_{large}$*	83.7/78.6	80.1
Our Models		
$BERT_{base}$* + DCMN	73.2/64.2	67.0
$BERT_{large}$* + DCMN	79.2/72.1	74.1
$BERT_{large}$* + DCMN + P_{SS} + A_{OI}	79.3/74.4	**75.8**
$XLNet_{large}$* + DCMN + P_{SS} + A_{OI}	86.5/81.3	**82.8**
Human Performance		
Turkers	85.1/69.4	73.3
Ceiling	95.4/94.2	94.5

Dual Multi-head Co-attention for Multi-choice Reading Comprehension



Dual Multi-head Co-attention for Multi-choice Reading Comprehension





$$\mathrm{Attention}(E^P, E^{QA}, E^{QA}) = \mathrm{softmax}\left(\frac{E^P (E^{QA})^T}{\sqrt{d_k}}\right) E^{QA}$$

$$\mathrm{head_i} = \mathrm{Attention}(E^P W_i^Q, E^{QA} W_i^K, E^{QA} W_i^V)$$

$$\mathrm{MHA}(E^P, E^{QA}, E^{QA}) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h}) W^O$$

$$\mathrm{MHA_1} = \mathrm{MHA}(E^P, E^{QA}, E^{QA})$$

$$\mathrm{MHA_2} = \mathrm{MHA}(E^{QA}, E^P, E^P)$$

$$\mathrm{DUMA}(E^P, E^{QA}) = \mathrm{Fuse}(\mathrm{MHA_1}, \mathrm{MHA_2}) \qquad (1)$$

Dual Multi-head Co-attention for Multi-choice Reading Comprehension

Decoder

$$O_i = \mathrm{DUMA}(E^P, E^{QA_i})$$

$$L(A_r|P,Q) = -\log \frac{\exp(W^T O_r)}{\sum_{i=1}^{s} \exp(W^T O_i)}$$

model	dev	test
FTLM++ [Sun *et al.*, 2019a]	58.1*	58.2*
$\mathrm{BERT_{large}}$ [Devlin *et al.*, 2018]	66.0*	66.8*
XLNet [Yang *et al.*, 2019]	-	72.0*
$\mathrm{RoBERTa_{large}}$ [Liu *et al.*, 2019]	85.4*	85.0*
$\mathrm{RoBERTa_{large}}$+MMM [Jin *et al.*, 2020]	88.0*	88.9*
$\mathrm{ALBERT_{xxlarge}}$ [Lan *et al.*, 2020]	89.2	88.5
Our model	**89.3**	**90.4**
Our model + multi-task learning[Wan, 2020]	-	**91.8**

model	test (M/H)	source
HAF [Zhu *et al.*, 2018a]	46.0(45.0/46.4)*	publication
MRU [Tay *et al.*, 2018]	50.4(57.7/47.4)*	publication
HCM [Wang *et al.*, 2018]	50.4(55.8/48.2)*	publication
MMN [Tang *et al.*, 2019]	54.7(61.1/52.2)*	publication
GPT [Radford *et al.*, 2018]	59.0(62.9/57.4)*	publication
RSM [Sun *et al.*, 2019b]	63.8(69.2/61.5)*	publication
OCN [Ran *et al.*, 2019]	71.7(76.7/69.6)*	publication
XLNet [Yang *et al.*, 2019]	81.8(85.5/80.2)*	publication
$\mathrm{XLNet_{xxlarge}}$+DCMN+ [Zhang *et al.*, 2020]	82.8(86.5/81.3)*	publication
XLNet + DCMN+	82.8(86.5/81.3)	leaderboard
RoBERTa	83.2(86.5/81.8)	leaderboard
DCMN+ (ensemble)	84.1(88.5/82.3)	leaderboard
RoBERTa + MMM	85.0(89.1/83.3)	leaderboard
ALBERT (single)	86.5(89.0/85.5)	leaderboard
ALBERT (ensemble)	89.4(91.2/88.6)	leaderboard
$\mathrm{ALBERT_{xxlarge}}$ [Lan *et al.*, 2020]	86.6(89.0/85.5) †	
$\mathrm{ALBERT_{xxlarge}}$+DUMA	**88.0(90.9/86.7)**	our model
$\mathrm{ALBERT_{xxlarge}}$+DUMA(ensemble)	**89.8(92.6/88.7)**	our model

RACE Leaderboard

Model	Report Time	Institute	RACE	RACE-M	RACE-H
Human Ceiling Performance	Apr 15, 2017	CMU	*94.5*	*95.4*	*94.2*
Amazon Mechanical Turker	Apr 15, 2017	CMU	73.3	85.1	69.4
Megatron-BERT (ensemble)	Mar 13, 2020	NVIDIA Research	**90.9**	**93.1**	**90.0**
ALBERT + DUMA (ensemble)	Mar 18, 2020	SJTU & Huawei Noah's Ark Lab	89.8	92.6	88.7
Megatron-BERT	Mar 13, 2020	NVIDIA Research	89.5	91.8	88.6
ALBERT (ensemble)	Sep 26, 2019	Google Research & TTIC	89.4	91.2	88.6
UnifiedQA	May 02, 2020	AI2 & UW	89.4	-	-
ALBERT + DUMA	Feb 08, 2020	SJTU & Huawei Noah's Ark Lab	88.0	90.9	86.7
T5*	May 02, 2020	Google	87.1	-	-

Leaderboard

Report Time	Model	Accuracy
	Human Ceiling Performance *Tencent & Cornell & UW & AI2* **Sun et al., 2019**	98.6
	Human Performance *Tencent & Cornell & UW & AI2* **Sun et al., 2019**	95.5
Feb 26, 2020	ALBERT-xxlarge + DUMA + Multi-Task Learning *IBM Research AI* **Wan et al., 2020**	**91.8**
Feb 05, 2020	ALBERT-xxlarge + DUMA *SJTU & Huawei Noah's Ark Lab* **Zhu et al., 2020**	90.4
Oct 01, 2019	RoBERTa-Large + MMM *MIT & Amazon Alexa AI* **Jin et al., 2019**	88.9
Jul 21, 2019	XLNet-Large *River Valley High School, Singapore* **https://github.com/NoviScl/XLNet_DREAM**	72.0