论文阅读笔记

Step5

MF1833063, 史鹏, spwannasing@gmail.com

2019 年 7 月 4 日

1 ERNIE:Enhanced Language Representation with Informative Entities

此文章是对bert的一次扩展，提出了知识图谱中的多信息实体（informative entity）可以作为外部知识改善语言表征。

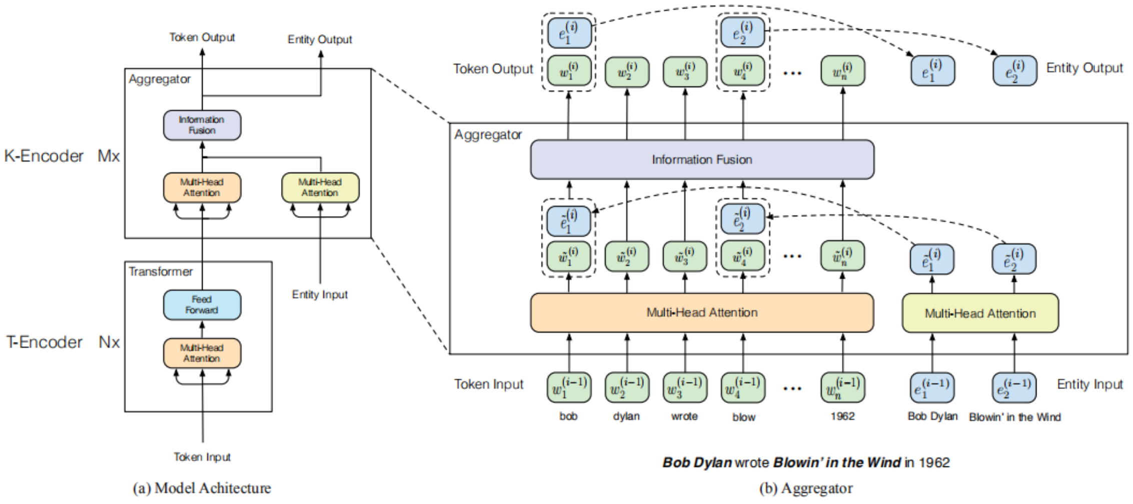
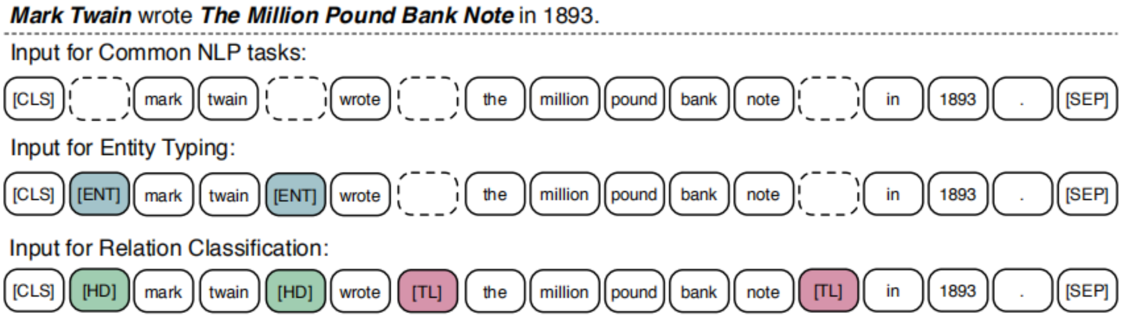


图 1: overview



Mark Twain wrote ***The Million Pound Bank Note*** in 1893.

Input for Common NLP tasks:

[CLS] mark twain wrote the million pound bank note in 1893 . [SEP]

Input for Entity Typing:

[CLS] [ENT] mark twain [ENT] wrote the million pound bank note in 1893 . [SEP]

Input for Relation Classification:

[CLS] [HD] mark twain [HD] wrote [TL] the million pound bank note [TL] in 1893 . [SEP]

图 2: finetune

Knowledgeable Encoder：

$$\left\{\tilde{\boldsymbol{w}}_1^{(i)},\ldots,\tilde{\boldsymbol{w}}_n^{(i)}\right\}=\mathrm{MH-ATT}\left(\left\{\boldsymbol{w}_1^{(i-1)},\ldots,\boldsymbol{w}_n^{(i-1)}\right\}\right) \tag{1.1}$$

$$\left\{\tilde{e}_1^{(i)},\ldots,\tilde{e}_m^{(i)}\right\}=\mathrm{MH-ATT}\left(\left\{e_1^{(i-1)},\ldots,e_m^{(i-1)}\right\}\right) \tag{1.2}$$

对于和entity对齐的token：

$$\begin{aligned}\boldsymbol{h}_j&=\sigma\left(\tilde{\boldsymbol{W}}_t^{(i)}\tilde{\boldsymbol{w}}_j^{(i)}+\tilde{\boldsymbol{W}}_e^{(i)}\tilde{\boldsymbol{e}}_k^{(i)}+\tilde{\boldsymbol{b}}^{(i)}\right)\\ \boldsymbol{w}_j^{(i)}&=\sigma\left(\boldsymbol{W}_t^{(i)}\boldsymbol{h}_j+\boldsymbol{b}_t^{(i)}\right)\\ \boldsymbol{e}_k^{(i)}&=\sigma\left(\boldsymbol{W}_e^{(i)}\boldsymbol{h}_j+\boldsymbol{b}_e^{(i)}\right)\end{aligned} \tag{1.3}$$

else：

$$
\begin{aligned}
\boldsymbol{h}_j &= \sigma\left(\tilde{\boldsymbol{W}}_t^{(i)} \tilde{\boldsymbol{w}}_j^{(i)} + \tilde{\boldsymbol{b}}^{(i)}\right) \\
\boldsymbol{w}_j^{(i)} &= \sigma\left(\boldsymbol{W}_t^{(i)} \boldsymbol{h}_j + \boldsymbol{b}_t^{(i)}\right)
\end{aligned} \tag{1.4}
$$

对于引入的信息的pre-training目标：

$$
p\left(e_j \mid w_i\right) = \frac{\exp\left(\text{ linear }\left(\boldsymbol{w}_i^o\right) \cdot \boldsymbol{e}_j\right)}{\sum_{k=1}^m \exp\left(\text{ linear }\left(\boldsymbol{w}_i^o\right) \cdot \boldsymbol{e}_k\right)} \tag{1.5}
$$

2 ERNIE: Enhanced Representation through Knowledge Integration

ERNIE 通过建模海量数据中的词、实体及实体关系，学习真实世界的语义知识。相较于 BERT 学习原始语言信号，ERNIE 直接对先验语义知识单元进行建模，增强了模型语义表示能力。这里我们举个例子：

Learnt by BERT ：哈 [mask] 滨是 [mask] 龙江的省会，[mask] 际冰 [mask] 文化名城。

Learnt by ERNIE：[mask] [mask] [mask] 是黑龙江的省会，国际 [mask] [mask] 文化名城。

在 BERT 模型中，我们通过『哈』与『滨』的局部共现，即可判断出『尔』字，模型没有学习与『哈尔滨』相关的任何知识。而 ERNIE 通过学习词与实体的表达，使模型能够建模出『哈尔滨』与『黑龙江』的关系，学到『哈尔滨』是『黑龙江』的省会以及『哈尔滨』是个冰雪城市。

训练数据方面，除百科类、资讯类中文语料外，ERNIE 还引入了论坛对话类数据，利用 DLM（Dialogue Language Model）建模 Query-Response 对话结构，将对话 Pair 对作为输入，引入 Dialogue Embedding 标识对话的角色，利用 Dialogue Response Loss 学习对话的隐式关系，进一步提升模型的语义表示能力。

我们在自然语言推断，语义相似度，命名实体识别，情感分析，问答匹配 5 个公开的中文数据集合上进行了效果验证，ERNIE 模型相较 BERT 取得了更好的效果。

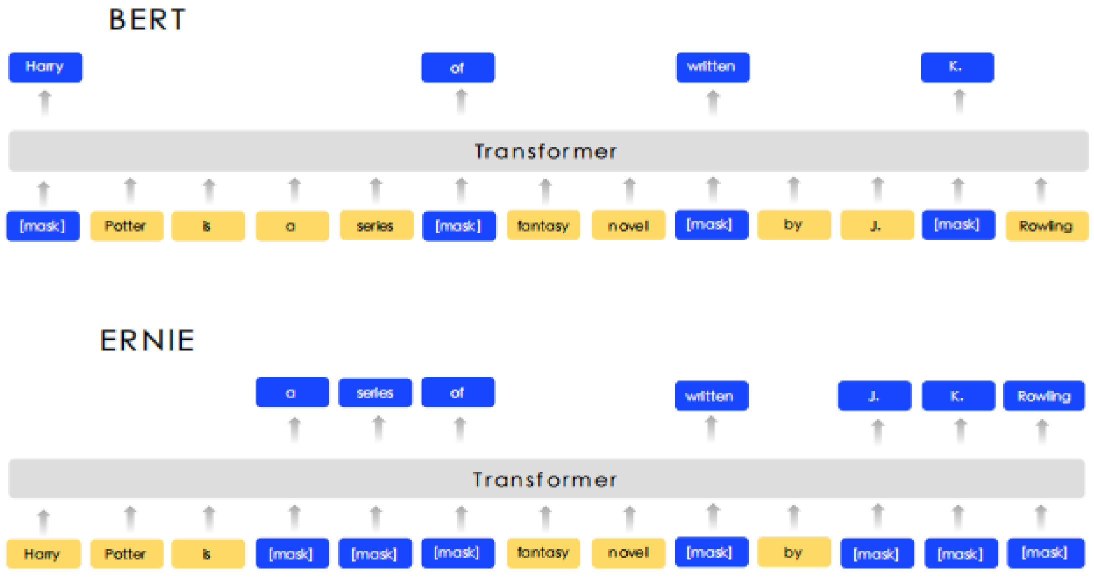
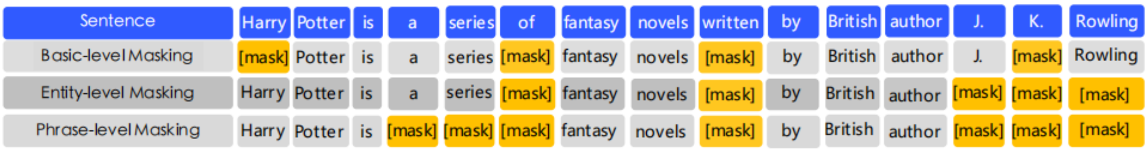


Figure 1: The different masking strategy between BERT and ERNIE



Sentence	Harry	Potter	is	a	series	of	fantasy	novels	written	by	British	author	J.	K.	Rowling
Basic-level Masking	[mask]	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	J.	[mask]	Rowling
Entity-level Masking	Harry	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]
Phrase-level Masking	Harry	Potter	is	[mask]	[mask]	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]

Figure 2: Different masking level of a sentence

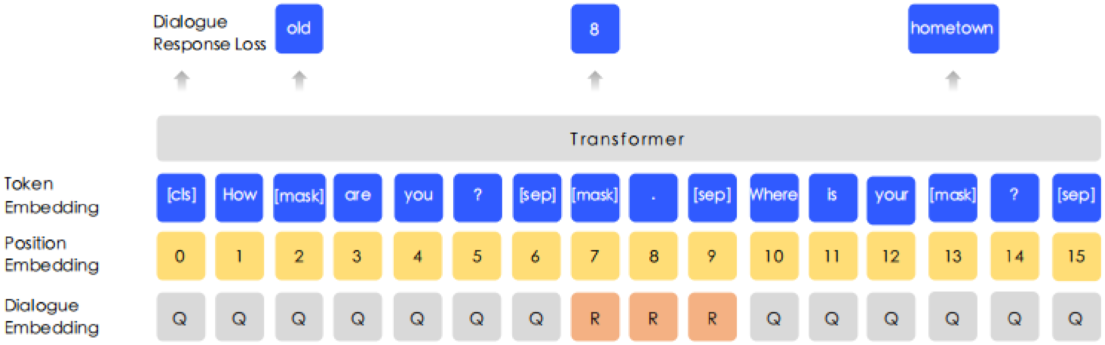
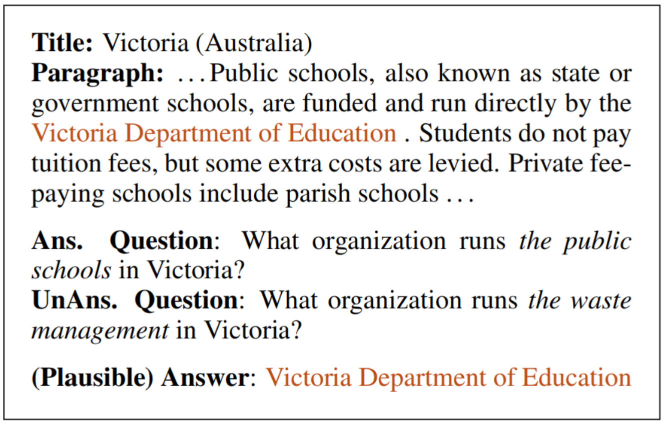


Figure 3: Dialogue Language Model. Source sentence: [cls] How [mask] are you [sep] 8 . [sep] Where is your [mask] ? [sep]. Target sentence (words the predict): old, 8, hometown)

3 Learning to Ask Unanswerable Questions for Machine Reading Comprehension

提出一种数据增强技术，根据与包含答案的相应段落配对的可回答问题自动生成相关的无法回答的问题。所提出的结构为“pair-to-sequence”。



Title: Victoria (Australia)
Paragraph: . . . Public schools, also known as state or government schools, are funded and run directly by the Victoria Department of Education . Students do not pay tuition fees, but some extra costs are levied. Private fee-paying schools include parish schools . . .

Ans. Question: What organization runs *the public schools* in Victoria?
UnAns. Question: What organization runs *the waste management* in Victoria?

(Plausible) Answer: Victoria Department of Education

Figure 1: An example taken from the SQuAD 2.0 dataset. The annotated (plausible) answer span in the paragraph is used as a pivot to align the pair of answerable and unanswerable questions.

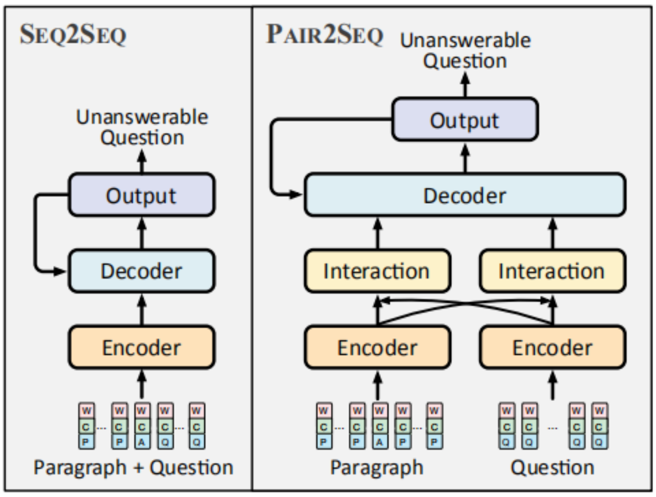


Figure 2: Diagram of the proposed pair-to-sequence model and sequence-to-sequence model. The input embeddings is the sum of the word embeddings, the character embeddings and the token type embeddings. The input questions are all answerable.

4 Course Concept Expansion in MOOCs with External Knowledge and Interactive Game

随着大规模在线开放课程(MOOC)的日益普及，为MOOC用户自动提供课外知识成为可能。语义漂移和知识缺乏在复杂的MOOC环境下，现有的方法不能有效地扩展课程概念。本文首先在通过外部知识库搜索新概念的过程中建立一个新的边界，然后利用异构特征来验证高质量的结果。具体算法没有特别仔细地看。

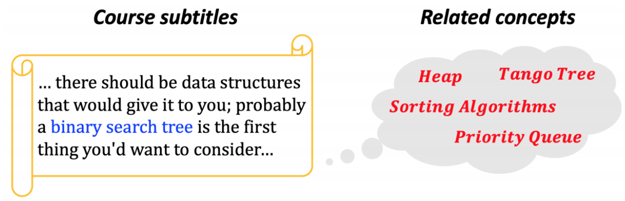

Course subtitles
Related concepts
... there should be data structures that would give it to you; probably a binary search tree is the first thing you'd want to consider...
Heap
Tango Tree
Sorting Algorithms
Priority Queue

Figure 1: An example of “out-of-teaching” concepts in the course “*Data Structure and Algorithm*”.

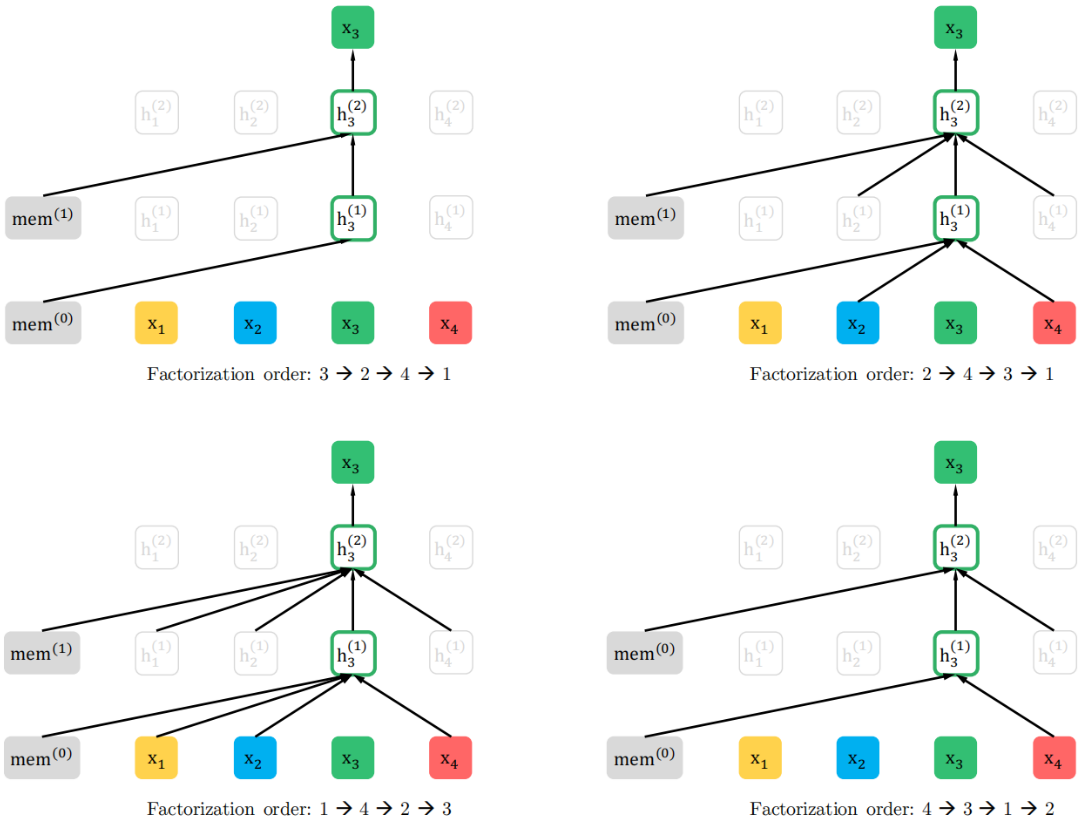
5 XLNet: Generalized Autoregressive Pretraining for Language Understanding

本文提出了一个新的预训练模型，在20个任务上刷新了Bert的记录。其相对bert的主要改进是结合了AR(Transformer-XL)和AE的优点，抛弃了bert的Masked language model，使用了Permutation language model。对于每一种排列的序列，并不是改变原始的序列（即position embedding不变），只是改变mask（bert里的mask是直接mask一个word，这里的是在Transformer的计算过程中“mask”住不需要的部分）。 e.g 在图二中，对于序列3241，1能“看见”324，所以都不用mask掉，而对于2，只能看见3，所以要mask掉14。

虽然排列语言模型能满足目前的目标，但是对于普通的transformer结构来说是存在一定的问题的，为什么这么说呢，看个例子，假设我们要求这样的一个对数似然，$p_\theta\left(X_{z_t}|x_{z_{<t}}\right)$如果采用标准的softmax的话，那么

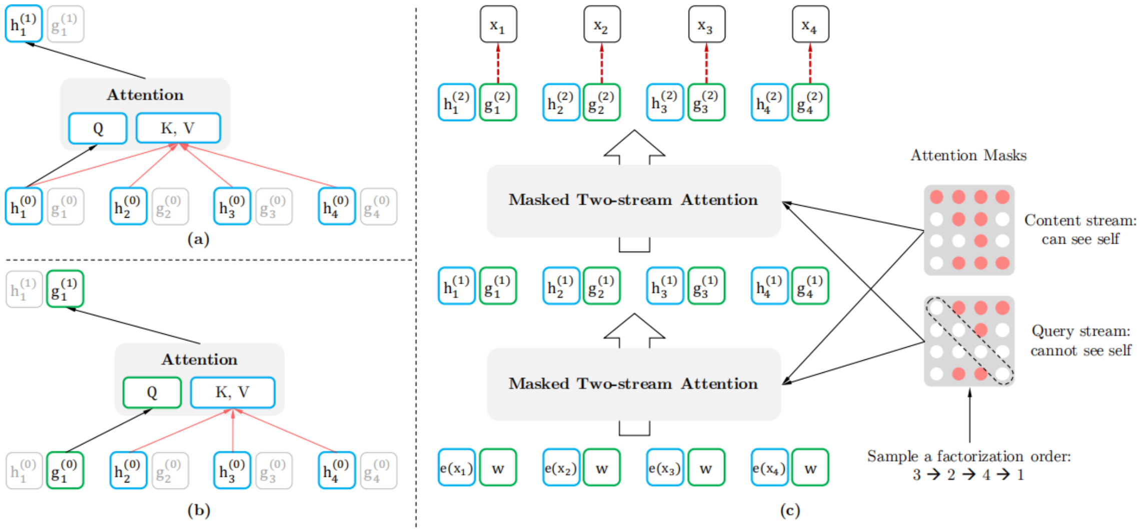
$$p_\theta\left(X_{z_t}|x_{z_{<t}}\right) = \frac{\exp\left(e(x)^T h_\theta\left(x_{z_z}\right)\right)}{\sum_{x'} \exp\left(e\left(x'\right)^T h_\theta\left(x_{z_{<t}}\right)\right)}$$

其中$h_\theta\left(x_{z_{<t}}\right)$表示的是添加了mask后的transformer的输出值，可以发现$h_\theta\left(x_{z_{<t}}\right)$并不依赖于其要预测的内容的位置信息，因为无论预测目标的位置在哪里，因式分解后得到的所有情况都是一样的，并且transformer的权重对于不同的情况是一样的，因此无论目标位置怎么变都能得到相同的分布结果，如下图所示，假如我们的序列index表示为[1,2,3]，对于目标2与3来说，其因式分解后的结果是一样的，那么经过transformer之后得到的结果肯定也是一样的。


x₃
h₁⁽²⁾ h₂⁽²⁾ h₃⁽²⁾ h₄⁽²⁾
mem⁽¹⁾ h₁⁽¹⁾ h₂⁽¹⁾ h₃⁽¹⁾ h₄⁽¹⁾
mem⁽⁰⁾ x₁ x₂ x₃ x₄
Factorization order: 3 → 2 → 4 → 1
Factorization order: 2 → 4 → 3 → 1
Factorization order: 1 → 4 → 2 → 3
Factorization order: 4 → 3 → 1 → 2

因此就提出了基于目标感知表征的双流自注意力。

$$
\begin{aligned}
g_{z_t}^{(m)} &\leftarrow \text{Attention}\left(\mathrm{Q}=g_{z_t}^{(m-1)}, \mathrm{KV}=\mathbf{h}_{z<t}^{(m-1)}; \theta\right) \\
h_{z_t}^{(m)} &\leftarrow \text{Attention}\left(\mathrm{Q}=h_{z_t}^{(m-1)}, \mathrm{KV}=\mathrm{h}_{z\leq t}^{(m-1)}; \theta\right)
\end{aligned}
$$


Attention
Q
K, V
(a)
(b)
(c)
Masked Two-stream Attention
Attention Masks
Content stream: can see self
Query stream: cannot see self
Sample a factorization order: 3 → 2 → 4 → 1
x1 x2 x3 x4
e(x1) w e(x2) w e(x3) w e(x4) w

reference：

https://zhuanlan.zhihu.com/p/70257427

https://blog.csdn.net/u012526436/article/details/93196139

6 Star-Transformer

本文提出了一种对于Transformer改进的结构“star-Transformer”。将标准Transformer中全连接的部分调整为星型的拓扑结构，在实际任务中速度提升了4.5倍。

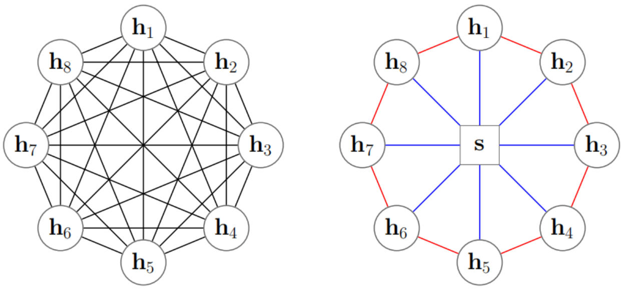


Figure 1: Left: Connections of one layer in Transformer, circle nodes indicate the hidden states of input tokens. Right: Connections of one layer in Star-Transformer, the square node is the virtual relay node. Red edges and blue edges are ring and radical connections, respectively.

添加了一个虚拟的节点作为relay node。由此可以分为两种连接，Radical Connections和Ring Connections。环连接可以有效地减少局部和非局部成分的非偏置学习负担，提高模型的泛化能力。将标准的Transformer复杂度 $O(n^2)$降低为$O(n)$。

根式连接集中在非局部构图上，环连接集中在局部构图上。因此，star-Transformer适用于尺寸适中的数据集，不依赖于繁重的预培训。

令$\mathbf{s}^t \in \mathbb{R}^{1\times d}$ and $\mathbf{H}^t \in \mathbb{R}^{n\times d}$ 分别为relay node和所有的satellite node。

初始化$\mathbf{H}^0 = \mathbf{E}$ 和 $s^0 = \text{average}(\mathbf{E})$

$$\begin{aligned}\mathbf{C}_i^t &= \left[\mathbf{h}_{i-1}^{t-1}; \mathbf{h}_i^{t-1}; \mathbf{h}_{i+1}^{t-1}; \mathbf{e}^i; \mathbf{s}^{t-1}\right]\\ \mathbf{h}_i^t &= \text{ MultiAtt }\left(\mathbf{h}_i^{t-1}, \mathbf{C}_i^t\right)\end{aligned}$$

$$\mathbf{h}_i^t = \text{ LayerNorm }\left(\text{ReLU}\left(\mathbf{h}_i^t\right)\right), i \in [1, n]$$

$$\begin{aligned}\mathbf{s}^t &= \text{ MultiAt }\left(\mathbf{s}^{t-1}, [\mathbf{s}^{t-1}; \mathbf{H}^t]\right)\\ \mathbf{s}^t &= \text{ LayerNorm }\left(\text{ReL U}\left(\mathbf{s}^t\right)\right)\end{aligned}$$