统计自然语言处理——论文

战蕾

51174506099

Learning to Rank Question Answer Pairs with Holographic Dual LSTM Architecture

Yi Tay
Nanyang Technological University
ytay017@e.ntu.edu.sg

Minh C. Phan
Nanyang Technological University
phan0005@e.ntu.edu.sg

Luu Anh Tuan
Institute for Infocomm Research
at.luu@i2r.a-star.edu.sg

Siu Cheung Hui
Nanyang Technological University
asschui@ntu.edu.sg

SIGIR'17, August 7-11, 2017, Shinjuku, Tokyo, Japan

目录

01 Background

02 Model

03 Results

04 Conclusion

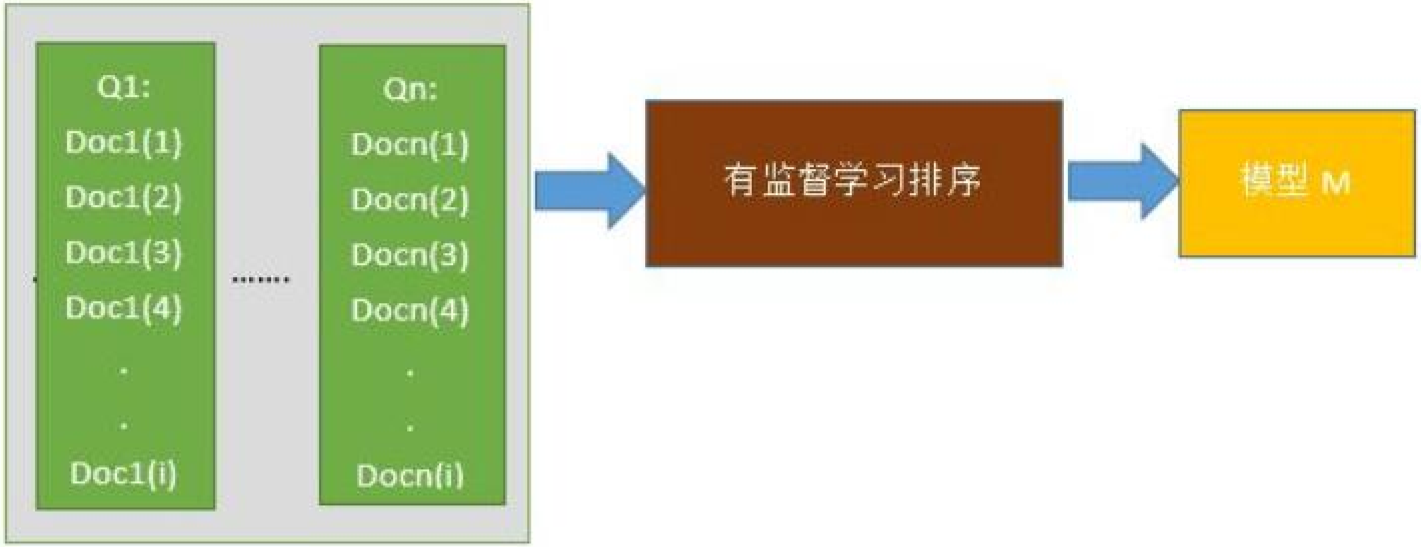
Background——Learning to Rank

Learning to Rank

Learning to rank有个中文名，叫做排序学习。

它是一种基于Supervised Learning的排序方法。

训练阶段：



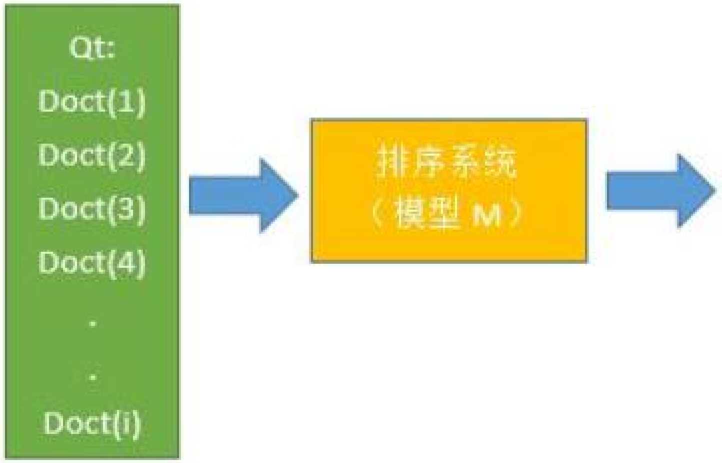
Background——Learning to Rank

Learning to Rank

在互联网不断发展的今天，更多复杂而有效设计的特征被应用到检索计算里面，比如查询与文档深层次匹配，网页pagerank等。

人工参数调整已经不能满足需求，此时机器学习被应用到这项任务中，同时由于互联网海量数据的特点，比如展现点击日志，基于大数据的learning to rank逐渐成为热门的领域。

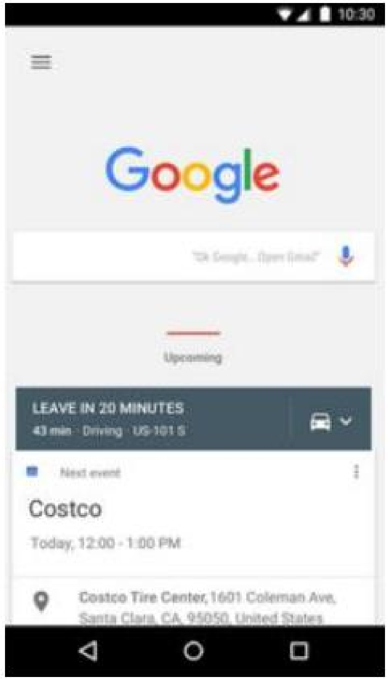
测试阶段：



Background——问答系统

Question Answering (QA) systems answer natural language questions.

Google Now



10:30

"Ok Google... Open Gmail"

Upcoming

LEAVE IN 20 MINUTES
43 min · Driving · US-101 S

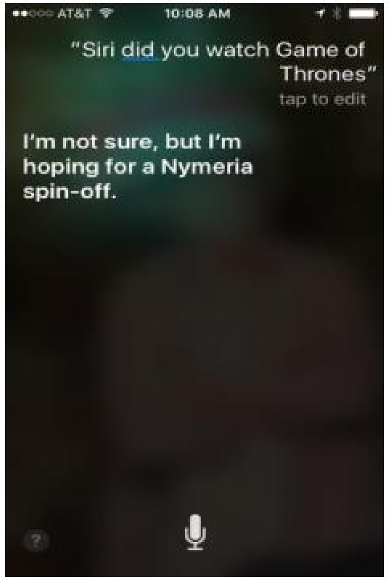
Next event

Costco

Today, 12:00 - 1:00 PM

Costco Tire Center, 1601 Coleman Ave, Santa Clara, CA 95050, United States

Apple Siri



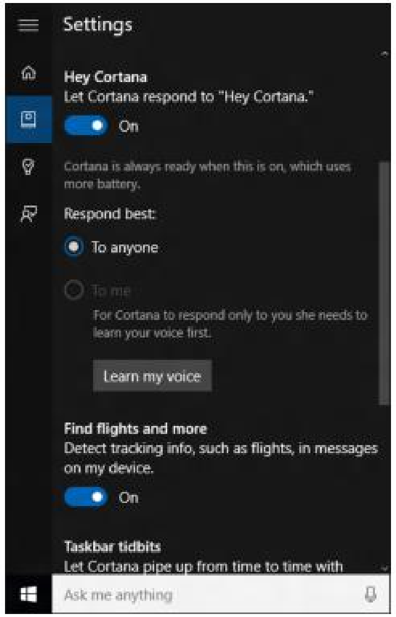
AT&T 10:08 AM

"Siri did you watch Game of Thrones"
tap to edit

I'm not sure, but I'm hoping for a Nymeria spin-off.

Amazon Alexa

Microsof Cortana



Settings

Hey Cortana
Let Cortana respond to "Hey Cortana."
On

Cortana is always ready when this is on, which uses more battery.

Respond best:

● To anyone

○ To me

For Cortana to respond only to you she needs to learn your voice first.

Learn my voice

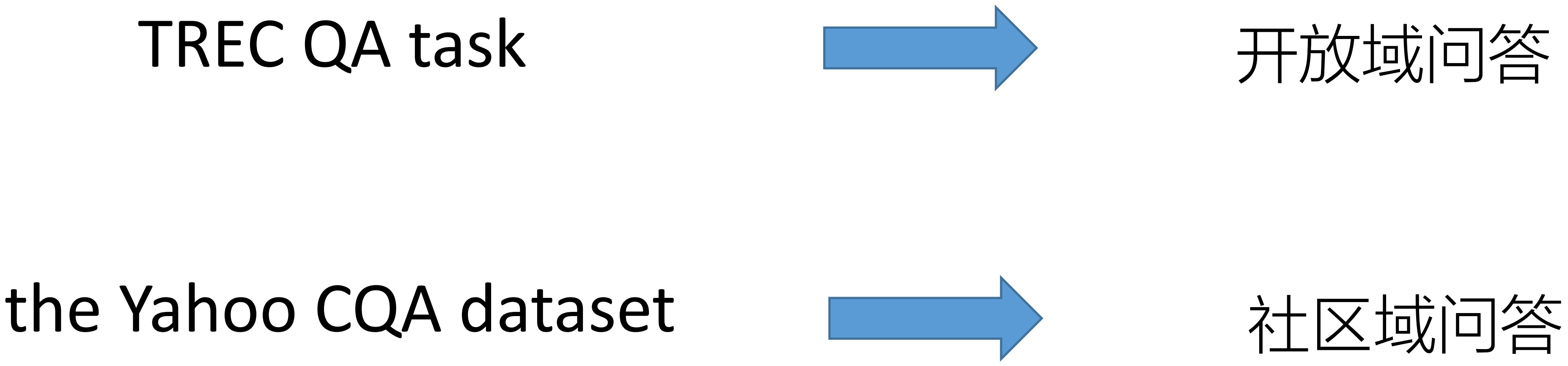
Find flights and more
Detect tracking info, such as flights, in messages on my device.
On

Taskbar tidbits
Let Cortana pipe up from time to time with

Ask me anything

Background——问答系统

Question Answering (QA) systems answer natural language questions.



TREC QA task → 开放域问答

the Yahoo CQA dataset → 社区域问答

Background——Motivation

This is highly related to many search and information retrieval tasks such as **traditional document retrieval and text matching**.

A key difference

Background——Motivation

This is highly related to many search and information retrieval tasks such as **traditional document retrieval and text matching**.

Much shorter compared to full-fledged documents

↓

? ? ?

Background——Motivation

This is highly related to many search and information retrieval tasks such as **traditional document retrieval and text matching**.

Much shorter compared to full-fledged documents

↓

designing features

Background——Motivation

Challenges:

1.Feature representations of questions and answers have to be learned or designed.

2.A similarity function has to be defined to match questions to answers

Background——Motivation

Recently:

Deep learning architectures

Background——Motivation

Recently:

Example CQA

Convolutional neural tensor network (CNTN)——a tensor layer is used to model the relationship between the representations using an additional tensor parameter

Background——Motivation

Recently:

Example CQA

Convolutional neural tensor network (CNTN)——a tensor layer is used to model the relationship between the representations using an additional tensor parameter

1.increases the number of parameters and inevitably increases the risk of overfitting.

Background——Motivation

Recently:

Example CQA

Convolutional neural tensor network (CNTN)——a tensor layer is used to model the relationship between the representations using an additional tensor parameter

1.increases the number of parameters and inevitably increases the risk of overfitting.

2.this significantly increases computational and memory cost of the overall network.

Background——Motivation

Recently:

Example CQA

Convolutional neural tensor network (CNTN)——a tensor layer is used to model the relationship between the representations using an additional tensor parameter

1.increases the number of parameters and inevitably increases the risk of overfitting.

2.this significantly increases computational and memory cost of the overall network.

3.restricts the expressiveness of the QA representations

Background——Motivated

The prime contributions of the paper

1.We adopt holographic composition for modeling the interaction between representations of QA pairs

Background——Motivated

The prime contributions of the paper

1.We adopt holographic composition for modeling the interaction between representations of QA pairs

2.We present a novel deep learning architecture,HD-LSTM (Holographic Dual LSTM) for learning to rank QA pairs

Background——Motivated

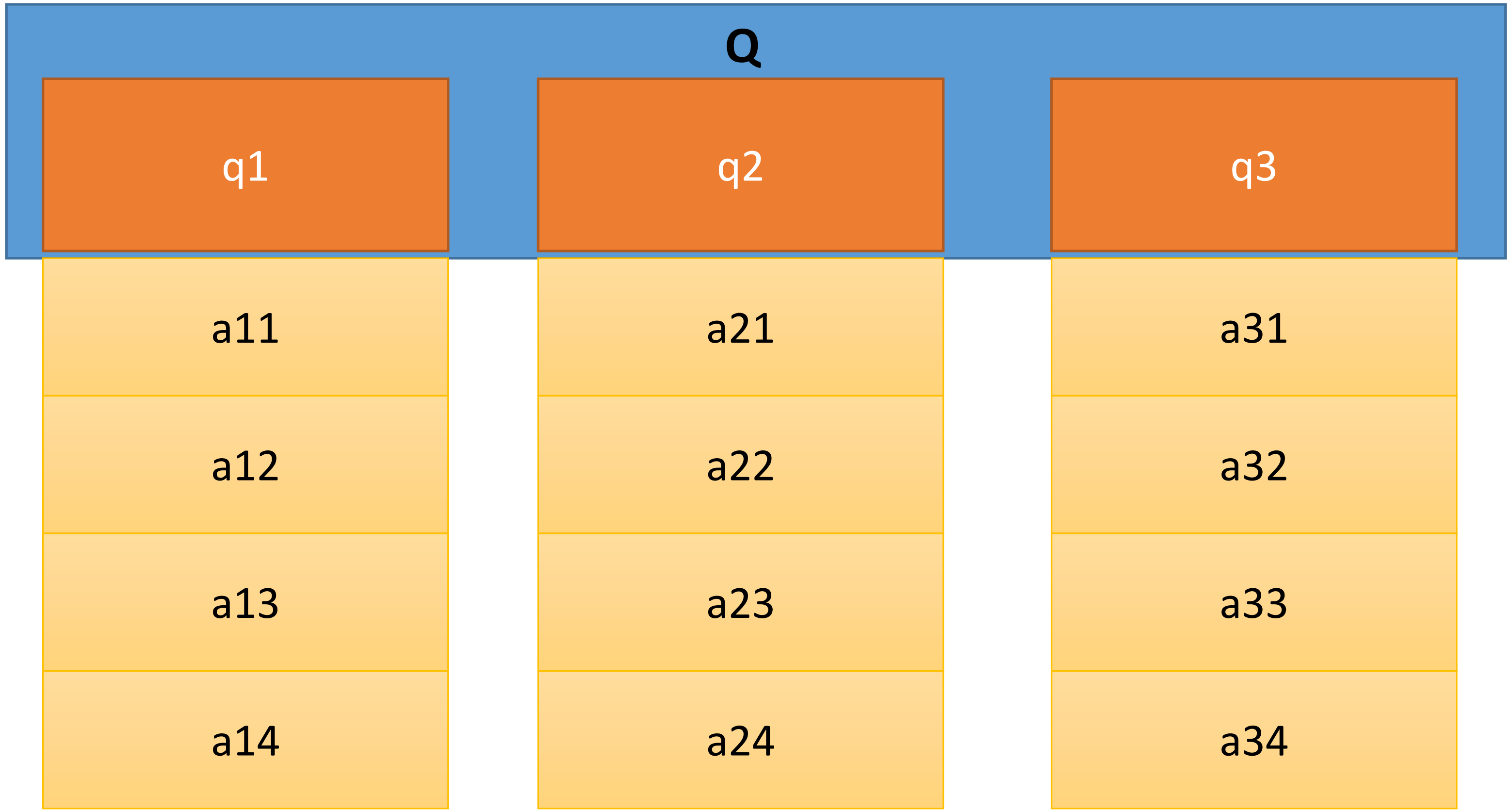
The prime contributions of the paper

1.We adopt holographic composition for modeling the interaction between representations of QA pairs

2.We present a novel deep learning architecture,HD-LSTM (Holographic Dual LSTM) for learning to rank QA pairs

3.We provide extensive experimental evidence of the effectiveness of our model on both factoid question answering and community-based question answering.

Background——Problem Statement



Q

q1	q2	q3
a11	a21	a31
a12	a22	a32
a13	a23	a33
a14	a24	a34

a function: $f(q, a) \in [0, 1]$ → rank a list of possible candidates

Background——Approach

Three ways for **supervised text ranking**

1.Pointwise

2.Pairwise

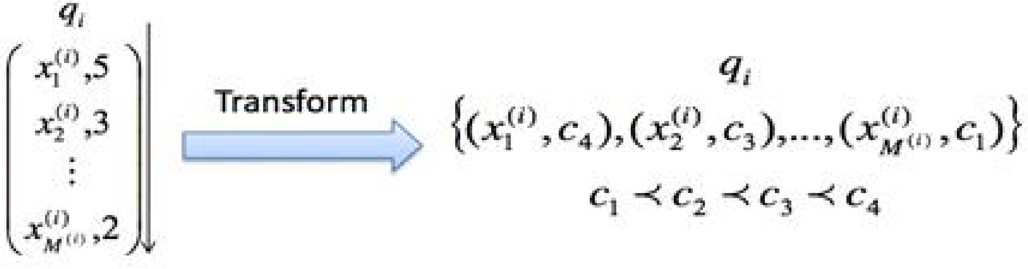
3.Listwise

Background——Approach

Three ways for **supervised text ranking**

1.Pointwise(this paper)

单文档方法的处理对象是单独的一篇文档，将文档转换为特征向量后，机器学习系统根据从训练数据中学习到的分类或者回归函数对文档打分，打分结果即是搜索结果。

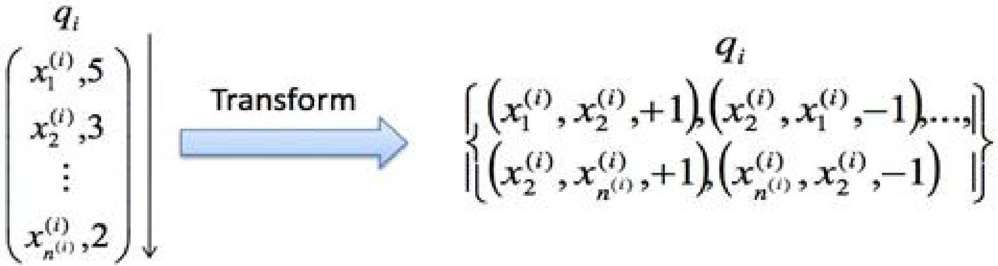

q_i
$x_1^{(i)},5$
$x_2^{(i)},3$
$\vdots$
$x_{M^{(i)}}^{(i)},2$
Transform
q_i
$\{(x_1^{(i)},c_4),(x_2^{(i)},c_3),\ldots,(x_{M^{(i)}}^{(i)},c_1)\}$
$c_1 \prec c_2 \prec c_3 \prec c_4$

Background——Approach

Three ways for **supervised text ranking**

2.Pairwise

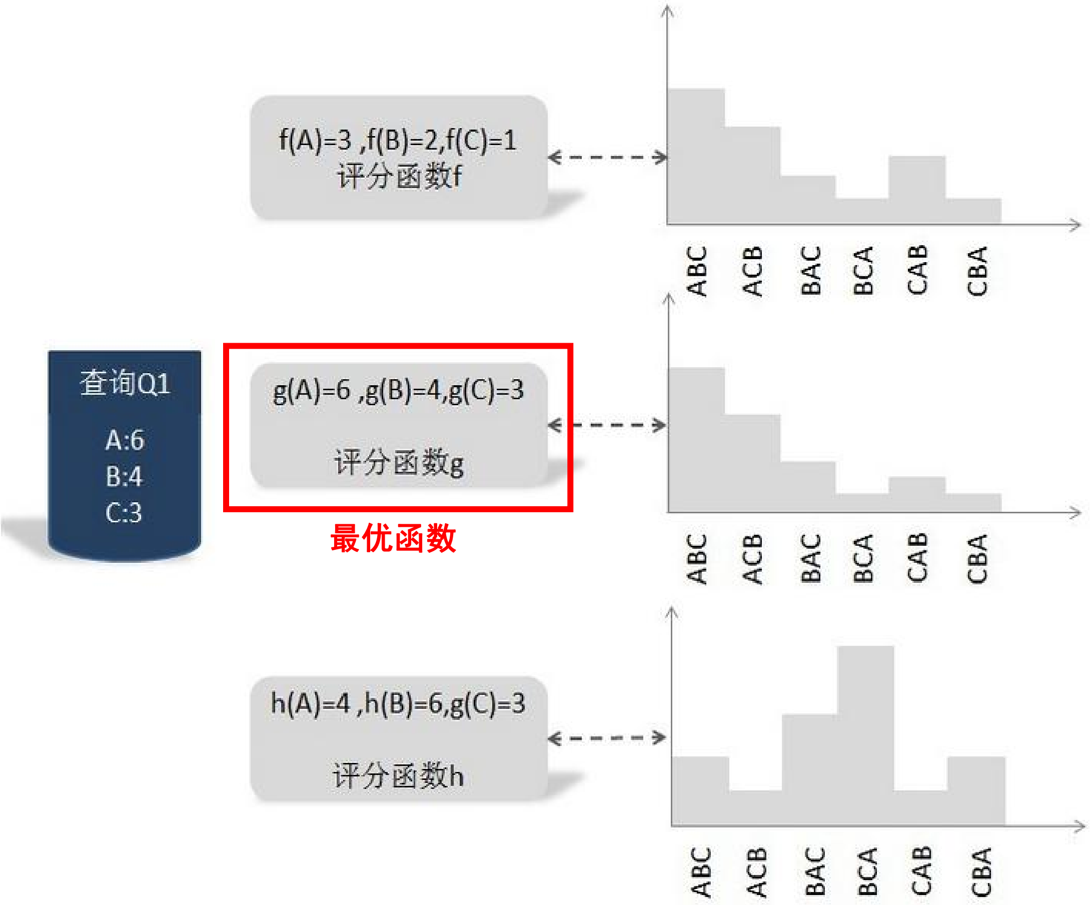
对于搜索系统来说，系统接收到用户查询后，返回相关文档列表，所以问题的关键是确定文档之间的先后顺序关系。单文档方法完全从单个文档的分类得分角度计算，没有考虑文档之间的顺序关系。文档对方法将排序问题转化为多个pair的排序问题，比较不同文章的先后顺序。


q_i
$x_1^{(i)},5$
$x_2^{(i)},3$
$\vdots$
$x_{n^{(i)}}^{(i)},2$
Transform
q_i
$(x_1^{(i)},x_2^{(i)},+1),(x_2^{(i)},x_1^{(i)},-1),\ldots,$
$(x_2^{(i)},x_{n^{(i)}}^{(i)},+1),(x_{n^{(i)}}^{(i)},x_2^{(i)},-1)$

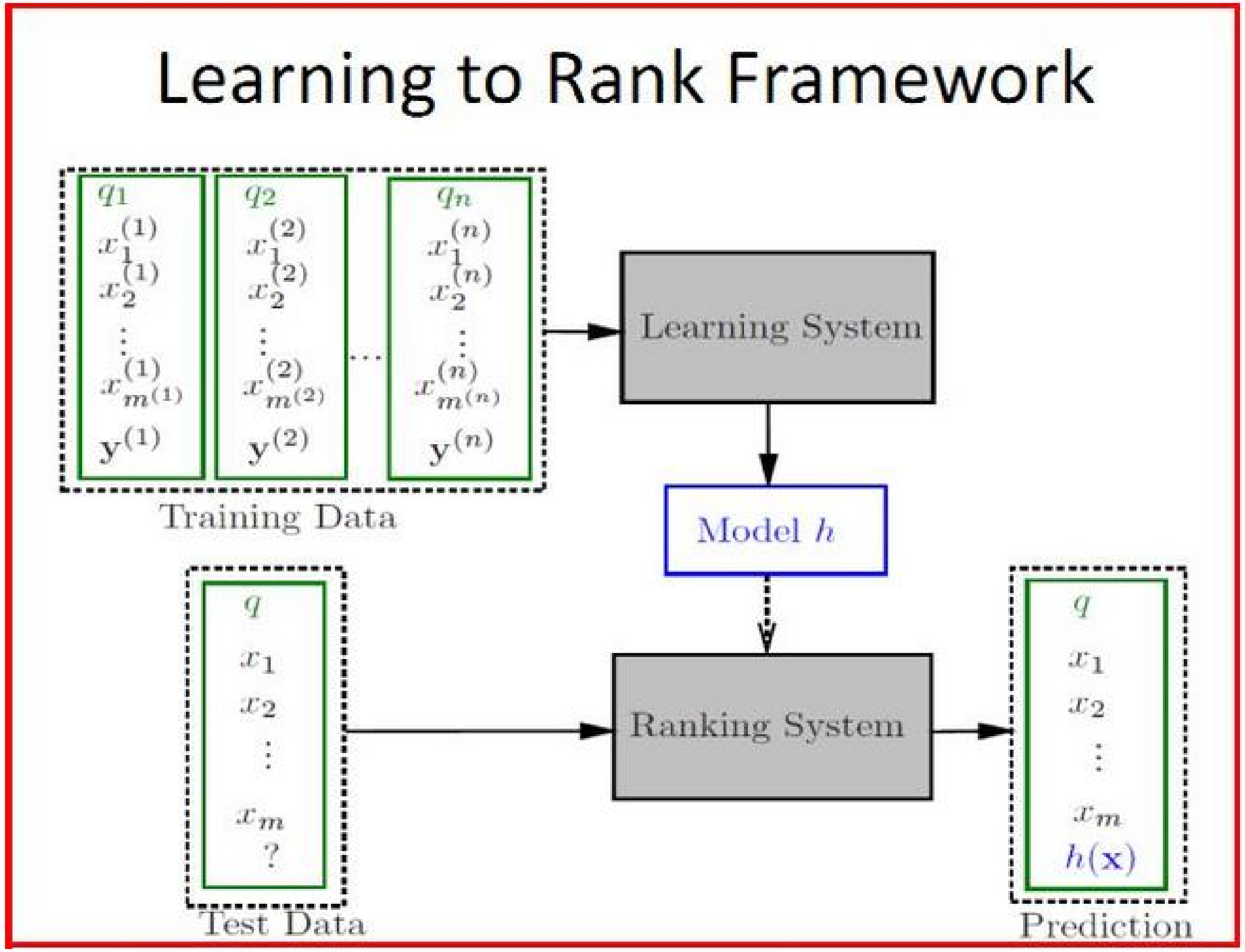
Three ways for **supervised text ranking**

3.Listwise

文档列表方法根据K个训练实例（一个查询及其对应的所有搜索结果评分作为一个实例）训练得到最优评分函数F，对于一个新的用户查询，函数F 对每一个文档打分，之后按照得分顺序由高到低排序，就是对应的搜索结果。



Background——Framework



Model——HD-LSTM

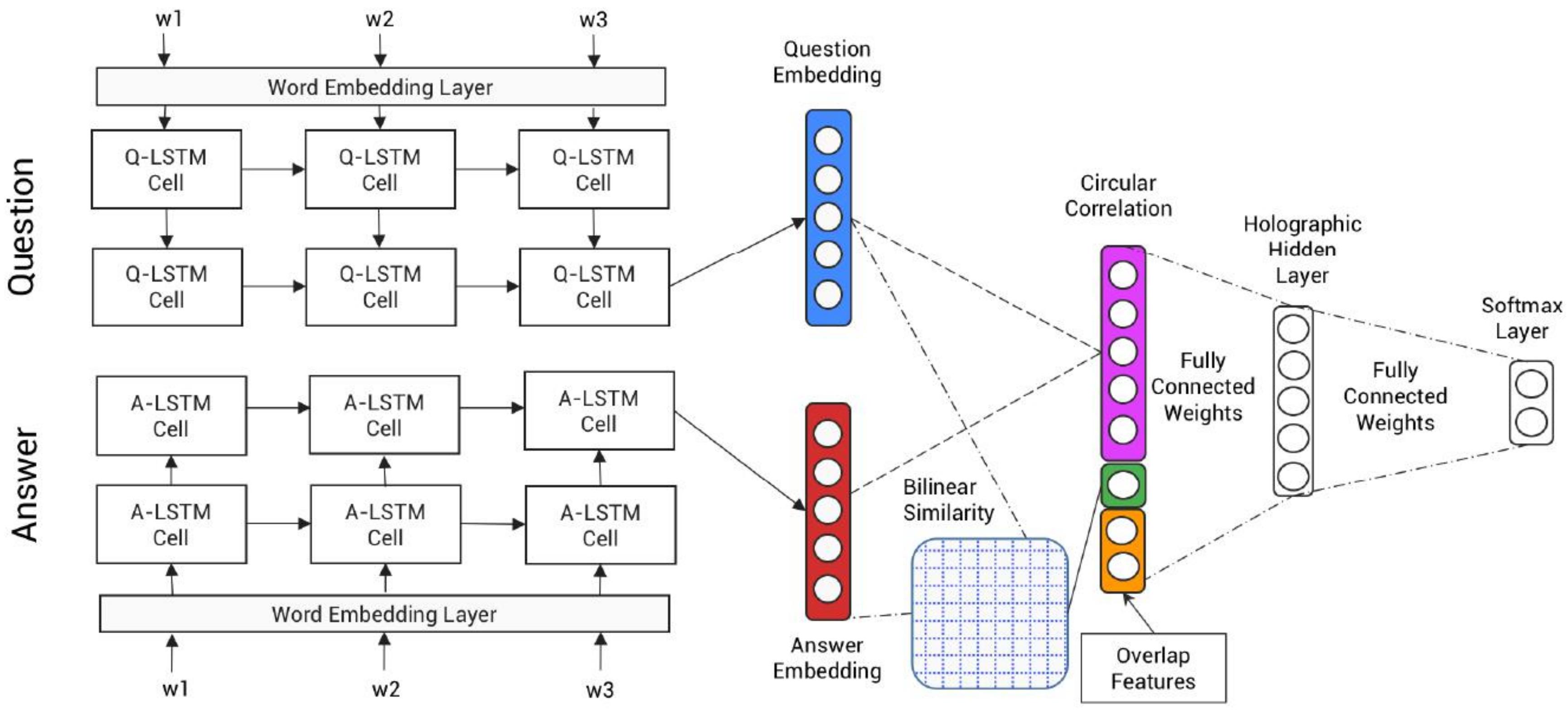


Figure 1: Holographic Dual LSTM Deep Learning Model for Ranking of QA Pairs

Model——Embedding

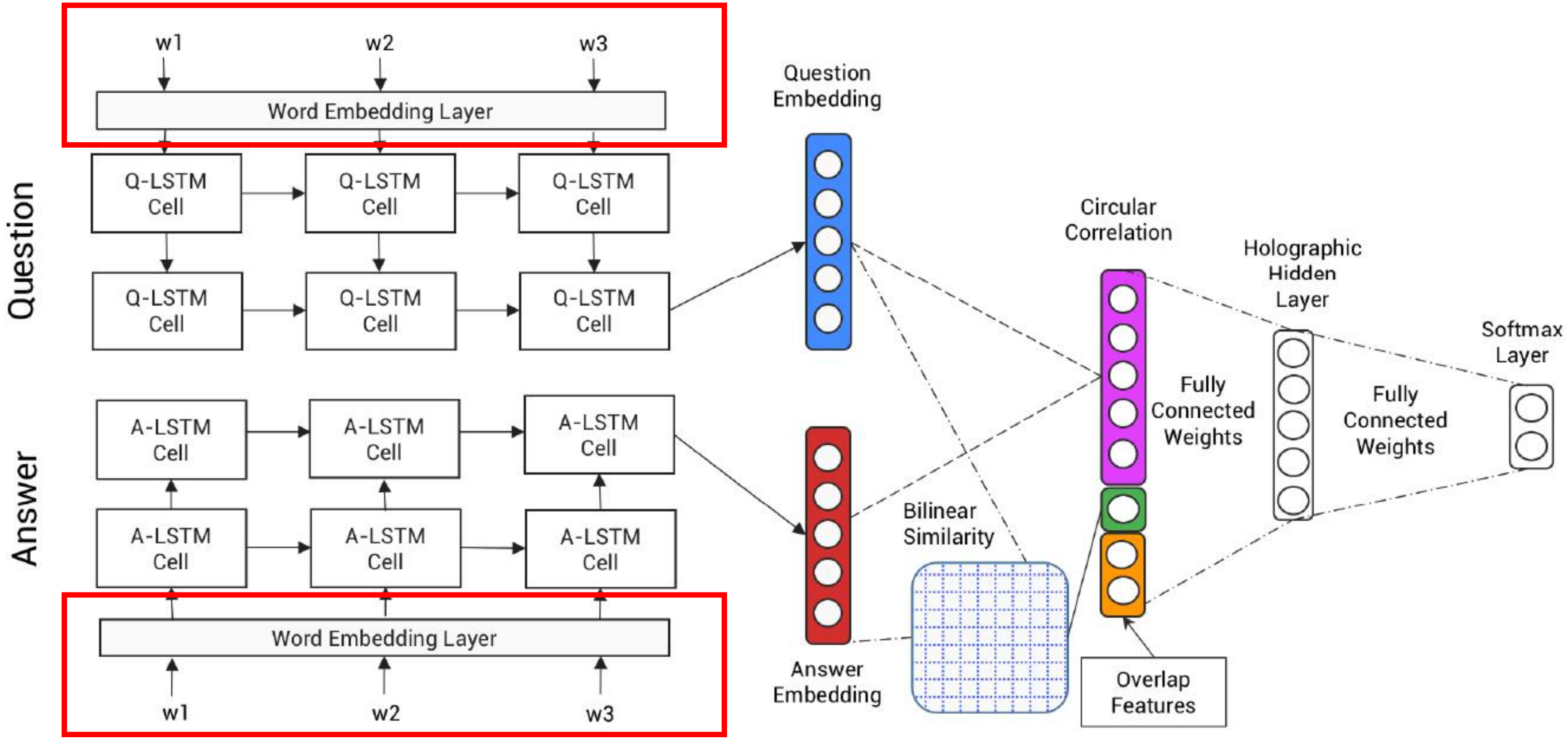
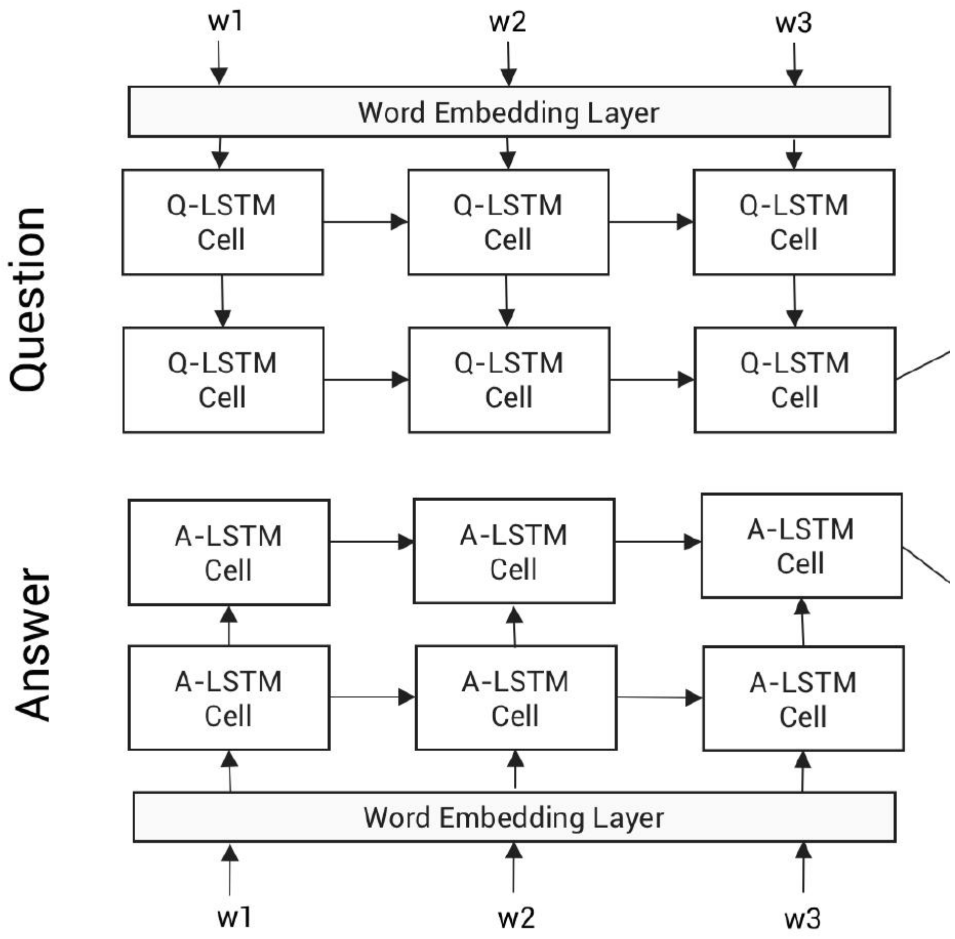


Figure 1: Holographic Dual LSTM Deep Learning Model for Ranking of QA Pairs

Model——Embedding



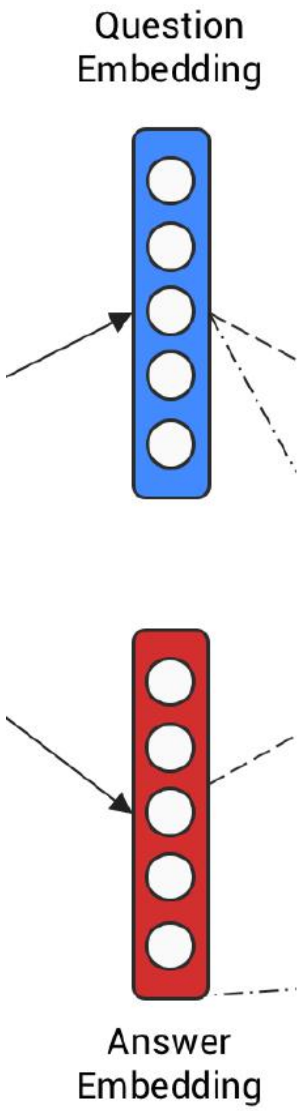
The parameters of this layer are $W \in \mathbb{R}^{|V| \times n}$

initialize W： pretrained SkipGran embeddings

Output:

the last hidden output from Q-LSTM and A-LSTM are taken to be **the final representation for question and answer respectively**

Model——Embedding



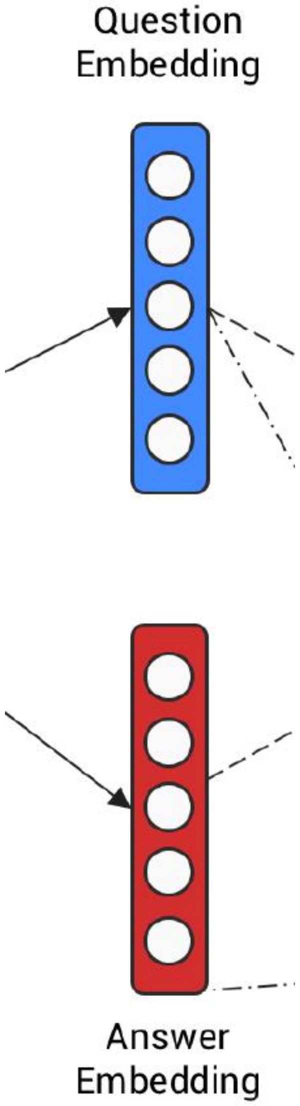
The parameters of this layer are $W \in \mathbb{R}^{|V|\times n}$

initialize W： pretrained SkipGran embeddings

Output:

the last hidden output from Q-LSTM and A-LSTM are taken to be **the final representation for question and answer respectively**

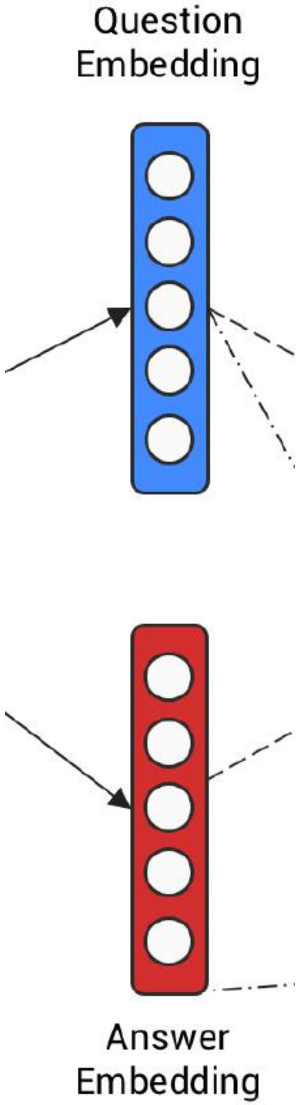
Model——Holographic Matching of QA pairs



We denote q**o**a as a compositional operator applied to vectors q and a.

$$q \circ a$$

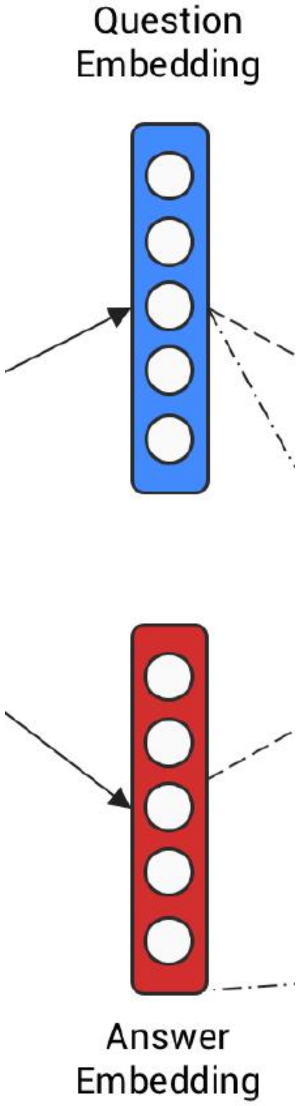
Model——Holographic Matching of QA pairs



We denote q**o**a as a compositional operator applied to vectors q and a.

$$q \circ a = q \star a$$

Model——Holographic Matching of QA pairs



We employ the **circular correlation** of vectors to learn relationships between question and answer embeddings

$$q \circ a = q \star a$$

Model——Holographic Matching of QA pairs

$$q \circ a = q \star a$$

We employ the **circular correlation** of vectors to learn relationships between question and answer embeddings

$$\star : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$$

Model——Holographic Matching of QA pairs

$$q \circ a = q \star a$$

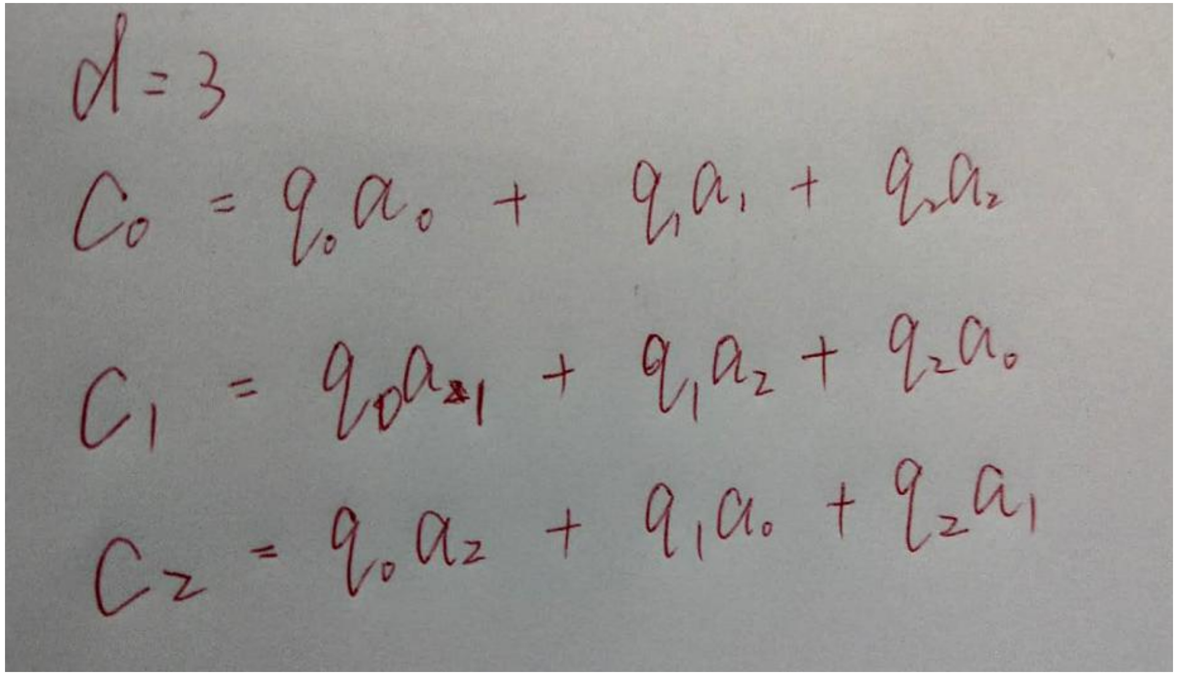
We employ the **circular correlation** of vectors to learn relationships between question and answer embeddings

$$[q \star a]_k = \sum_{i=0}^{d-1} q_i \, a_{(k+i) \mod d}$$

Model——Holographic Matching of QA pairs

$$[q \star a]_k = \sum_{i=0}^{d-1} q_i \, a_{(k+i) \bmod d}$$

We employ the **circular correlation** of vectors to learn relationships between question and answer embeddings



$d=3$

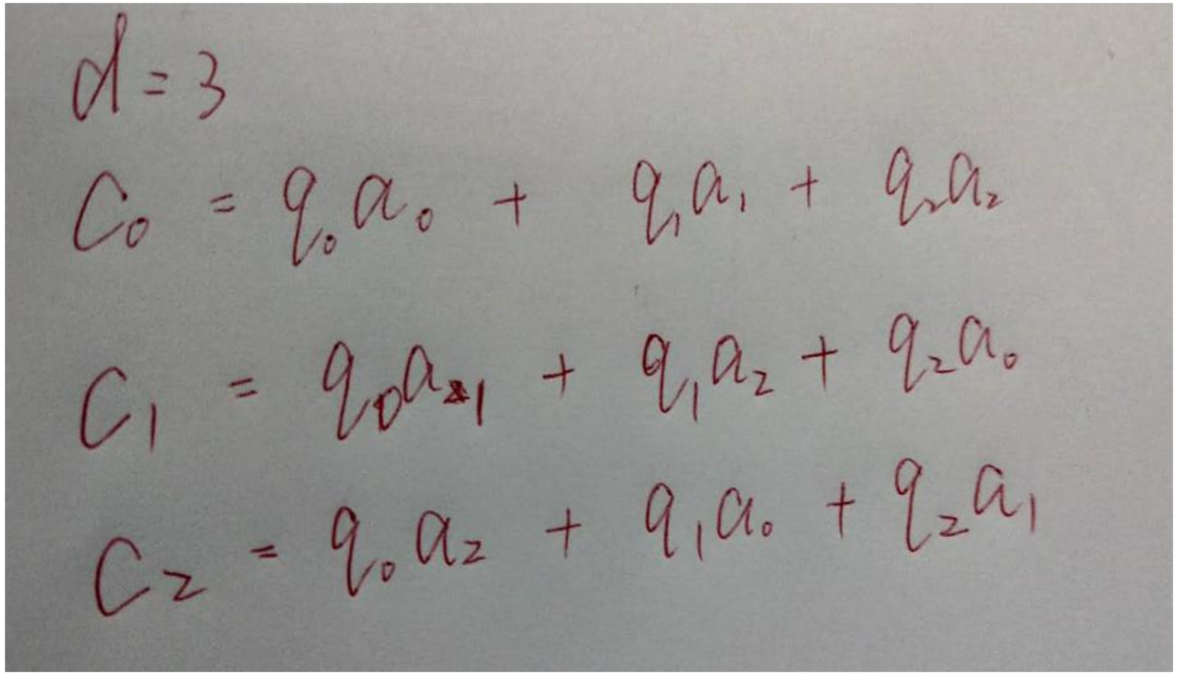
$c_0 = q_0 a_0 + q_1 a_1 + q_2 a_2$

$c_1 = q_0 a_1 + q_1 a_2 + q_2 a_0$

$c_2 = q_0 a_2 + q_1 a_0 + q_2 a_1$

Model——Holographic Matching of QA pairs

We employ the **circular correlation** of vectors to learn relationships between question and answer embeddings

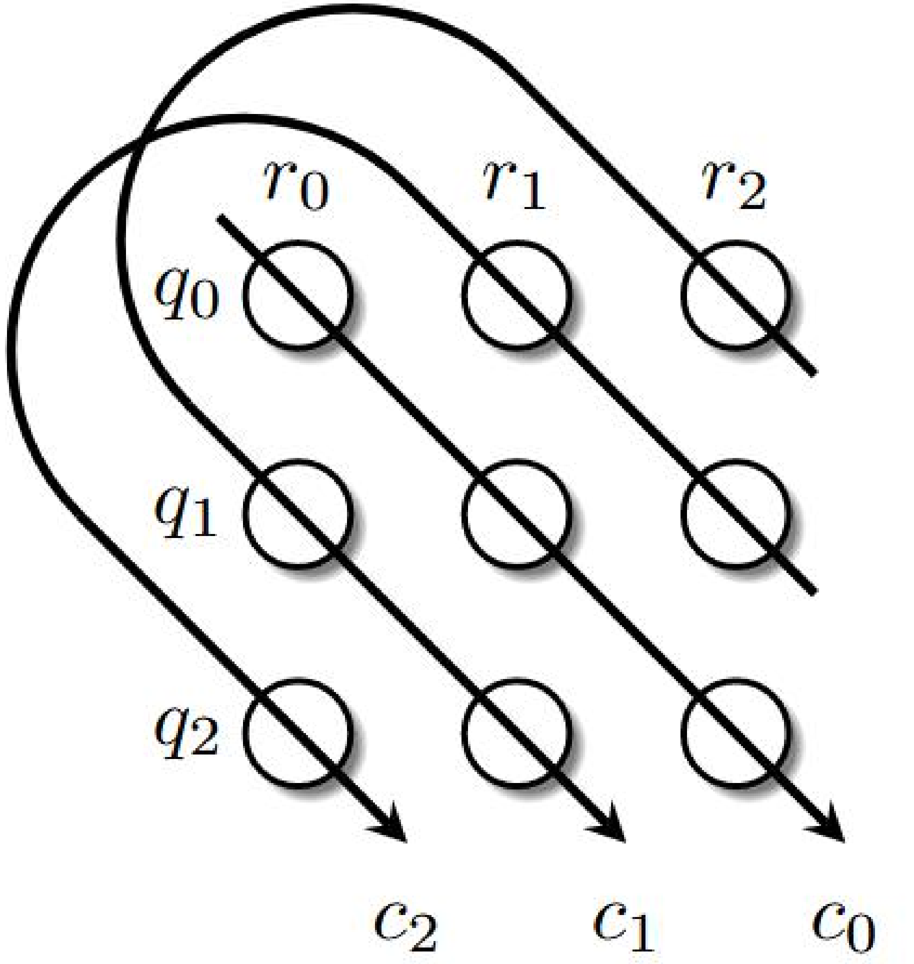


$d=3$

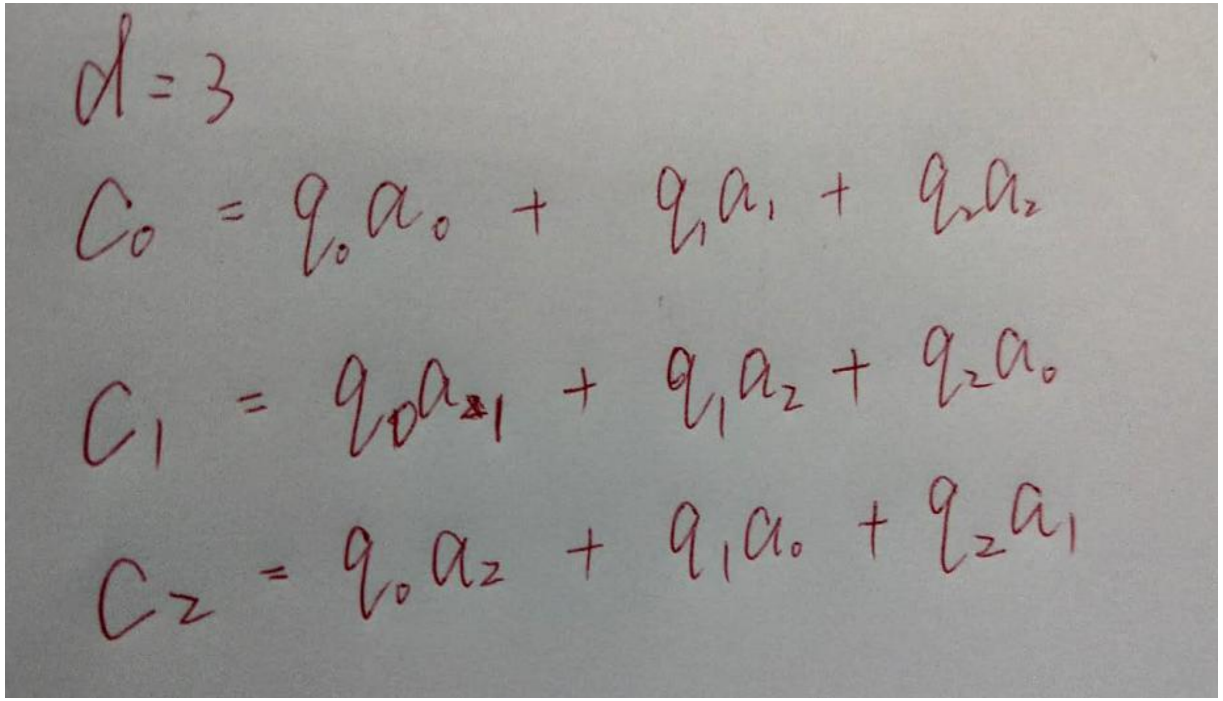
$c_0 = q_0 a_0 + q_1 a_1 + q_2 a_2$

$c_1 = q_0 a_1 + q_1 a_2 + q_2 a_0$

$c_2 = q_0 a_2 + q_1 a_0 + q_2 a_1$



Model——Holographic Matching of QA pairs

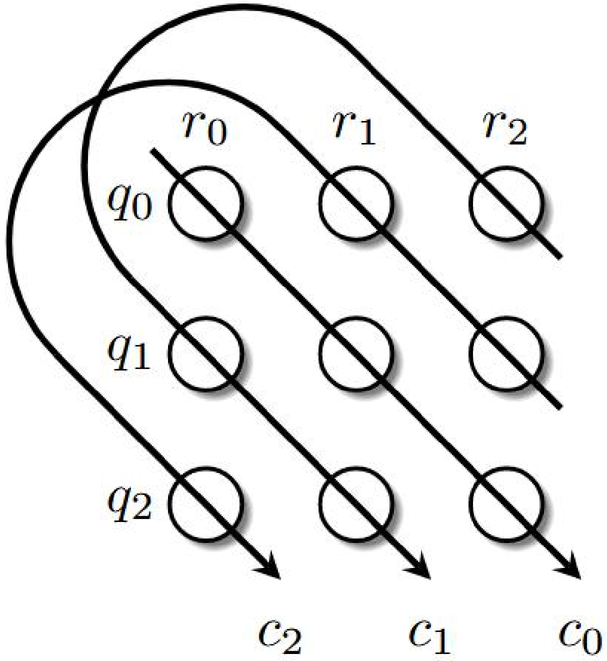


$d=3$

$c_0 = q_0a_0 + q_1a_1 + q_2a_2$

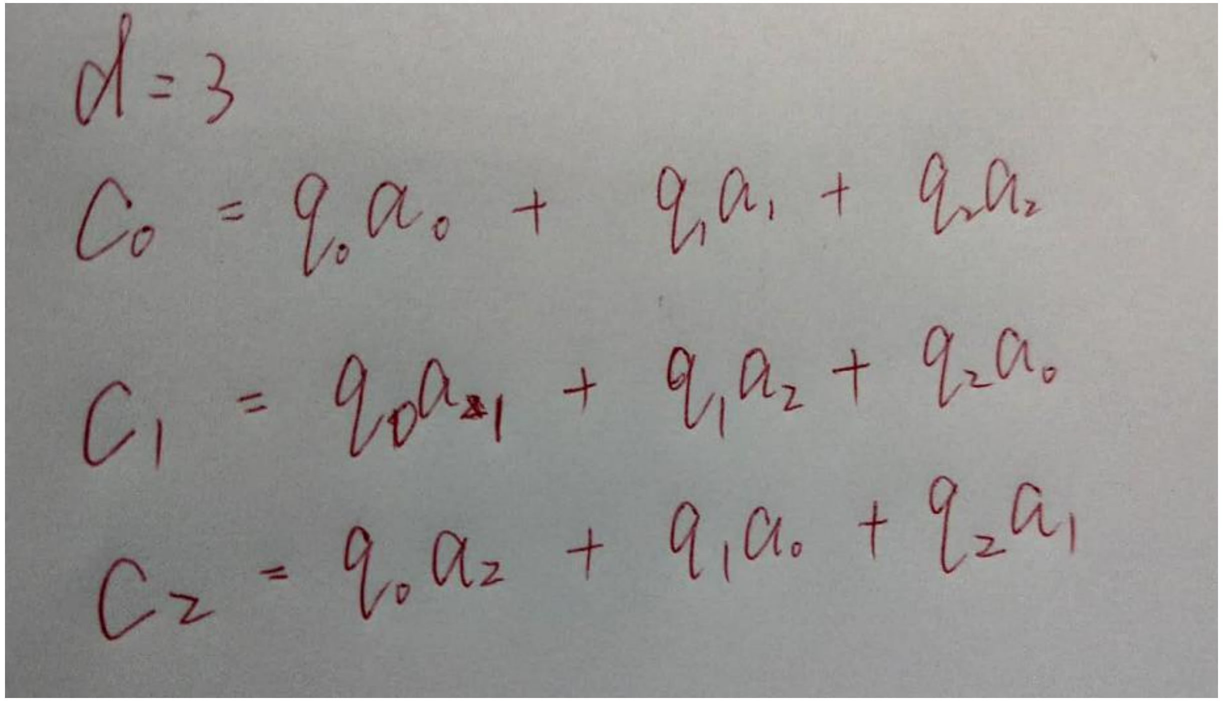
$c_1 = q_0a_1 + q_1a_2 + q_2a_0$

$c_2 = q_0a_2 + q_1a_0 + q_2a_1$



$q \circ a = q \star a$ → $[q \star a]_k = \sum_{i=0}^{d-1} q_i \, a_{(k+i) \bmod d}$ → $q \star a = \mathcal{F}^{-1}(\overline{\mathcal{F}(q)} \odot \mathcal{F}(a))$

Model——Holographic Matching of QA pairs

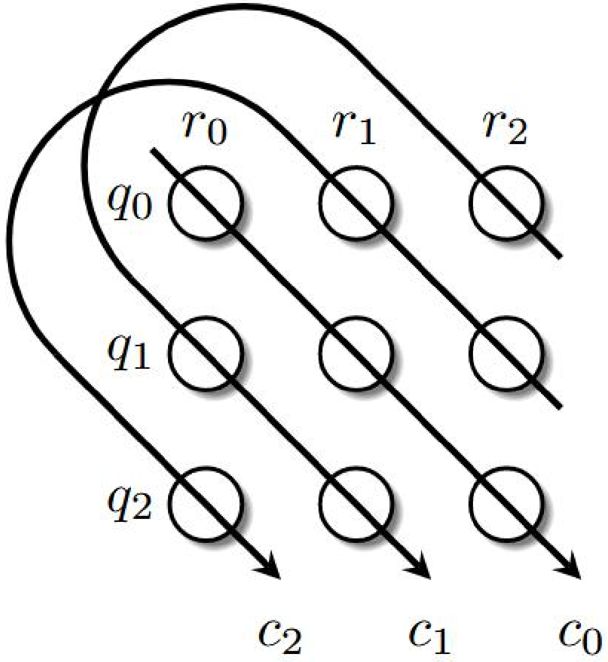


$d=3$

$c_0 = q_0 a_0 + q_1 a_1 + q_2 a_2$

$c_1 = q_0 a_1 + q_1 a_2 + q_2 a_0$

$c_2 = q_0 a_2 + q_1 a_0 + q_2 a_1$

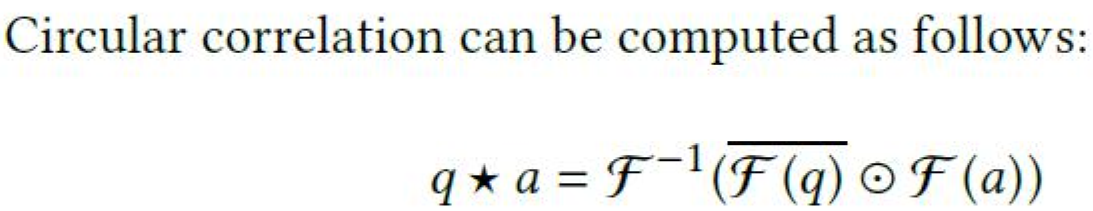


$$q \circ a = q \star a$$

$$[q \star a]_k = \sum_{i=0}^{d-1} q_i \, a_{(k+i) \bmod d}$$

$$\boxed{q \star a = \mathcal{F}^{-1}(\overline{\mathcal{F}(q)} \odot \mathcal{F}(a))}$$

Model——Holographic Matching of QA pairs



Circular correlation can be computed as follows:

$$q \star a = \mathcal{F}^{-1}(\overline{\mathcal{F}(q)} \odot \mathcal{F}(a))$$

$\mathcal{F}(.)$	fast Fourier transform (FFT)
$\mathcal{F}^{-1}(.)$	inverse fast Fourier transform
$\overline{\mathcal{F}(q)}$	the complex conjugate

vector c is the result of composing q and a with circular correlation.

Model——Holographic Matching of QA pairs

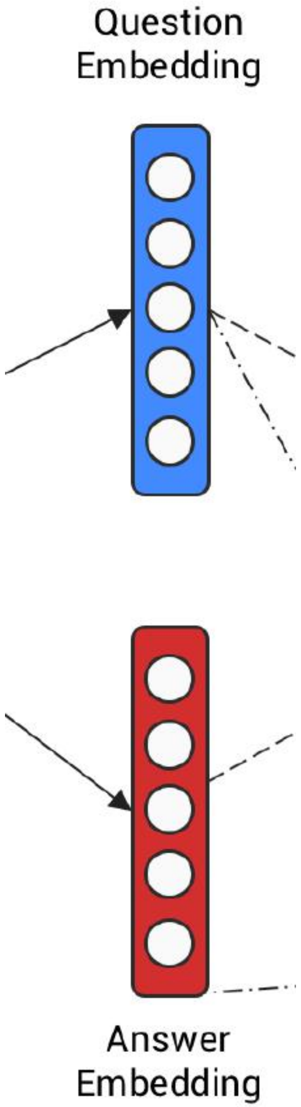
Circular correlation can be computed as follows:

$$q \star a = \mathcal{F}^{-1}(\overline{\mathcal{F}(q)} \odot \mathcal{F}(a))$$

$\mathcal{F}(.)$	fast Fourier transform (FFT)
$\mathcal{F}^{-1}(.)$	inverse fast Fourier transform
$\overline{\mathcal{F}(q)}$	the complex conjugate

One key advantage of this composition method is that there are **no increase in parameters**.

Model——Holographic Matching of QA pairs



In the case where question and answer representations are of **different dimensions**, we can simply zero-pad the vectors to make them the **same length**.

Model——Holographic Matching of QA pairs

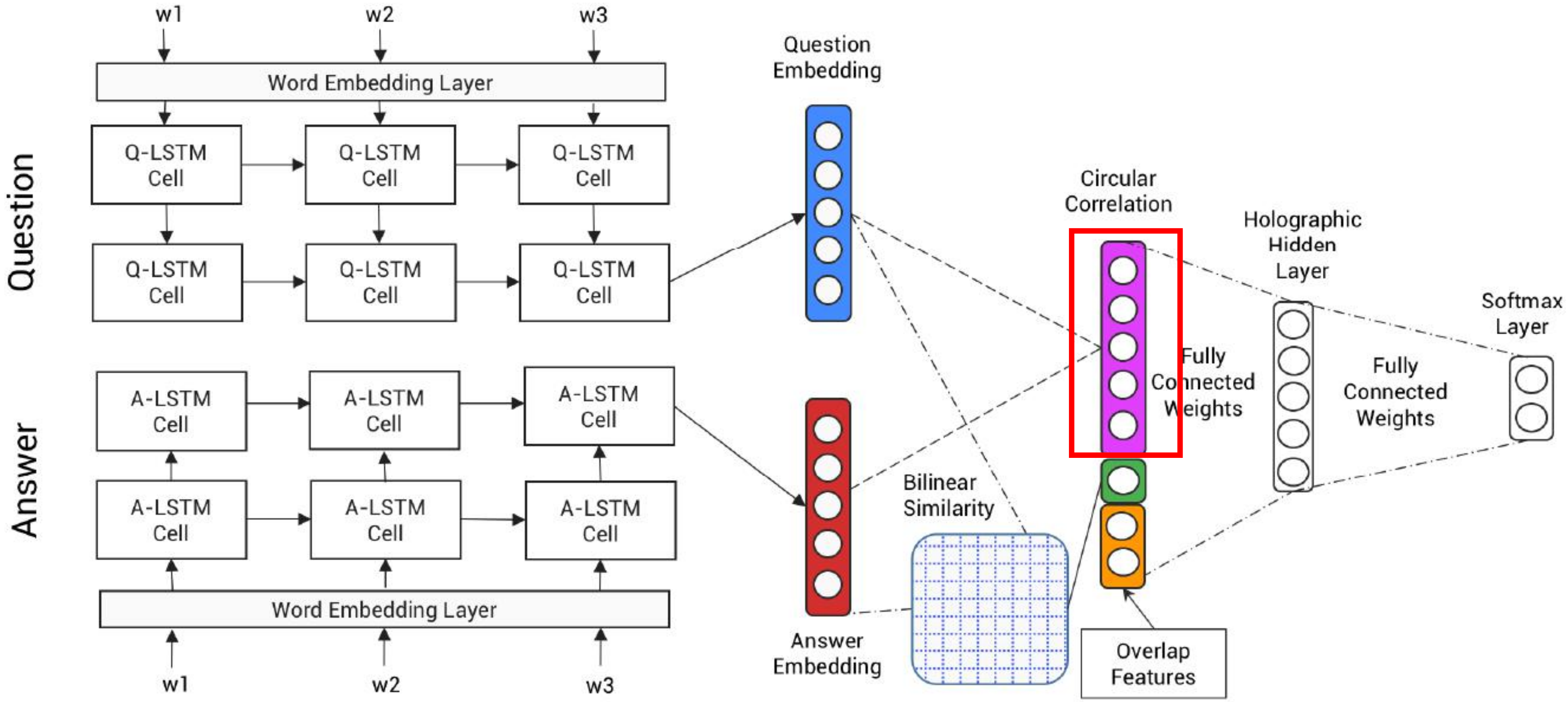


Figure 1: Holographic Dual LSTM Deep Learning Model for Ranking of QA Pairs

Model——Holographic Hidden Layer

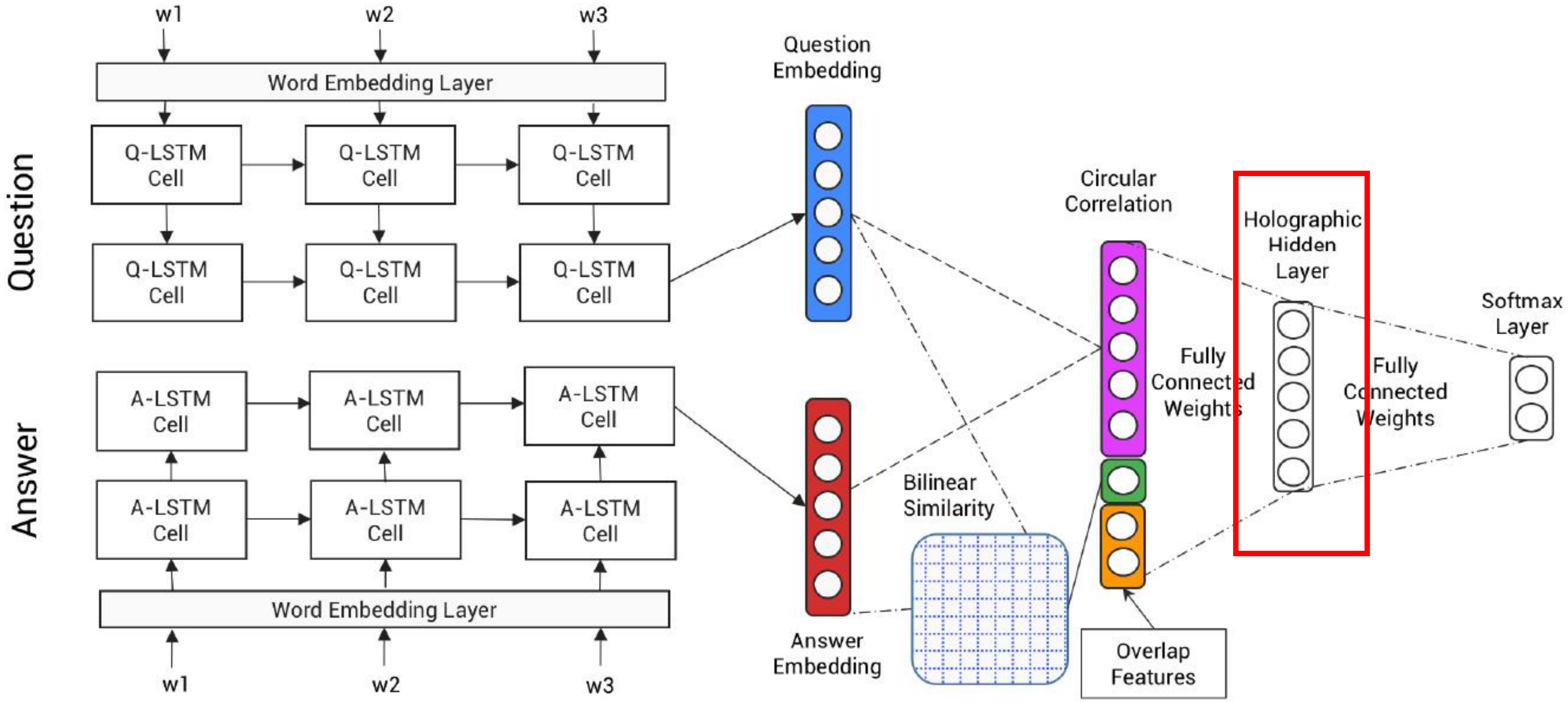


Figure 1: Holographic Dual LSTM Deep Learning Model for Ranking of QA Pairs

Model——Holographic Hidden Layer

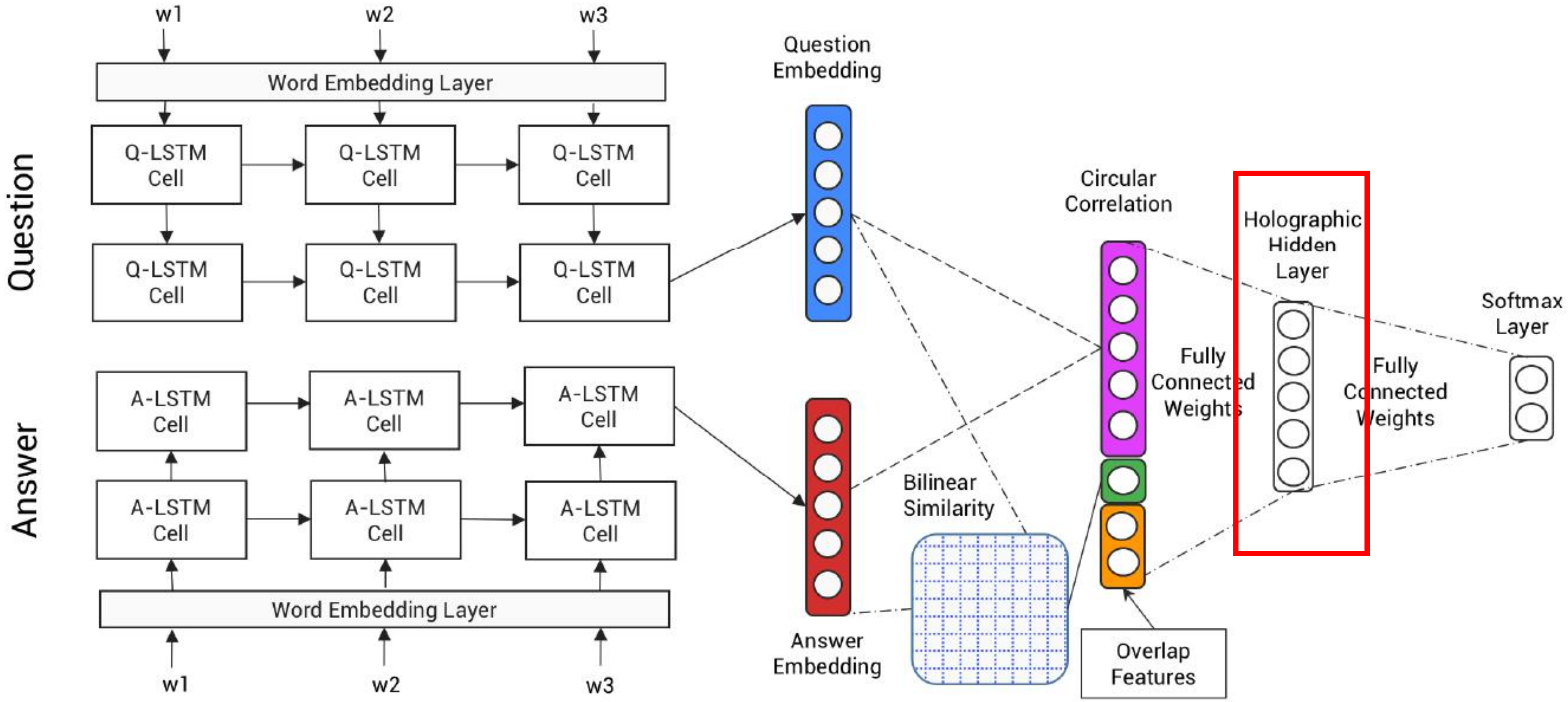
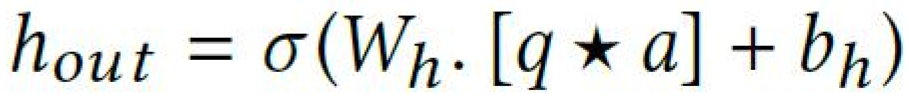


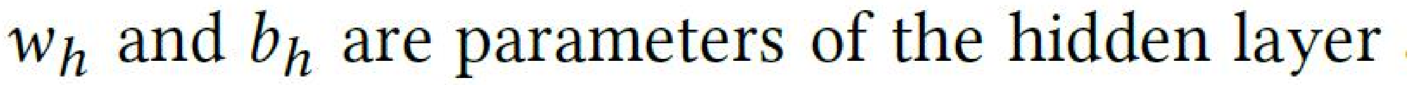
Figure 1: Holographic Dual LSTM Deep Learning Model for Ranking of QA Pairs

a fully connected hidden layer follows our compositional operator

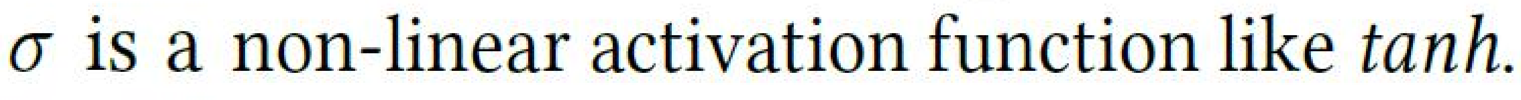
Model——Holographic Hidden Layer



$$h_{out} = \sigma(W_h . [q \star a] + b_h)$$



w_h and b_h are parameters of the hidden layer



σ is a non-linear activation function like *tanh*.

a fully connected hidden layer follows our compositional operator

Model——Holographic Hidden Layer

$$h_{out} = \sigma(W_h.\,[q \star a] + b_h)$$

Traditionally composition operator

$$\oplus : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \rightarrow \mathbb{R}^{d_1+d_2}$$

Our composition operator

$$\star : \mathbb{R}^{d} \times \mathbb{R}^{d} \rightarrow \mathbb{R}^{d}$$

Concatenation does not consider the relationship between latent features of QA embeddings.

Model——Softmax Layer

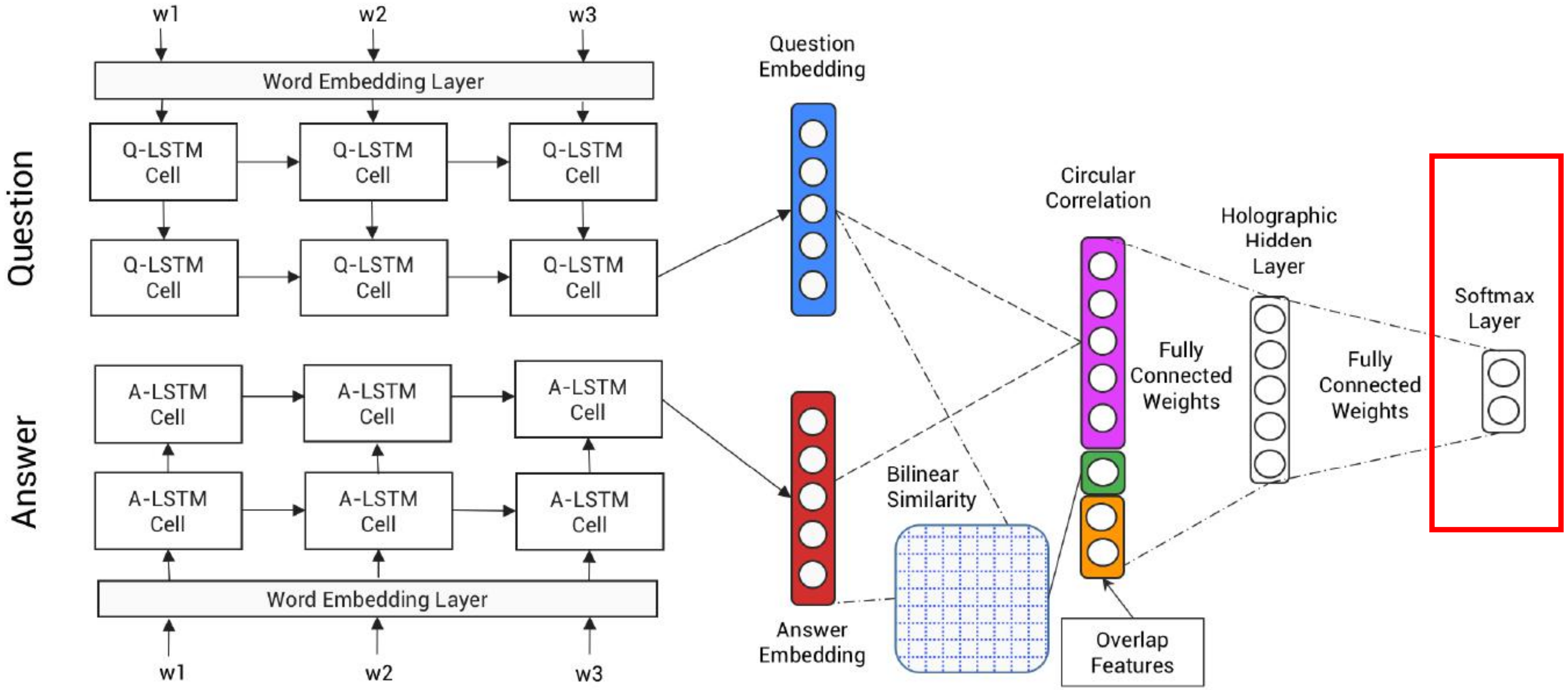


Figure 1: Holographic Dual LSTM Deep Learning Model for Ranking of QA Pairs

Model——Softmax Layer

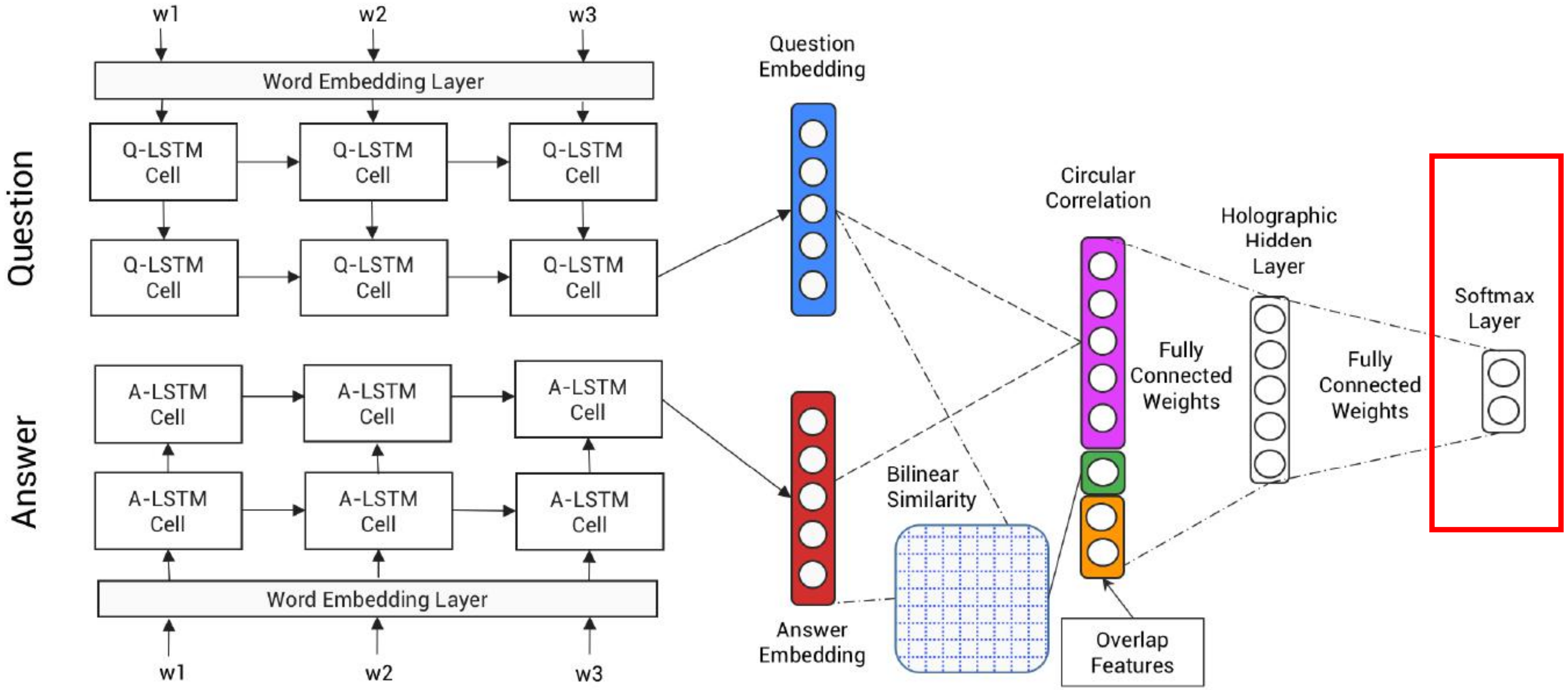


Figure 1: Holographic Dual LSTM Deep Learning Model for Ranking of QA Pairs

$$p = softmax(W_f . h_{out} + b_f)$$

Model——Incorporating Additional Features

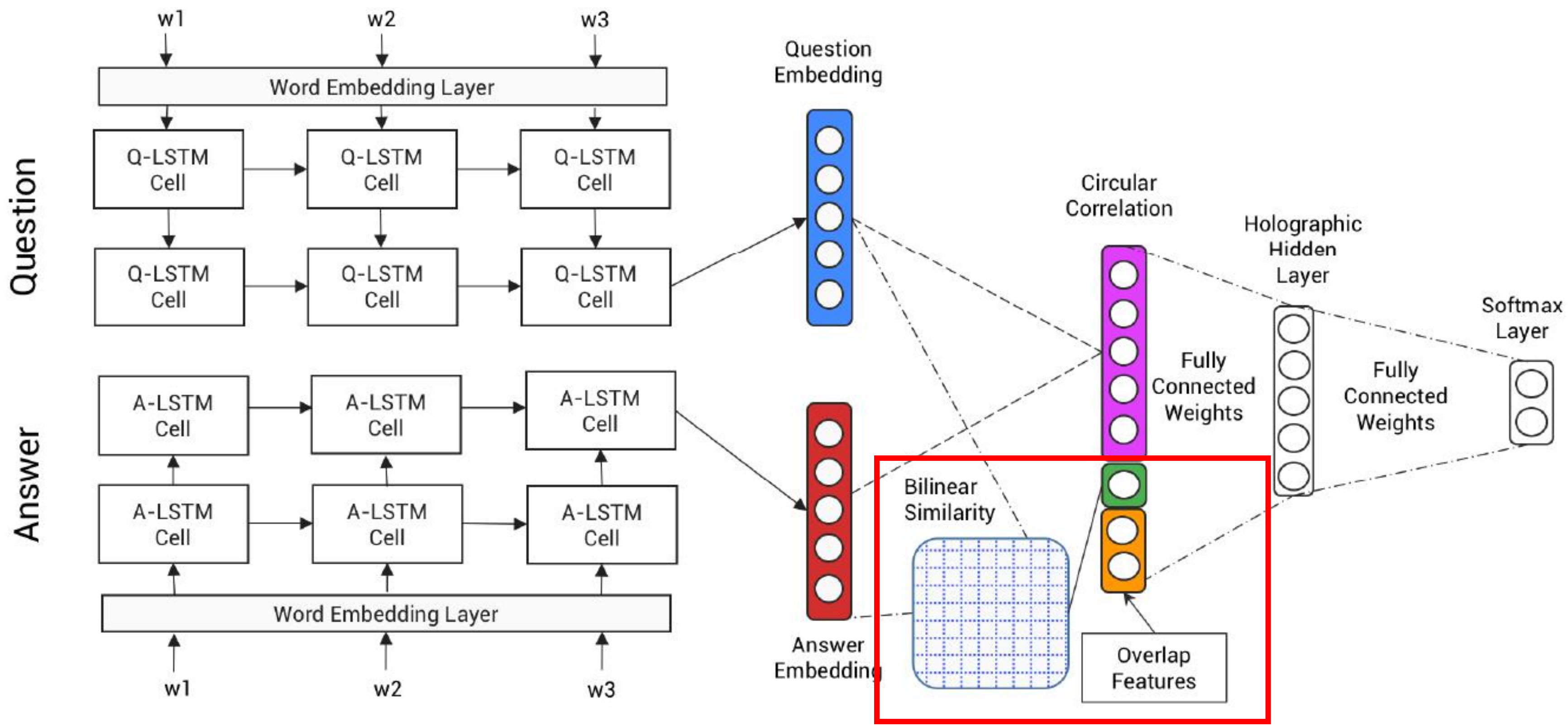
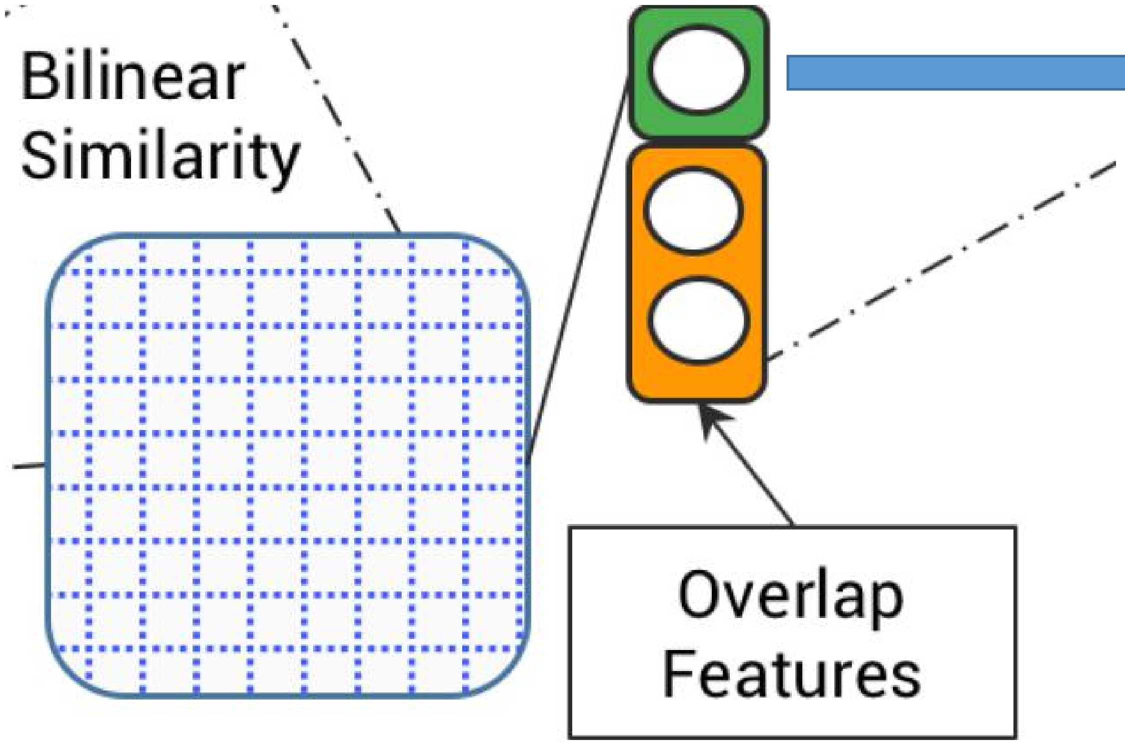


Figure 1: Holographic Dual LSTM Deep Learning Model for Ranking of QA Pairs

Model——Incorporating Additional Features

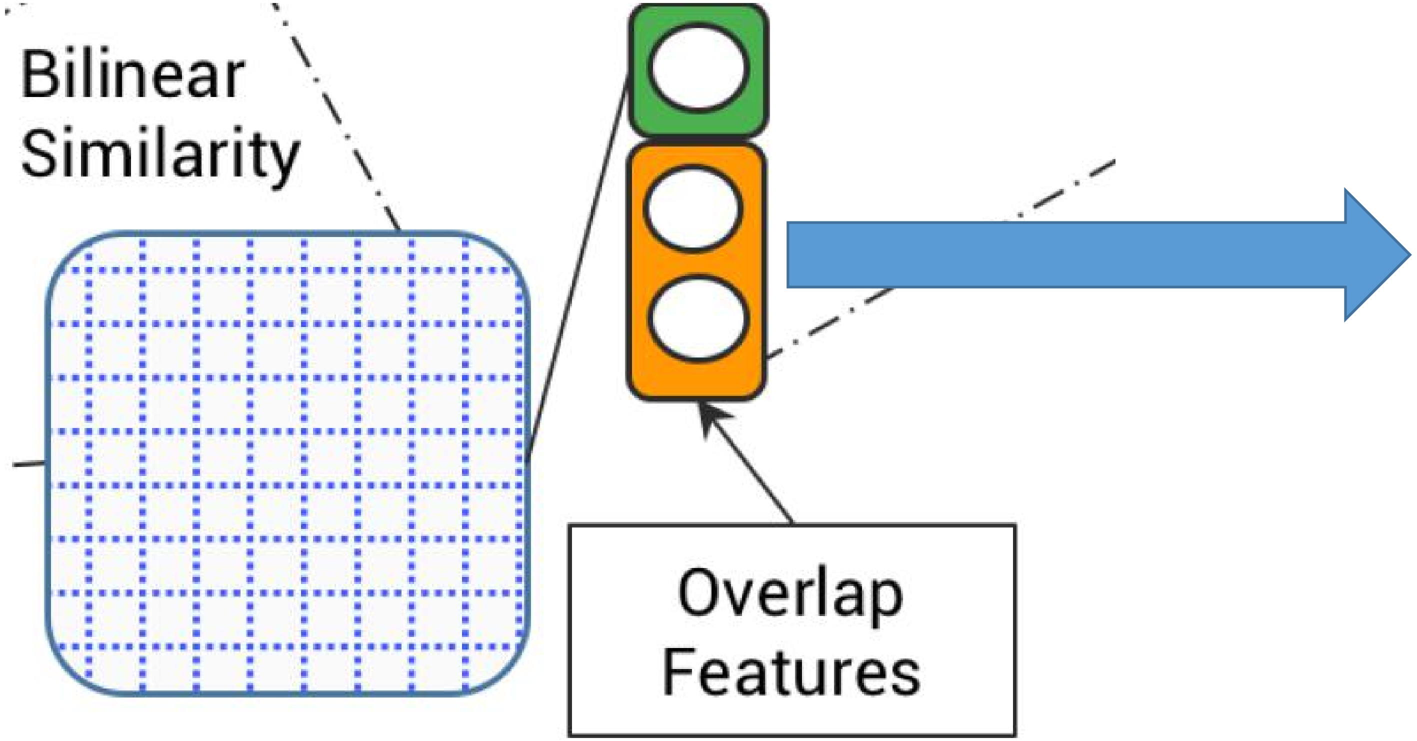


an additional **similarity measure** in our model between QA embeddings

$$sim(q, a) = \vec{q}^{\,T} M \vec{a}$$

$M \in \mathbb{R}^{n \times n}$ is a similarity matrix

Model——Incorporating Additional Features



X_{feat}

word overlap features

Model——Incorporating Additional Features

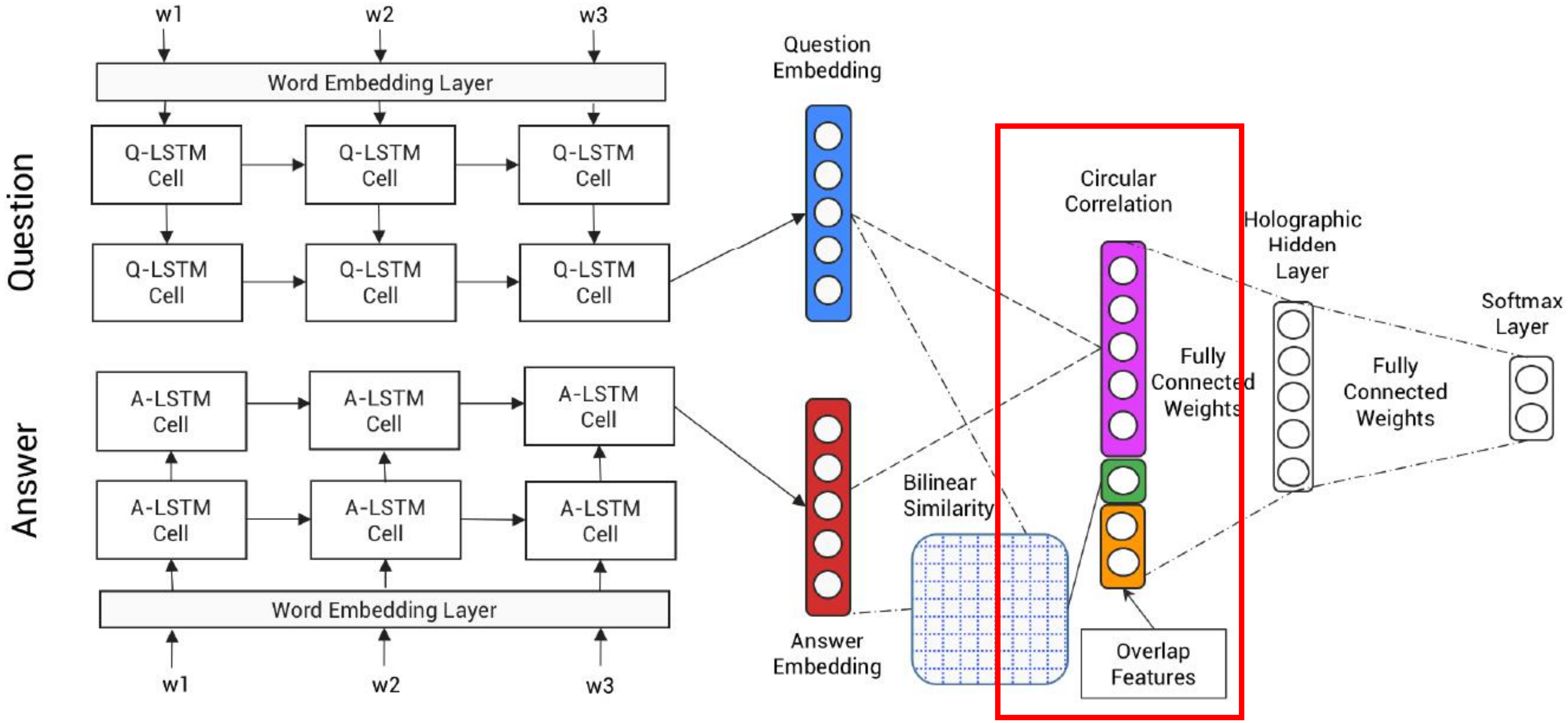


Figure 1: Holographic Dual LSTM Deep Learning Model for Ranking of QA Pairs

$$[[q \star a], sim(q,a), X_{feat}].$$

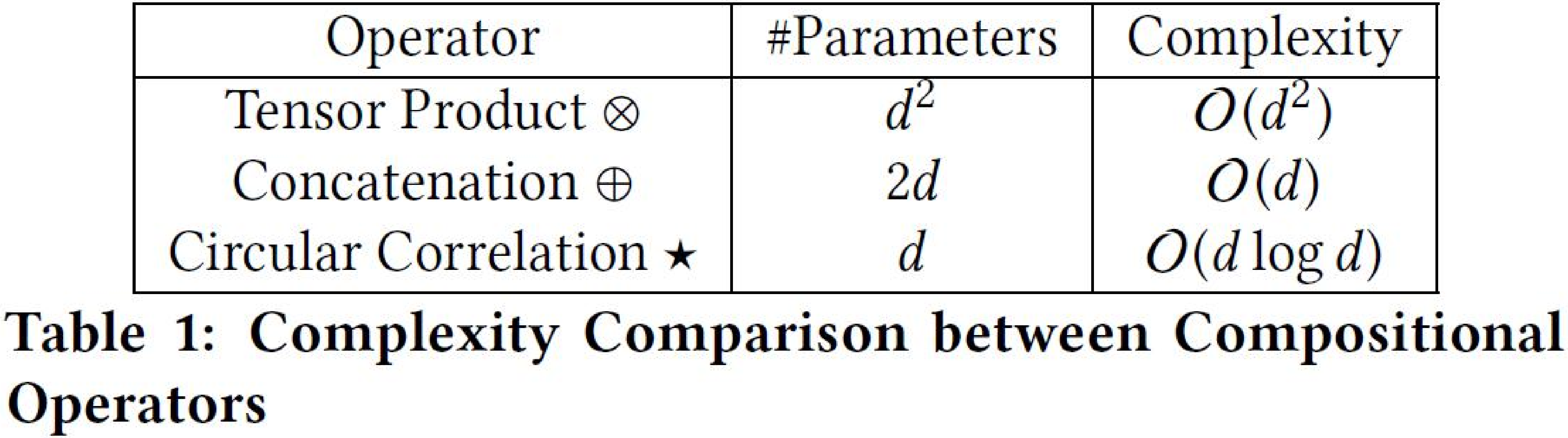
The prime contributions of the paper

~~1.We adopt holographic composition for modeling the interaction between representations of QA pairs~~

~~2.We present a novel deep learning architecture,HD-LSTM (Holographic Dual LSTM) for learning to rank QA pairs~~

3.We provide extensive experimental evidence of the effectiveness of our model on both factoid question answering and community-based question answering.

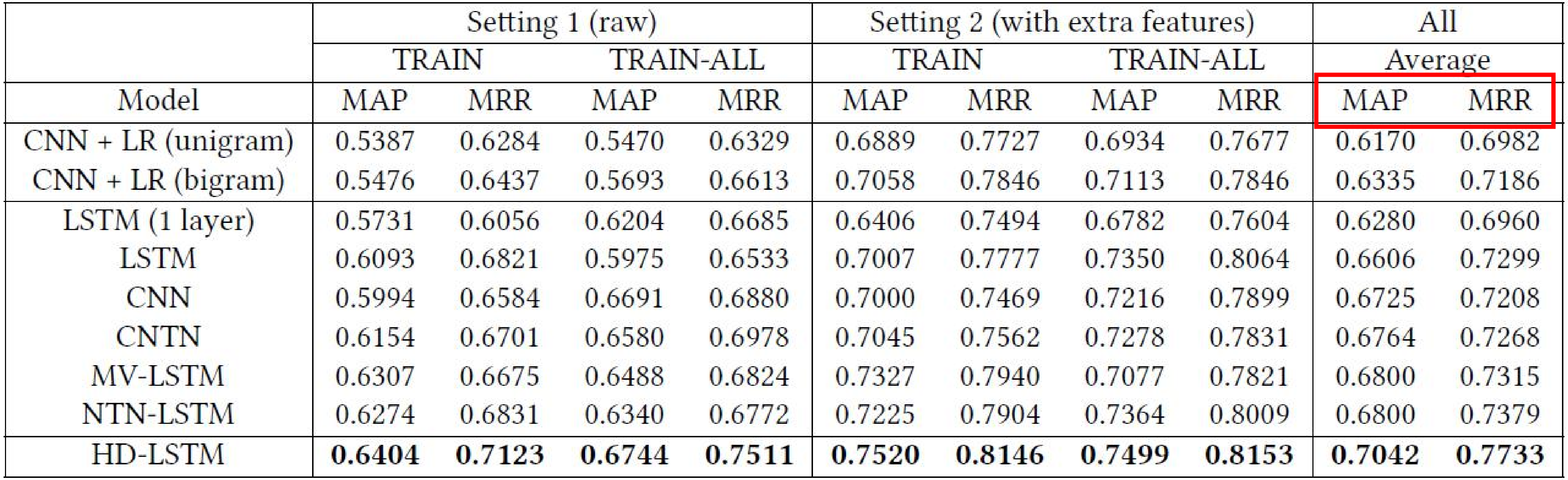
Results——Complexity Analysis



Operator	#Parameters	Complexity
Tensor Product $\otimes$	d^2	$\mathcal{O}(d^2)$
Concatenation $\oplus$	$2d$	$\mathcal{O}(d)$
Circular Correlation $\star$	d	$\mathcal{O}(d \log d)$

Table 1: Complexity Comparison between Compositional Operators

Results——Experimental Results on TREC QA



	Setting 1 (raw)				Setting 2 (with extra features)				All	
	TRAIN		TRAIN-ALL		TRAIN		TRAIN-ALL		Average	
Model	MAP	MRR	MAP	MRR	MAP	MRR	MAP	MRR	MAP	MRR
CNN + LR (unigram)	0.5387	0.6284	0.5470	0.6329	0.6889	0.7727	0.6934	0.7677	0.6170	0.6982
CNN + LR (bigram)	0.5476	0.6437	0.5693	0.6613	0.7058	0.7846	0.7113	0.7846	0.6335	0.7186
LSTM (1 layer)	0.5731	0.6056	0.6204	0.6685	0.6406	0.7494	0.6782	0.7604	0.6280	0.6960
LSTM	0.6093	0.6821	0.5975	0.6533	0.7007	0.7777	0.7350	0.8064	0.6606	0.7299
CNN	0.5994	0.6584	0.6691	0.6880	0.7000	0.7469	0.7216	0.7899	0.6725	0.7208
CNTN	0.6154	0.6701	0.6580	0.6978	0.7045	0.7562	0.7278	0.7831	0.6764	0.7268
MV-LSTM	0.6307	0.6675	0.6488	0.6824	0.7327	0.7940	0.7077	0.7821	0.6800	0.7315
NTN-LSTM	0.6274	0.6831	0.6340	0.6772	0.7225	0.7904	0.7364	0.8009	0.6800	0.7379
HD-LSTM	**0.6404**	**0.7123**	**0.6744**	**0.7511**	**0.7520**	**0.8146**	**0.7499**	**0.8153**	**0.7042**	**0.7733**

Results——Experimental Results on TREC QA

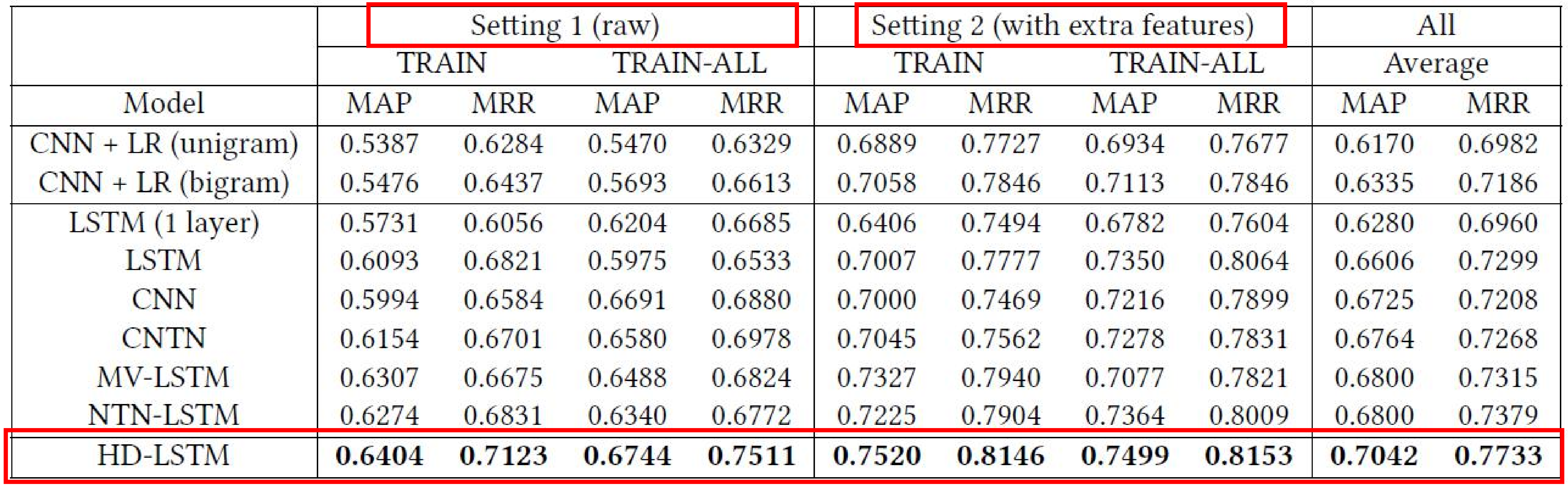
MAP(Mean Average Precision):

假设有两个主题，主题1有4个相关网页，主题2有5个相关网页。某系统对于主题1检索出4个相关网页，其rank分别为1, 2, 4, 7；对于主题2检索出3个相关网页，其rank分别为1,3,5。对于主题1，平均准确率为(1/1+2/2+3/4+4/7)/4=0.83。对于主题2，平均准确率为(1/1+2/3+3/5+0+0)/5=0.45。则MAP= (0.83+0.45)/2=0.64。

MRR(Mean Reciprocal Rank)：

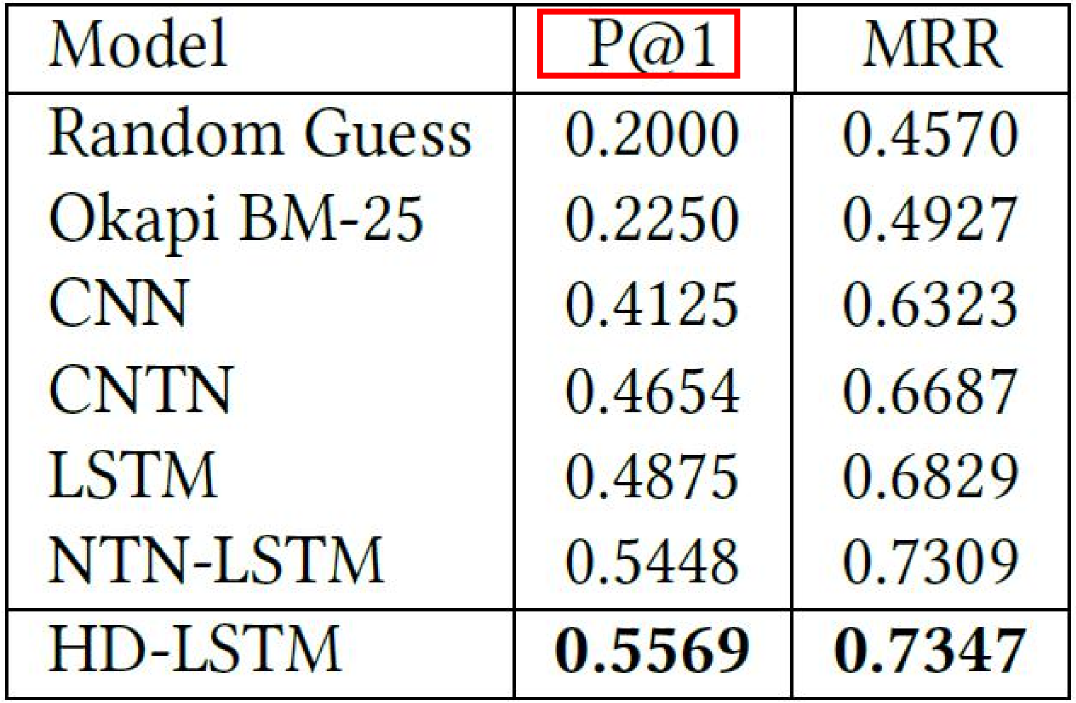
MRR（Mean reciprocal rank）是一个国际上通用的对搜索算法进行评价的机制，即第一个结果匹配，分数为1，第二个匹配分数为0.5，第n个匹配分数为1/n，如果没有匹配的句子分数为0。最终的分数为所有得分之和。

Results——Experimental Results on TREC QA



	Setting 1 (raw)				Setting 2 (with extra features)				All	
	TRAIN		TRAIN-ALL		TRAIN		TRAIN-ALL		Average	
Model	MAP	MRR	MAP	MRR	MAP	MRR	MAP	MRR	MAP	MRR
CNN + LR (unigram)	0.5387	0.6284	0.5470	0.6329	0.6889	0.7727	0.6934	0.7677	0.6170	0.6982
CNN + LR (bigram)	0.5476	0.6437	0.5693	0.6613	0.7058	0.7846	0.7113	0.7846	0.6335	0.7186
LSTM (1 layer)	0.5731	0.6056	0.6204	0.6685	0.6406	0.7494	0.6782	0.7604	0.6280	0.6960
LSTM	0.6093	0.6821	0.5975	0.6533	0.7007	0.7777	0.7350	0.8064	0.6606	0.7299
CNN	0.5994	0.6584	0.6691	0.6880	0.7000	0.7469	0.7216	0.7899	0.6725	0.7208
CNTN	0.6154	0.6701	0.6580	0.6978	0.7045	0.7562	0.7278	0.7831	0.6764	0.7268
MV-LSTM	0.6307	0.6675	0.6488	0.6824	0.7327	0.7940	0.7077	0.7821	0.6800	0.7315
NTN-LSTM	0.6274	0.6831	0.6340	0.6772	0.7225	0.7904	0.7364	0.8009	0.6800	0.7379
HD-LSTM	**0.6404**	**0.7123**	**0.6744**	**0.7511**	**0.7520**	**0.8146**	**0.7499**	**0.8153**	**0.7042**	**0.7733**

Results——Experimental Results on COMMUNITY-BASED QA



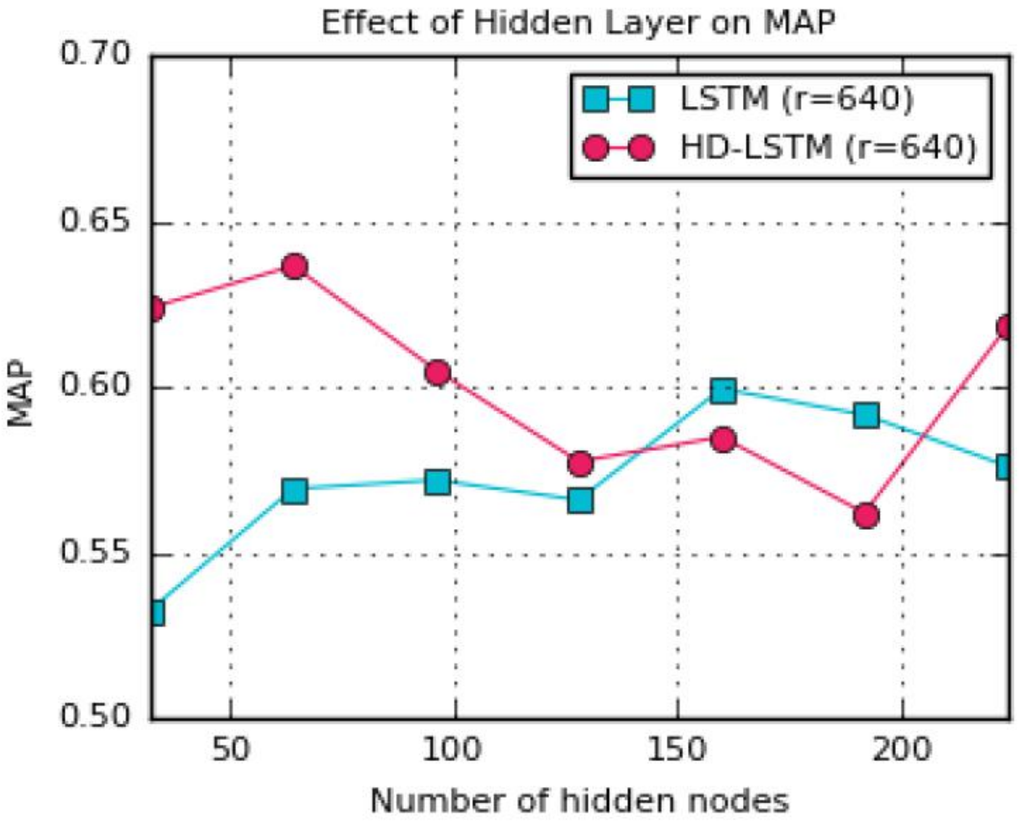
Model	P@1	MRR
Random Guess	0.2000	0.4570
Okapi BM-25	0.2250	0.4927
CNN	0.4125	0.6323
CNTN	0.4654	0.6687
LSTM	0.4875	0.6829
NTN-LSTM	0.5448	0.7309
HD-LSTM	**0.5569**	**0.7347**

Precision@1: $\frac{1}{N}\sum_{i=1}^{N}\delta(r(A^{+})=1)$

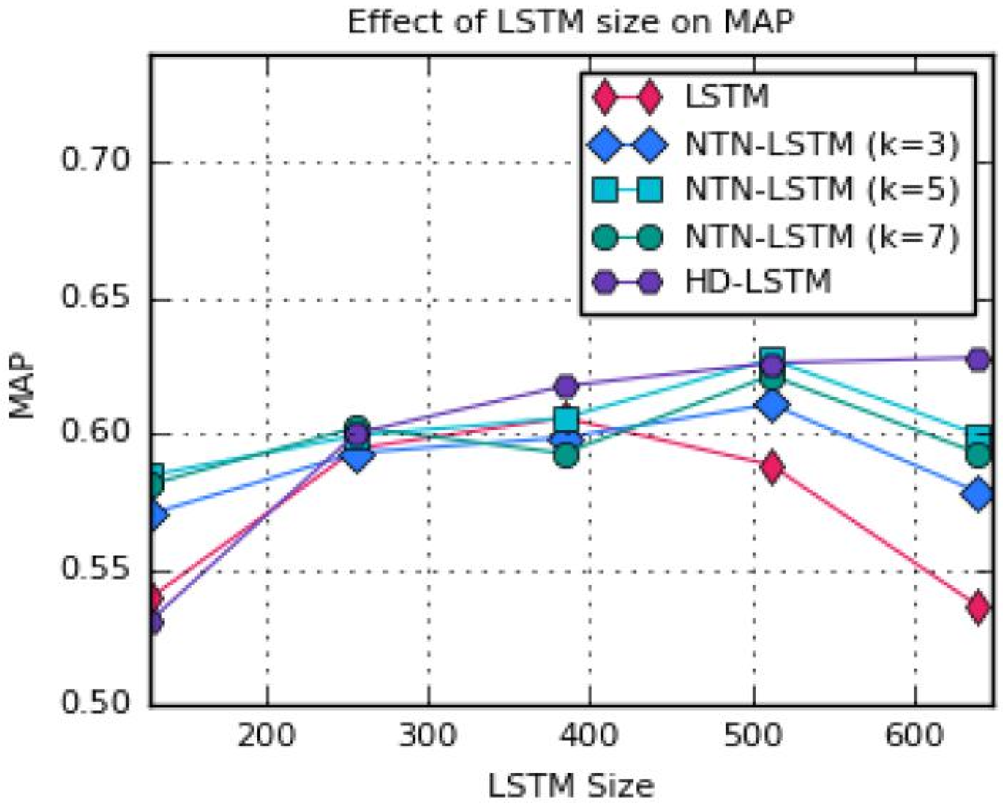
δ is the indicator function

A^{+} is the ground truth

Results——Effect of Hidden Layer Size



Results——Effect of Embedding Dimension



Background——Motivated

The prime contributions of the paper

~~1.We adopt holographic composition for modeling the interaction between representations of QA pairs~~

~~2.We present a novel deep learning architecture,HD-LSTM (Holographic Dual LSTM) for learning to rank QA pairs~~

~~3.We provide extensive experimental evidence of the effectiveness of our model on both factoid question answering and community-based question answering.~~

Conclusion

from the paper

1.丰富的实验

分别从两个数据集Trec QA与Yahoo CQA分析模型的通用性

每个数据集做了大量模型对比：
Trec QA 上7个模型对比、Yahoo CQA上5个模型对比且均分析不同点

能够从不同角度，发现并证明模型的闪光点

Conclusion

from the paper

2.清晰的表述能力

Conclusion

from the paper

3.开阔的学术视野