Attention-over-Attention Neural Networks for Reading Comprehension

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu*

来源：ACL2017

汇报人：徐军

日　期：2017/10/20

目录

- ✓ 问题定义与数据集介绍
- ✓ Attention基础介绍
- ✓ 阅读理解主流模型
- ✓ 改进的模型
- ✓ 实验
- ✓ 总结

问题定义与数据集介绍

完型填空任务可以描述为一个三元组:

$$< \boldsymbol{D}, \boldsymbol{Q}, \boldsymbol{A} >$$

这里D是指原文Document，Q是指问题Query，A是Answer，即问题的答案。这个答案是来自一个固定大小的词汇表A中的一个词或者一个短语。

我们要解决的问题就变成了：给定一个Document-Query对(D,Q)，从A中找到最合适的答案answer。

这个任务中，答案通常是文档中的一个单词，例如命名实体Obama，普通名词sunny，动词，介词。

主流数据集-CNN/Daily Mail

从新闻网站CNN和Daily Mail中获取数据源，用自动摘要的方法生成每篇新闻的摘要，用新闻原文作为Document，将摘要中去掉一个entity作为Query，被去掉的entity作为Answer，从而得到阅读理解的数据三元组$< \boldsymbol{D}, \boldsymbol{Q}, \boldsymbol{A} >$

这里存在一个问题，就是有的query并不需要联系到document，通过query中的上下文就可以预测出answer是什么，也就失去了阅读理解的意义。

举个例子，蓝天白__。因此，论文中提出了用一些标识替换entity和重新排列的方法将数据打乱，防止上面现象的出现。

主流数据集-CNN/Daily Mail

Original Version	Anonymised Version
Context	
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ...	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " ...
Query	
Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	Producer **X** will not press charges against *ent212*, his lawyer says.
Answer	
Oisin Tymon	*ent193*

主流数据集-Children’s Book Test(CBT)

在儿童读物中，故事叙述结构的清晰，上下文的作用更加突出。每篇文章只选用21句话，前20句作为Document，将第21句中去掉一个词之后作为Query，被去掉的词作为Answer，并且给定10个候选答案。

每个候选答案是从原文中随机选取的，并且这10个答案的词性是相同的，要是名词都是名词，要是命名实体都是实体，要是动词都是动词。

作者通过实验发现，动词和介词与上下文关联不大，可以使用常识来进行判断，所以大部分的研究重点在于命名实体和普通名词

主流数据集-Children’s Book Test(CBT)

"Well, Miss Maxwell, I think it only fair to tell you that you may have trouble with those boys when they do come. Forewarned is forearmed, you know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and it is hard to corner him."

"Are the boys big ?" queried Esther anxiously.

"Yes. Thirteen and fourteen and big for their age. You can't whip 'em -- that is the trouble. A man might, but they'd twist you around their fingers. You'll have your hands full, I'm afraid. But maybe they'll behave all right after all."

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that Mr. Baxter had exaggerated matters a little.

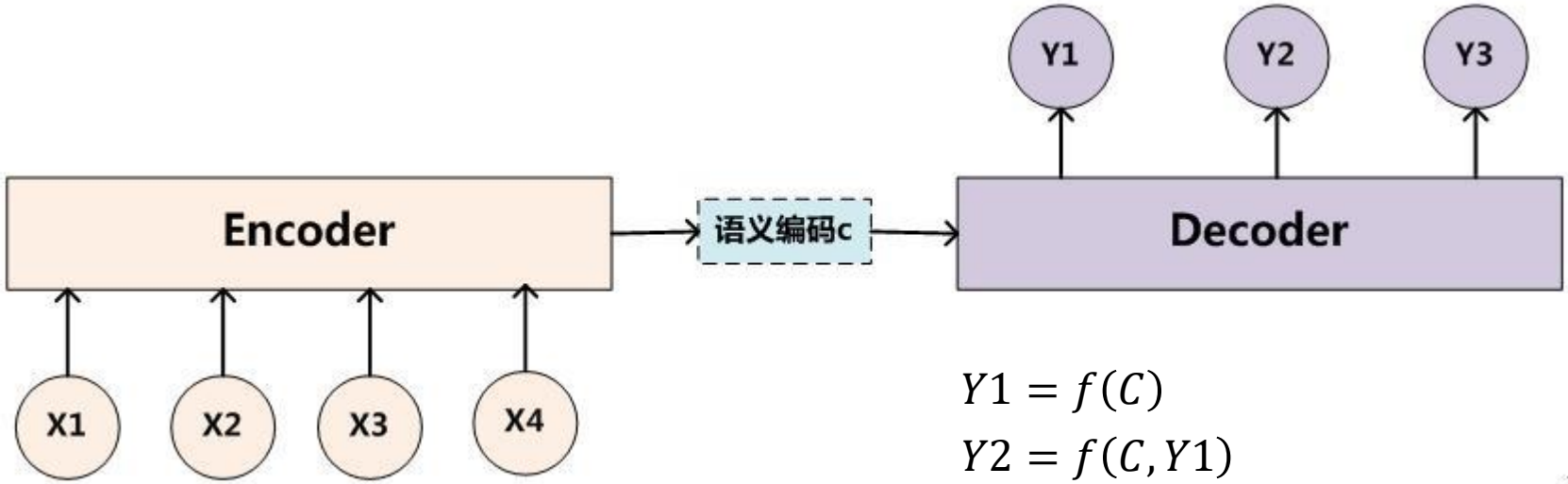
S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''
8 queried Esther anxiously .
9 `` Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly about her school and her work , hoped she was getting on well , and said he had two young rascals of his own to send soon .
20 Esther felt relieved .

q: She thought that Mr. _____ had exaggerated matters a little .

C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

a: Baxter

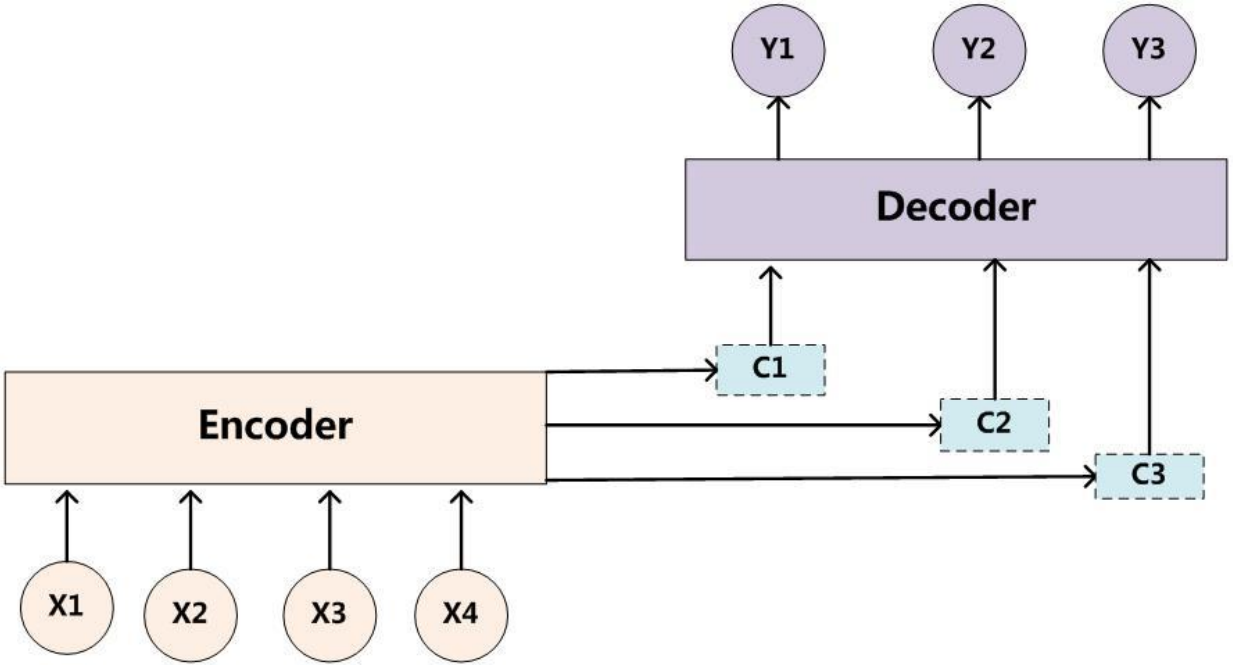
分心模型



$$Y1 = f(C)$$
$$Y2 = f(C, Y1)$$
$$Y3 = f(C, Y1, Y2)$$

在Encoder-Decoder模型下，首先将原文语义进行编码，得到中间结果C。解码时，对每一个Y来说，他们所利用的原文语义信息相同。存在两个问题，原文太长，语义会损失。原文中每个词对Y的影响力相同。

Attention模型



$$Y1 = f(C1)$$
$$Y2 = f(C2, Y1)$$
$$Y3 = f(C3, Y1, Y2)$$

Attention模型

举个例子，假设原文是“Tom chase Jerry”翻译之后是“汤姆追逐杰瑞”，计算输出时注意力的变化。

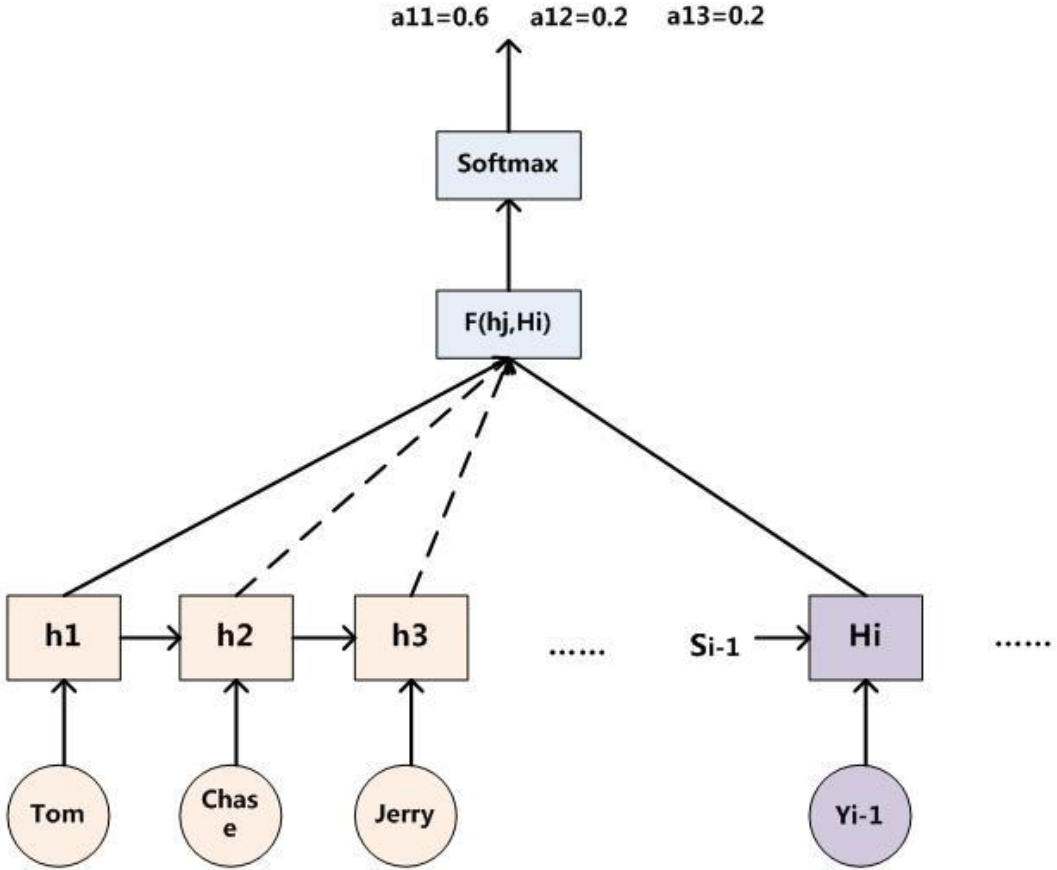
例如把“Tom”翻译“汤姆”时，Tom的注意力为0.6，chase的注意力为0.2，Jerry的注意力为0.2，这样就形成了一个基本的注意力模型。

$$C_{汤姆} = g(0.6 * f_2(" Tom "), 0.2 * f_2(" chase "), 0.2 * f_2(" Jerry "))$$

$$C_{追逐} = g(0.2 * f_2(" Tom "), 0.7 * f_2(" chase "), 0.1 * f_2(" Jerry "))$$

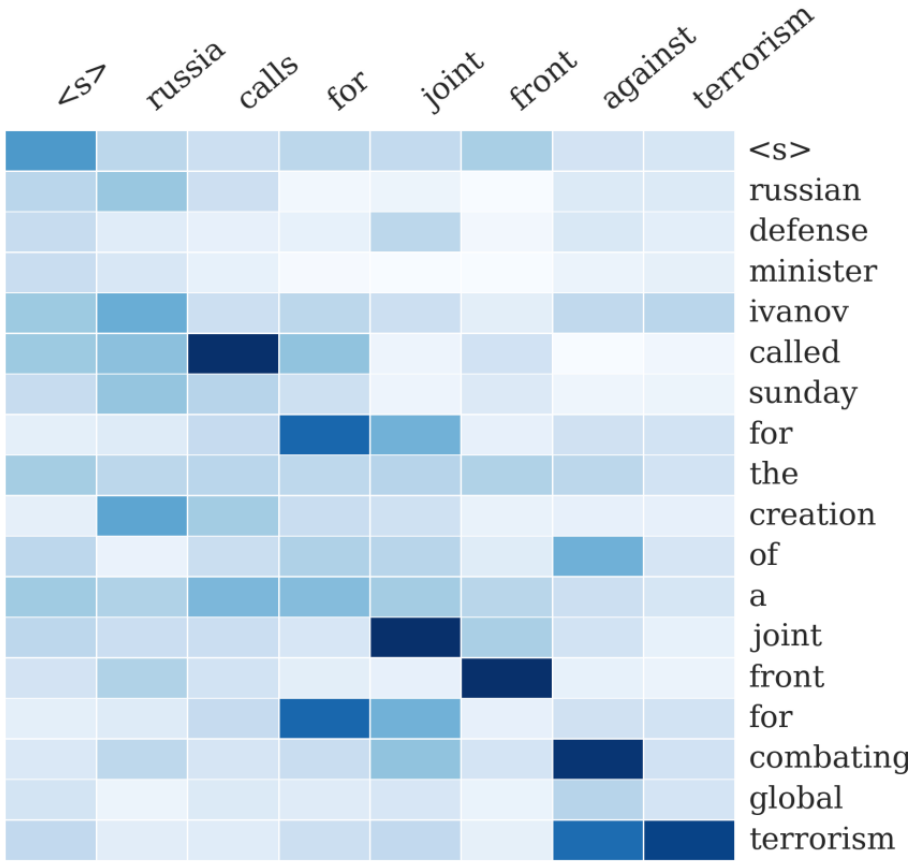
$$C_{杰瑞} = g(0.3 * f_2(" Tom "), 0.2 * f_2(" chase "), 0.5 * f_2(" Jerry "))$$

Attention模型



怎么计算每个单词的在原文中的注意力分布？

Attention模型

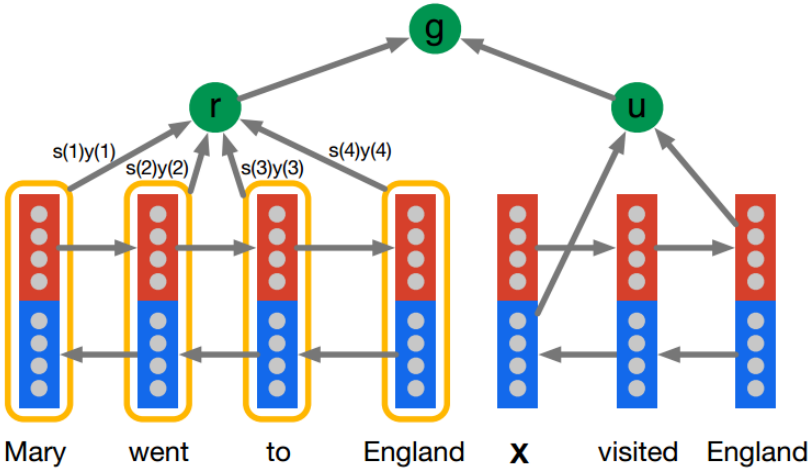


一个最直观的例子，横坐标表示摘要，纵坐标表示原文。

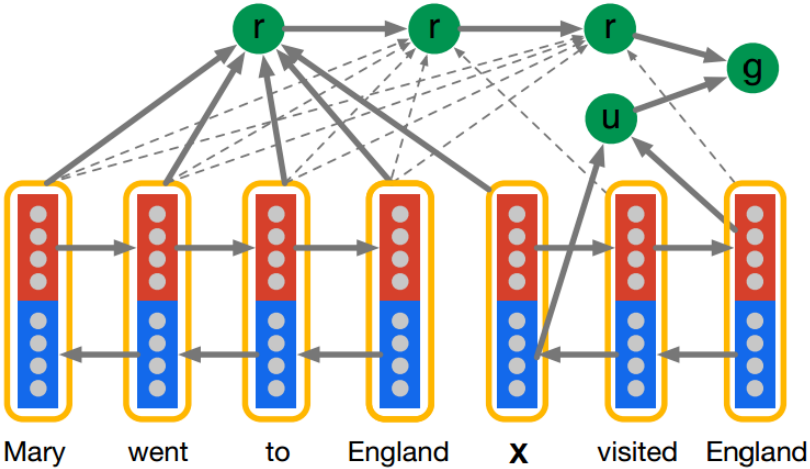
矩阵中每一列代表生成的目标单词对应输入句子每个单词的AM分配概率，颜色越深代表分配到的概率越大。

阅读理解主流模型

在（Hermann et al.,2015NIPS），CNN/CBT的作者对这两个数据集设计了两个模型



(a) Attentive Reader.



(b) Impatient Reader.

阅读理解主流模型

Attentive Reader

基本的实现过程是，左边表示对document的双向LSTM编码，右边表示对query的双向LSTM编码，对于document这一部分，每个单词的两次编码直接拼接，得到隐含状态 ，再乘以各自的权重，这里的权重是通过神经网络学习得到的，再相加得到 ，对于query这一部分，只将正向的LSTM层的最后编码和反向的LSTM的最后一层编码拼接，得到隐含状态 。通过神经网络的训练，计算document中每一个attention词与query的关联度，这篇文章中作者将query的特征 与每一个带attention词向量相加，最后根据 计算概率值。

Impatient Reader

这一模型和 attentive reader 类似，但是每读入一个 query token 就迭代计算一次document 的权重分布

阅读理解主流模型

	CNN		Daily Mail	
	valid	test	valid	test
Maximum frequency	26.3	27.9	22.5	22.7
Exclusive frequency	30.8	32.6	27.3	27.7
Frame-semantic model	32.2	33.0	30.7	31.1
Word distance model	46.2	46.9	55.6	54.8
Deep LSTM Reader	49.0	49.9	57.1	57.3
Uniform attention	31.1	33.6	31.0	31.7
Attentive Reader	56.5	58.9	64.5	63.7
Impatient Reader	**57.0**	**60.6**	**64.8**	**63.9**

Table 5: Results for all the models and benchmarks on the CNN and Daily Mail datasets. The Uniform attention baseline sets all of the $m(t)$ parameters to be equal.

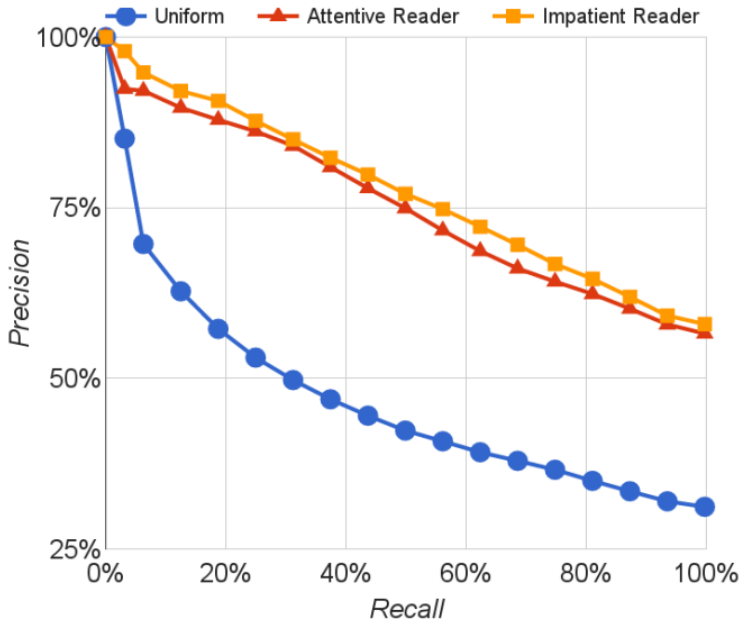
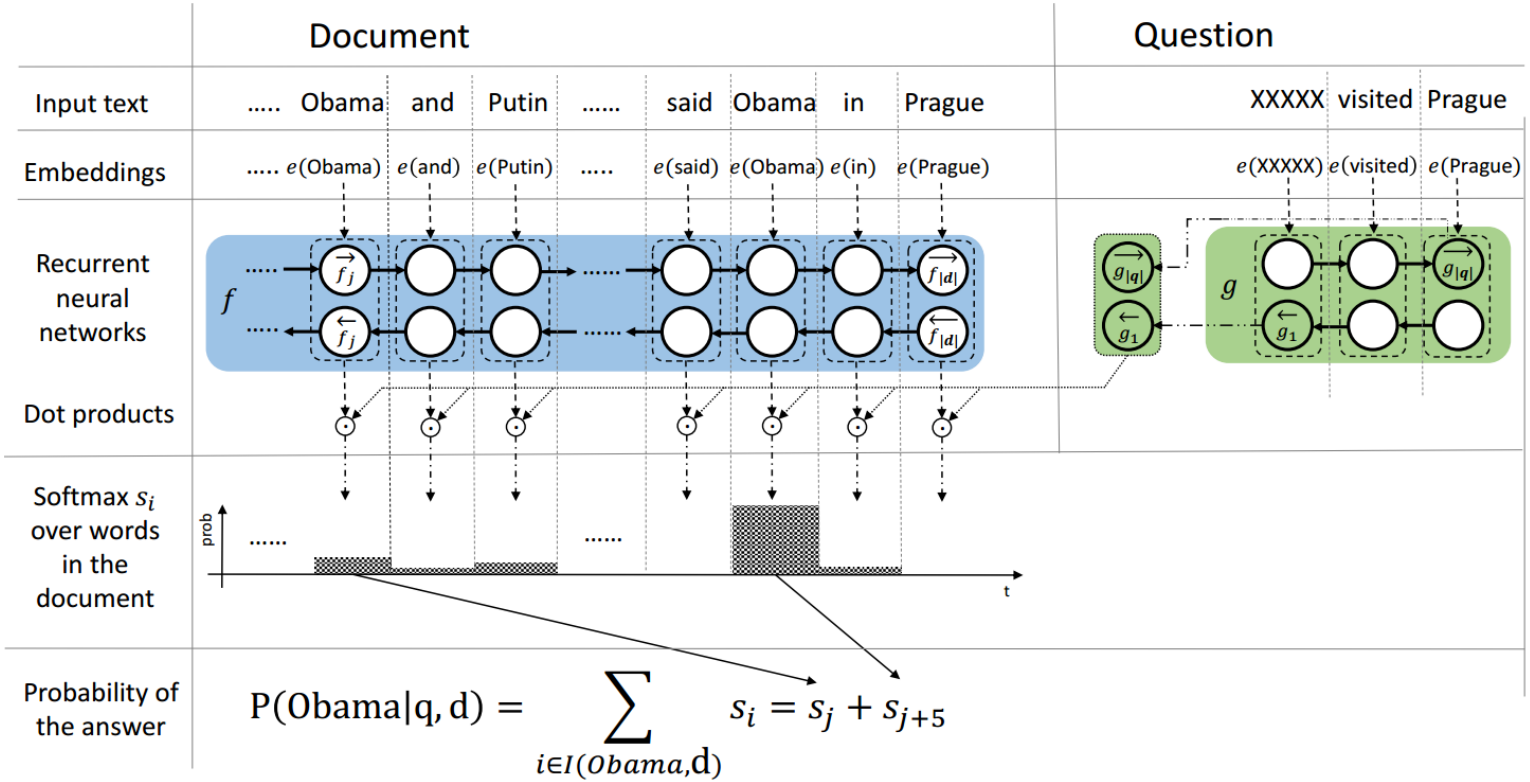


Figure 2: Precision@Recall for the attention models on the CNN validation data.

阅读理解主流模型

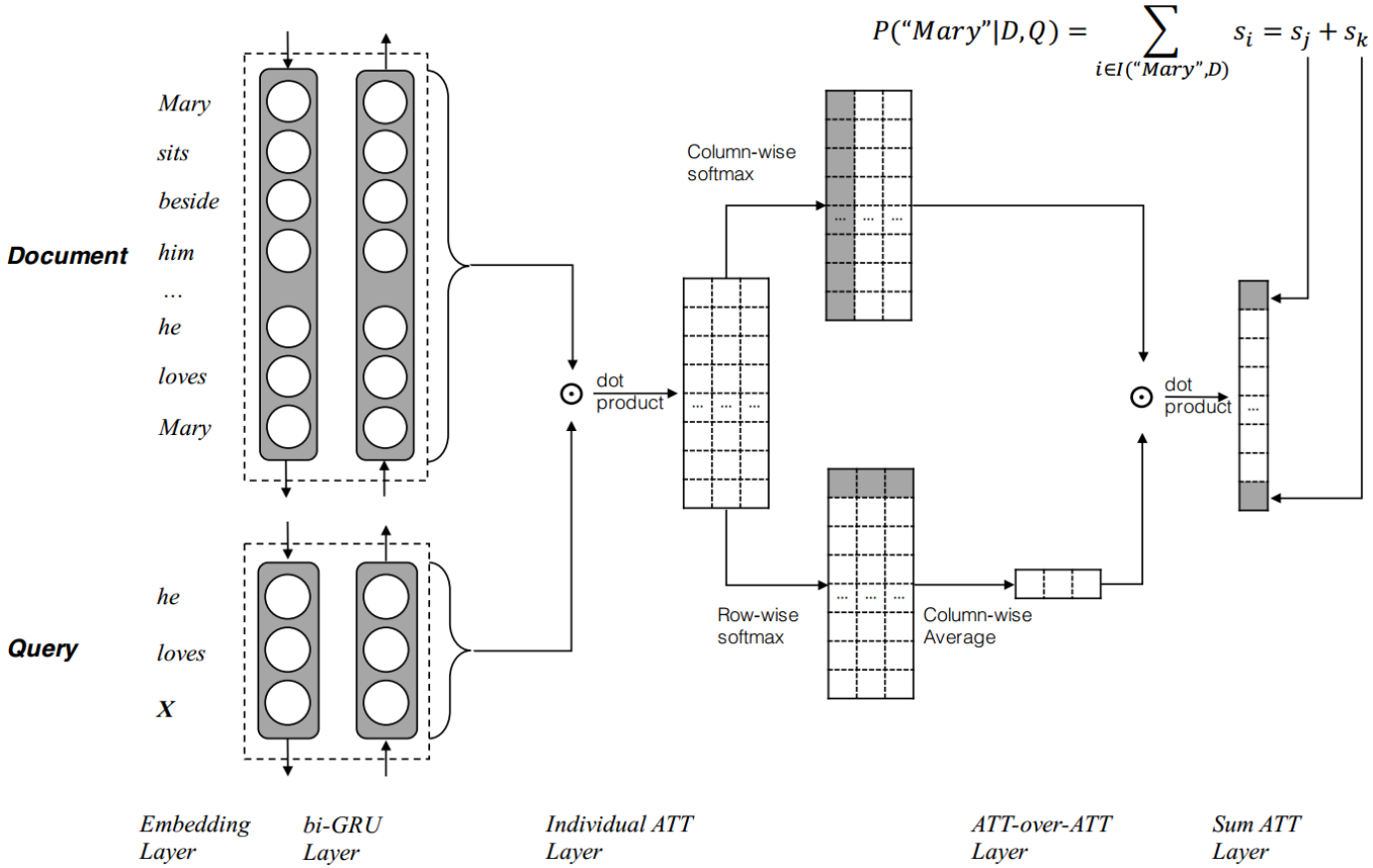
针对这个问题(Kadlec et al.,2016ACL)，设计了如下的模型



阅读理解主流模型

	CNN		Daily Mail	
	valid	test	valid	test
Attentive Reader †	61.6	63.0	70.5	69.0
Impatient Reader †	61.8	63.8	69.0	68.0
MemNNs (single model) ‡	63.4	66.8	NA	NA
MemNNs (ensemble) ‡	66.2	69.4	NA	NA
Dynamic Entity Repres. (max-pool) ♯	71.2	70.7	NA	NA
Dynamic Entity Repres. (max-pool + byway)♯	70.8	72.0	NA	NA
Dynamic Entity Repres. + w2v ♯	71.3	72.9	NA	NA
Chen et al. (2016) (single model)	72.4	72.4	76.9	75.8
AS Reader (single model)	68.6	69.5	75.0	73.9
AS Reader (avg for top 20%)	68.4	69.9	74.5	73.5
AS Reader (avg ensemble)	73.9	**75.4**	78.1	77.1
AS Reader (greedy ensemble)	74.5	74.8	78.7	**77.7**

改进的模型


$P(\text{“}Mary\text{”}|D,Q) = \sum_{i \in I(\text{“}Mary\text{”},D)} s_i = s_j + s_k$
Document
Mary
sits
beside
him
…
he
loves
Mary
Query
he
loves
X
dot product
Column-wise softmax
Row-wise softmax
Column-wise Average
dot product
Embedding Layer
bi-GRU Layer
Individual ATT Layer
ATT-over-ATT Layer
Sum ATT Layer

改进的模型

1、Contextual Embedding

首先将document和query中的每个词提取出来，通过查询训练好的word embedding 得到每个词的特征表示。使用双向GRU模型来获取每个部分的语义表示。作者这里对每个单词的编码是384维，GRU输出层为256维。所以单个GRU的输入就是384维，输出为256维。双层GRU拼接之后的输出就是512维。

2、Pair-wise Matching Score

计算 $\boldsymbol{h_{doc}}$ 中每个单词与 $\boldsymbol{h_{query}}$单词的匹配分数，得到分数矩阵。

$$\boldsymbol{M(i,j) = h_{doc}(i) \cdot h_{query}(j)^T}$$

最后得到的分数矩阵 $\boldsymbol{M \in R^{|D| \times |Q|}}$，上面的例子中document的维度是 7*512，query的维度是 3*512，进行点乘之后变成了 7*3，列表示document中的单词，行表示query中的单词，ASReader模型中，将query部分编码成 1*512，最后得到的也是的矩阵 7*1。

3、Individual Attentions (列归一化)

计算query中每个词在document中注意力分布，使用softmax进行归一化，直观的解释就是document中每个词在query中的重要性，得到的矩阵的形状没有改变，只是数值进行了归一化。计算方法为：

$$\alpha(t) = softmax(M(1,t), \ldots, M(|D|,t))$$
$$\alpha = [\alpha(1), \alpha(2), \ldots, \alpha(|Q|)]$$

4、Attention-over-Attention (行归一化)

计算document中每个单词在query中注意力的权值分布，使用softmax归一化。

直观的解释就是query中每个词在document中的重要性，然后在对每一列求平均，获取query中每个词在document中的权重，最直观的印象就是每个词在query中的权重。然后计算两个方向 attention 的点积，得到document中每个词在query中的重要性。

改进的模型

N-best Re-ranking Strategy

- N-best Decoding

相比于前面的选择一个概率最大的作为最后的答案，这里选择概率最大的N个单词作为候选词。

- Refill Candidate into Query

把每个候选词填入Query中，检测他在里面的语义，形成N个句子。

改进的模型

N-best Re-ranking Strategy

- Feature Scoring(对上面的N个句子进行评分，作者主要选择了三个评分特征)

  - Global N-gram LM: 计算候选句子的流畅性。例如需要计算“蓝天白云”和“蓝天白雾”的流畅性，通过训练所以所有的文本可以得到已知“蓝天白”后面出现每个词的概率，这个概率就表示其流畅度。

  - Local N-gram LM: 还是计算候选句子的流畅性。但是这次求概率时，不是计算所有的训练文本，而是从每个query对应的document来计算。

  - Word-class LM: 这个方法先将文档中所有单词，通过聚类的方法分为1000类，然后将候选句子中按照类别来计算流畅度，计算方法和Global N-gram LM相同。

改进的模型

N-best Re-ranking Strategy

- Weight Tuning

通过训练数据不断的调整这三个特征的权重，使得目标损失最小。

- Re-scoring and Re-ranking

计算每个句子在这三个特征下的加权评分，最后再次通过softmax选择概率最大的。

实验

	CNN News		CBTest NE		CBTest CN	
	Valid	Test	Valid	Test	Valid	Test
Deep LSTM Reader (Hermann et al., 2015)	55.0	57.0	-	-	-	-
Attentive Reader (Hermann et al., 2015)	61.6	63.0	-	-	-	-
Human (context+query) (Hill et al., 2015)	-	-	-	*81.6*	-	*81.6*
MemNN (window + self-sup.) (Hill et al., 2015)	63.4	66.8	70.4	66.6	64.2	63.0
AS Reader (Kadlec et al., 2016)	68.6	69.5	73.8	68.6	68.8	63.4
CAS Reader (Cui et al., 2016)	68.2	70.0	74.2	69.2	68.2	65.7
Stanford AR (Chen et al., 2016)	72.4	72.4	-	-	-	-
GA Reader (Dhingra et al., 2016)	73.0	73.8	74.9	69.0	69.0	63.9
Iterative Attention (Sordoni et al., 2016)	72.6	73.3	75.2	68.6	72.1	69.2
EpiReader (Trischler et al., 2016)	**73.4**	74.0	75.3	69.7	71.5	67.4
AoA Reader	73.1	**74.4**	**77.8**	**72.0**	**72.2**	**69.4**
AoA Reader + Reranking	-	-	**79.6**	**74.0**	**75.7**	**73.1**
MemNN (Ensemble)	66.2	69.4	-	-	-	-
AS Reader (Ensemble)	73.9	75.4	74.5	70.6	71.1	68.9
GA Reader (Ensemble)	76.4	77.4	75.5	71.9	72.1	69.4
EpiReader (Ensemble)	-	-	76.6	71.8	73.6	70.6
Iterative Attention (Ensemble)	74.5	75.7	76.9	72.0	74.1	**71.0**
AoA Reader (Ensemble)	-	-	**78.9**	**74.5**	**74.7**	70.8
AoA Reader (Ensemble + Reranking)	-	-	**80.3**	**75.6**	**77.0**	**74.1**

实验

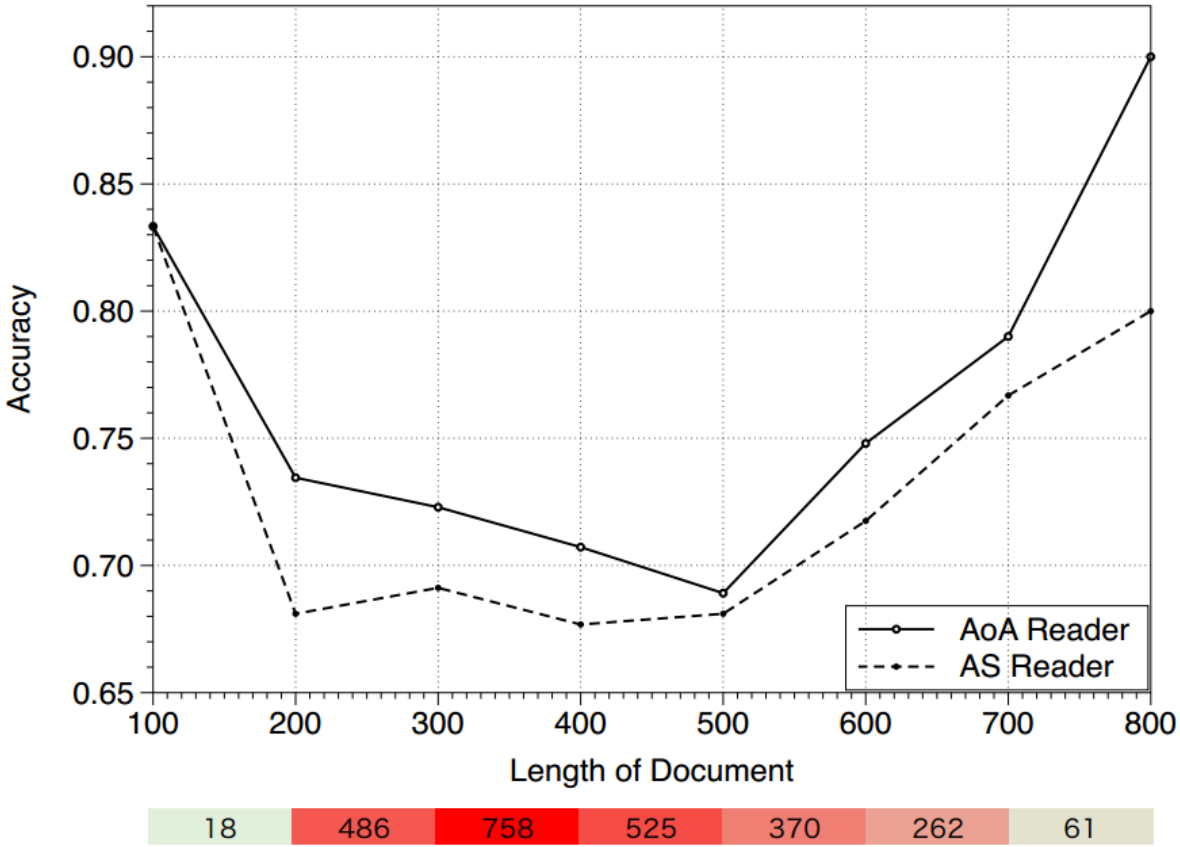
从上面的实验结果可以看出，Re-ranking对整个模型的效果有很多贡献。具体的贡献量如下：

	CBTest NE		CBTest CN	
	Valid	Test	Valid	Test
AoA Reader	77.8	72.0	72.2	69.4
+Global LM	78.3	72.6	73.9	71.2
+Local LM	79.4	73.8	74.7	71.7
+Word-class LM	79.6	74.0	75.7	73.1

在CBT NE测试集上Global LM增加了0.6，Local LM增加了0.12，Word-class LM增加了0.2。
在CBT CN测试集上Global LM增加了1.8，Local LM增加了0.5，Word-class LM增加了1.4。

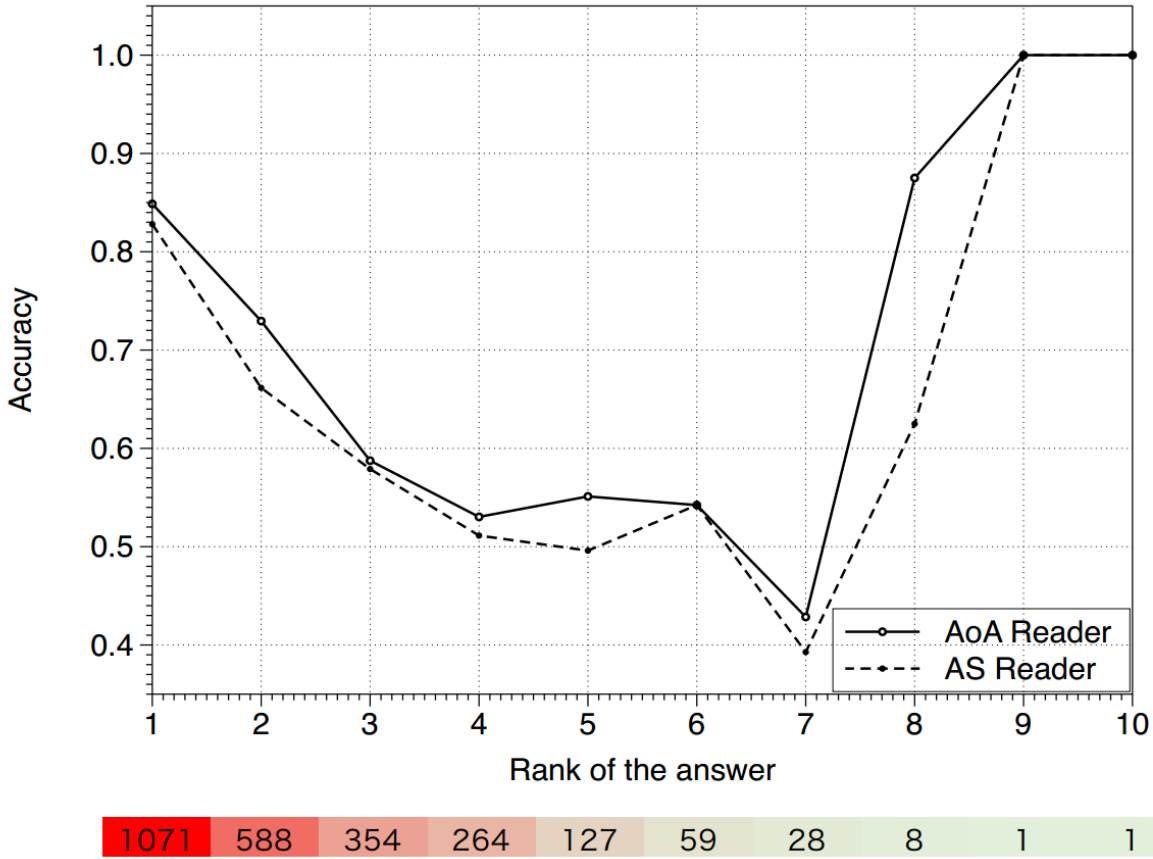
在NE这个数据集上，Local LM这个特征最为重要，回答问题的命名实体通常在对应的文章上。
在CN这个数据集上，Global LM和Word-class LM特征最为重要。

实验



从上最下面的横条表示每个区间的文档的数量，可以看出文档单词个数在400-500之间，整个模型的效果最差。这个效果和AS Reader模型类似，当单词个数超过700时，AoA模型精度提升非常大，可以看出AoA模型比AS模型更适合阅读长文本。

实验



1071	588	354	264	127	59	28	8	1	1

最下面的横条表示有多少个问句选择了答案所在的词频。例如有1071个句子的答案选择了词频最高的作为答案，测试的准确率为0.85左右。

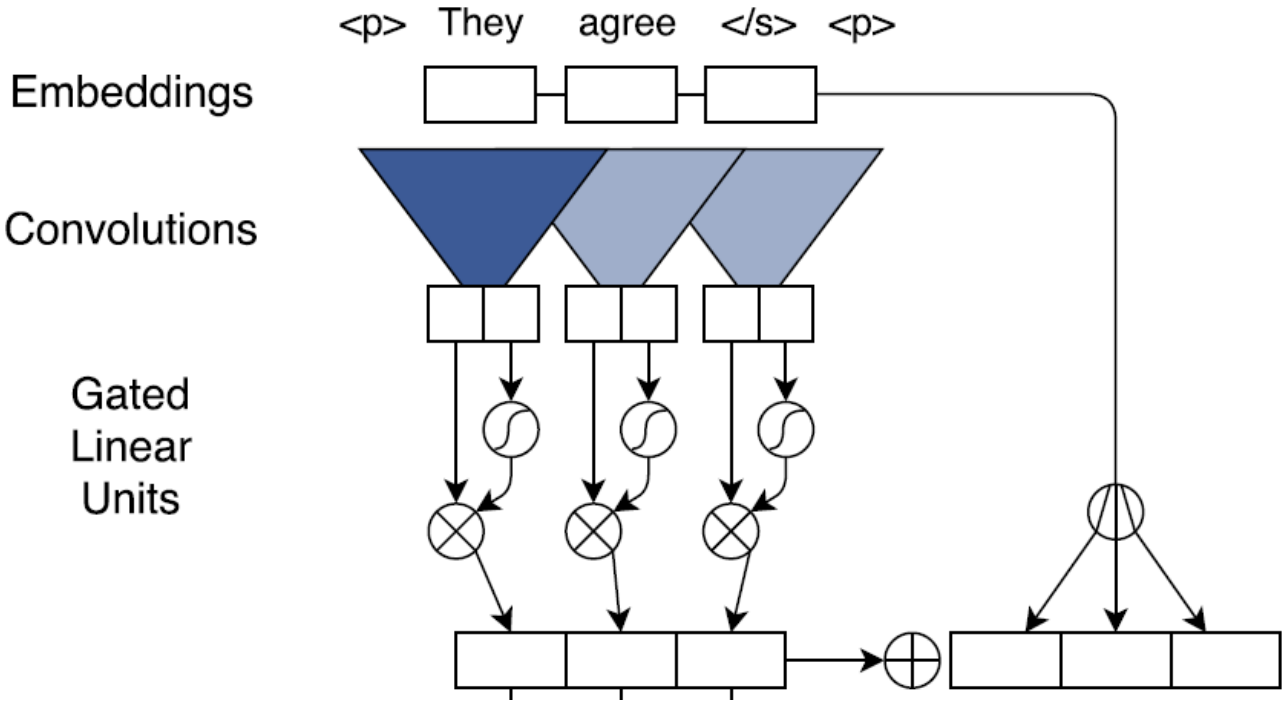
总结

创新点:

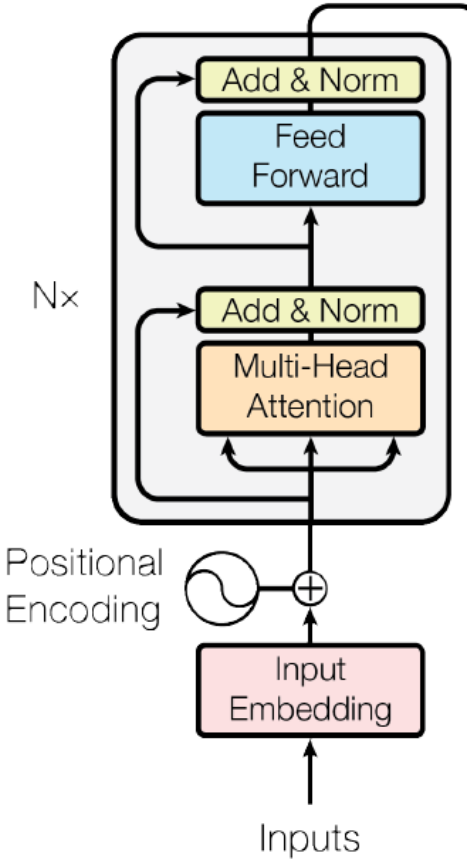
- 在这个任务上第一次使用 attention-over-attention。
- 模型思想简单便于理解，性能更好。
- 提出了N-best re-ranking 策略对候选集重新打分，提升了整个模型的性能

自己的想法

可以将RNN部分替换成CNN



自己的想法



全部使用Attention，不再使用神经网络