Datawhale

LCF: A Local Context Focus Mechanism for Aspect-Based Sentiment Classification

团队：Datawhale 深度学习团队

汇报人： 杨夕

2020.03.12

目录

01 动机

目录

01

02

03

02 方法介绍

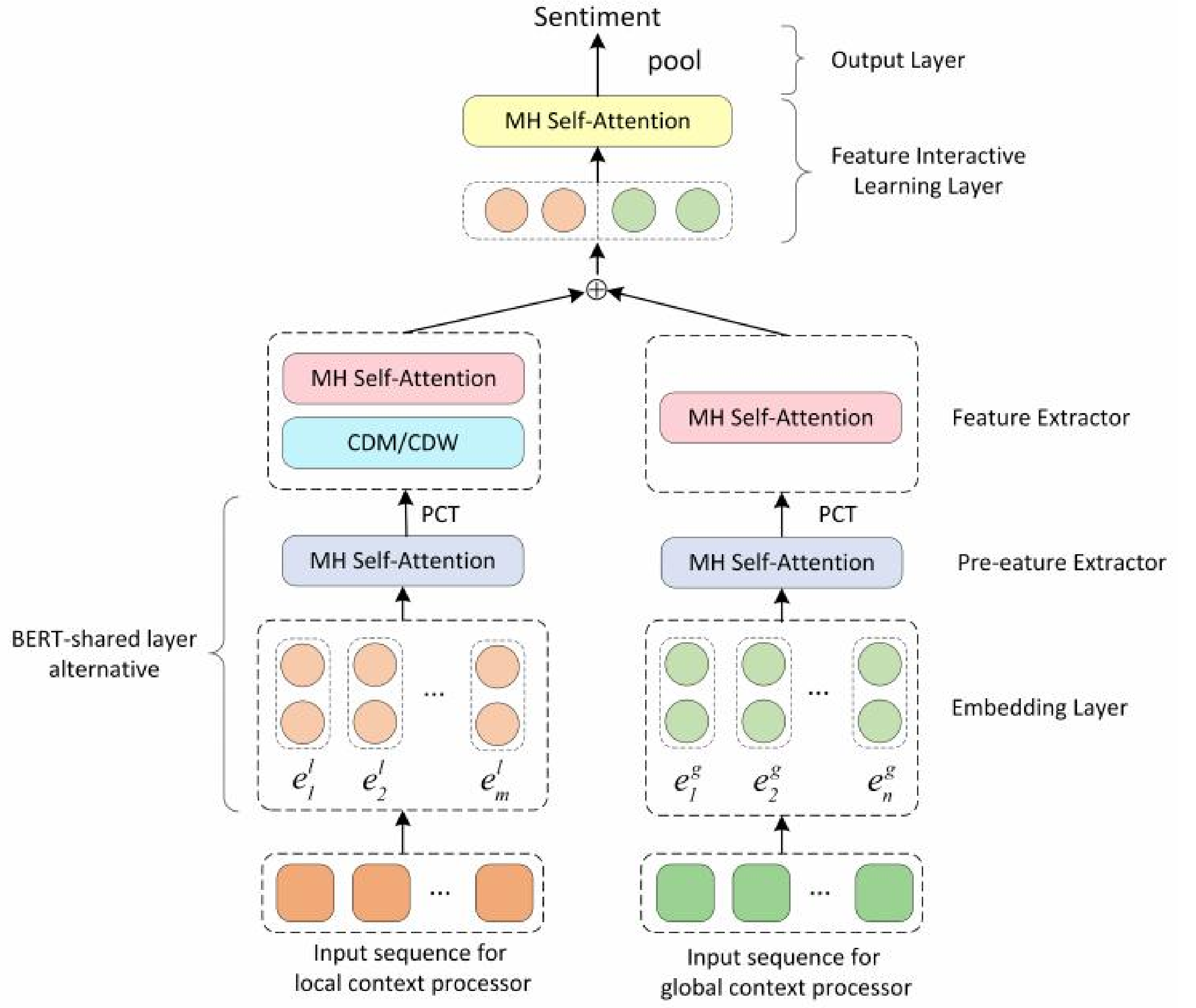


Figure 3. Overall architecture of LCF design. BERT-shared layer is alternative to substitute for embedding layer and Pre-Feature Extractor layer. MH Self-Attention: Multi-Head Self-Attention.

方法介绍——Semantic-Relative Distance

前人研究方法：将输入序列分为aspect序列和上下文序列，然后建立关系

方法存在问题：除全局上下文外，目标aspects的局部上下文也含重要信息

目标：确定目标aspects局部上下文所含的重要信息

while the food is so good and so popular that waiting can really be a nightmare

如何 阈值 设置 为 5 时，红色部分为局部上下文

02 方法介绍——GloVe Word Embedding

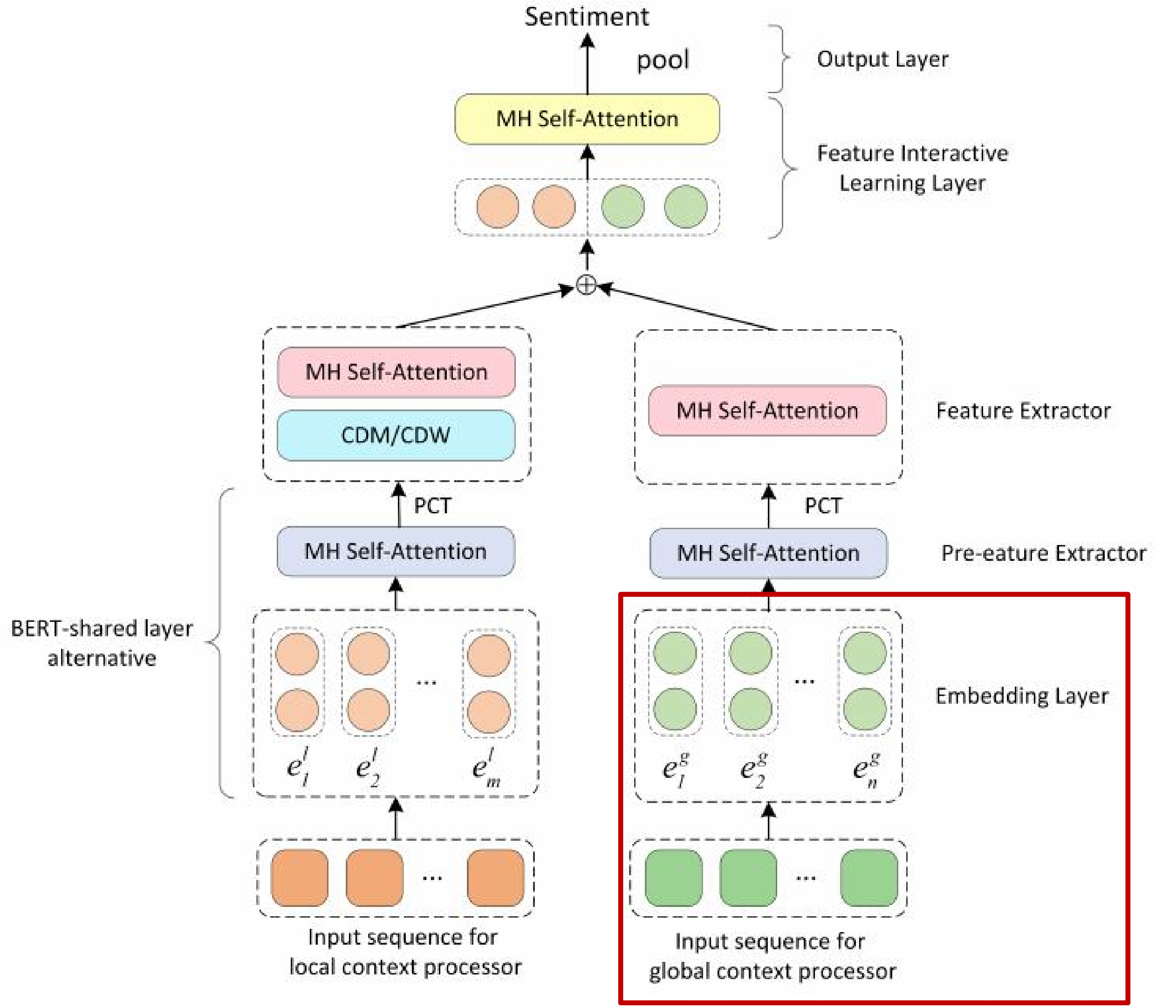


Figure 3. Overall architecture of LCF design. BERT-shared layer is alternative to substitute for embedding layer and Pre-Feature Extractor layer. MH Self-Attention: Multi-Head Self-Attention.

方法介绍——BERT-Shared Layer

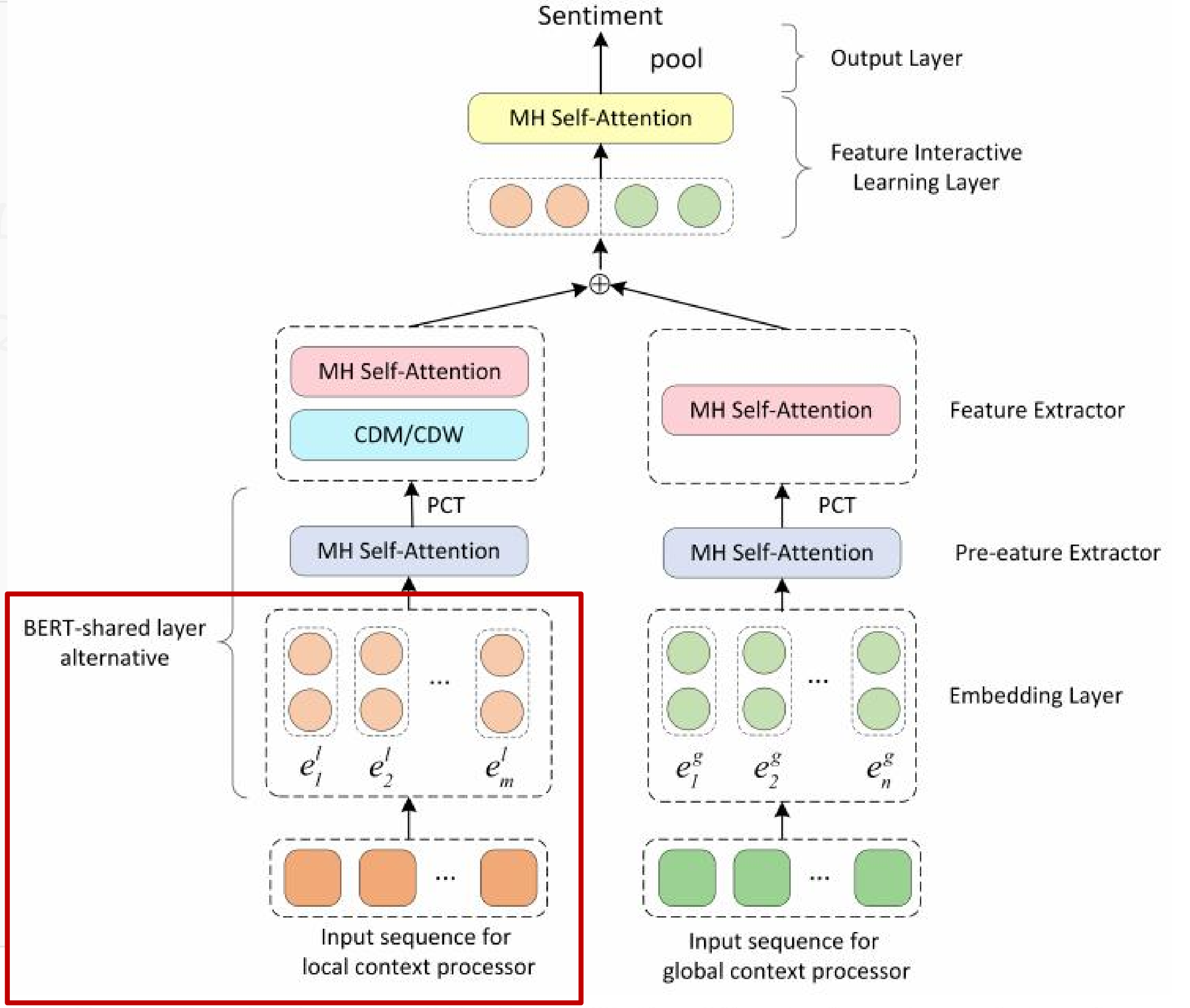


Figure 3. Overall architecture of LCF design. BERT-shared layer is alternative to substitute for embedding layer and Pre-Feature Extractor layer. MH Self-Attention: Multi-Head Self-Attention.

方法：利用两个独立的BERT -shared对 局部上下文序列特征和全局上下文特征进行建模

$$O^{l}_{BERT} = BERT^{l}\left(X^{l}\right)$$

$$O^{g}_{BERT} = BERT^{g}\left(X^{g}\right)$$

02 方法介绍——Pre-Feature Extractor

目录

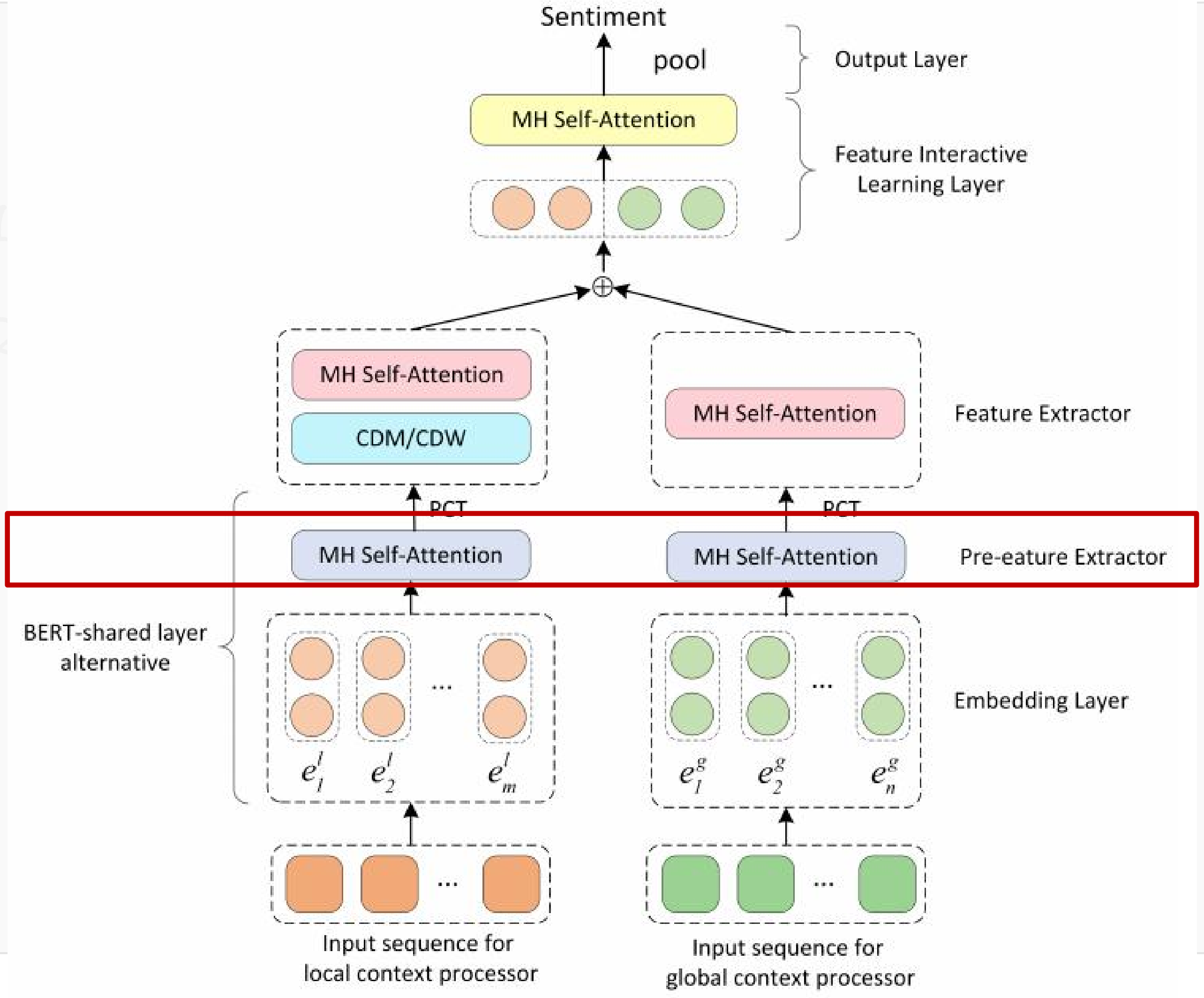


Figure 3. Overall architecture of LCF design. BERT-shared layer is alternative to substitute for embedding layer and Pre-Feature Extractor layer. MH Self-Attention: Multi-Head Self-Attention.

方法：利用 Multi-Head Self-Attention

$$\mathrm{SDA}(X_{\mathrm{SDA}}) = Softmax\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V$$

$$Q, K, V = f_x(X_{SDA})$$

$$f_x(X_{SDA}) = \begin{cases} Q = X_{SDA} \cdot W^q \\ K = X_{SDA} \cdot W^k \\ V = X_{SDA} \cdot W^v \end{cases}$$

$$MHSA(X) = Tanh\left(\{H_0; H_1; \ldots; H_h\} \cdot W^{MH}\right)$$

02 方法介绍——Pre-Feature Extractor

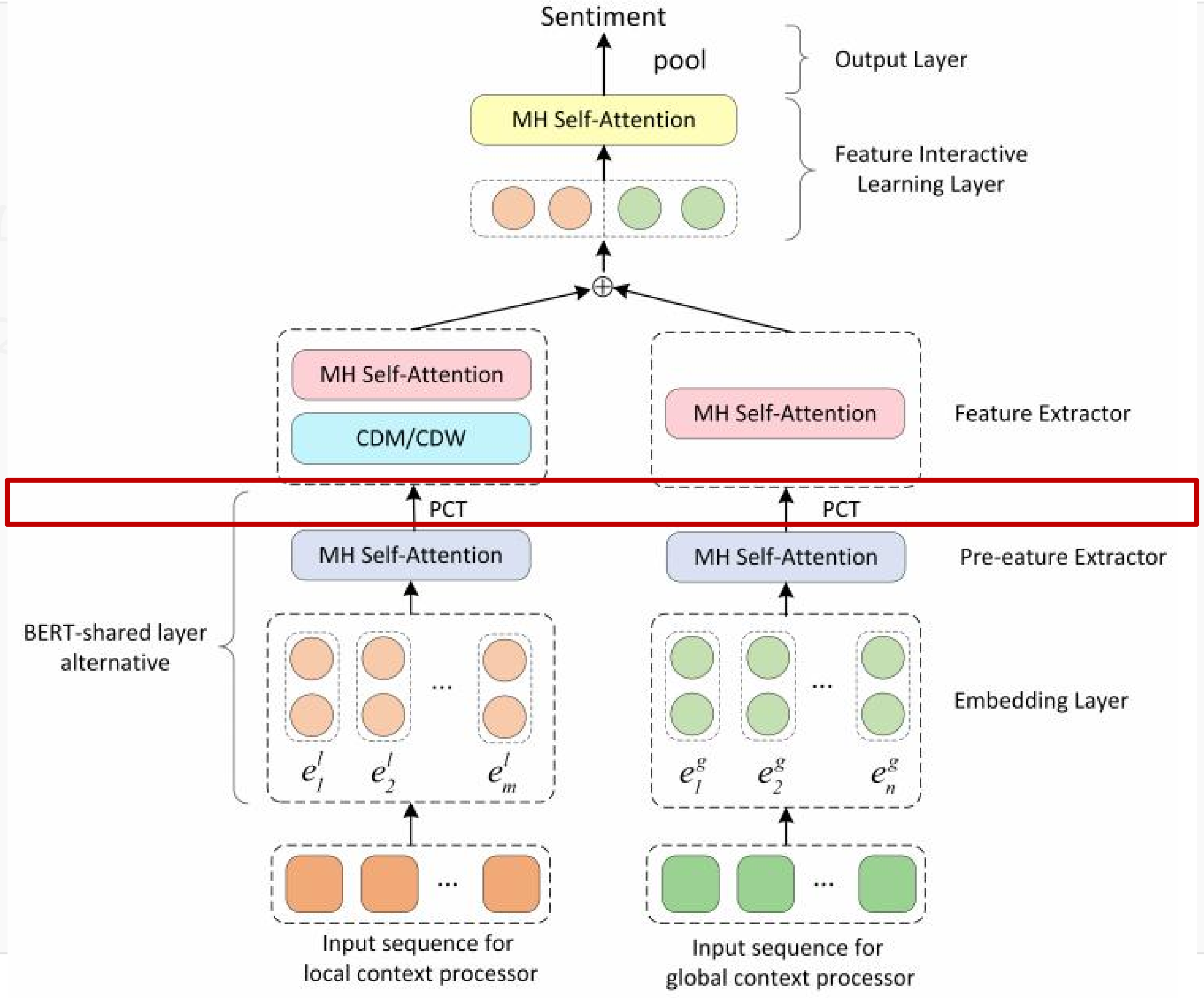


Figure 3. Overall architecture of LCF design. BERT-shared layer is alternative to substitute for embedding layer and Pre-Feature Extractor layer. MH Self-Attention: Multi-Head Self-Attention.

方法： Position-Wise Convolution Transformation (实验定律)

$$PCT\left(O_{mhsa}\right) = ReLU\left(O_{mhsa} * W_1 + b_1\right) * W_2 + b_2$$

$$O_{mhsa}^{l-embed} = MHSA^{l}\left(O_{embed}^{l}\right)$$

$$O_{mhsa}^{g-embed} = MHSA^{g}\left(O_{embed}^{l}\right)$$

$$O_{PFE}^{l} = PCT^{l}\left(O_{\mathrm{mhsa}}^{l-embed}\right)$$

$$O_{PFE}^{g} = PCT^{g}\left(O_{\mathrm{mhsa}}^{g-embed}\right)$$

02 方法介绍——Feature Extractor

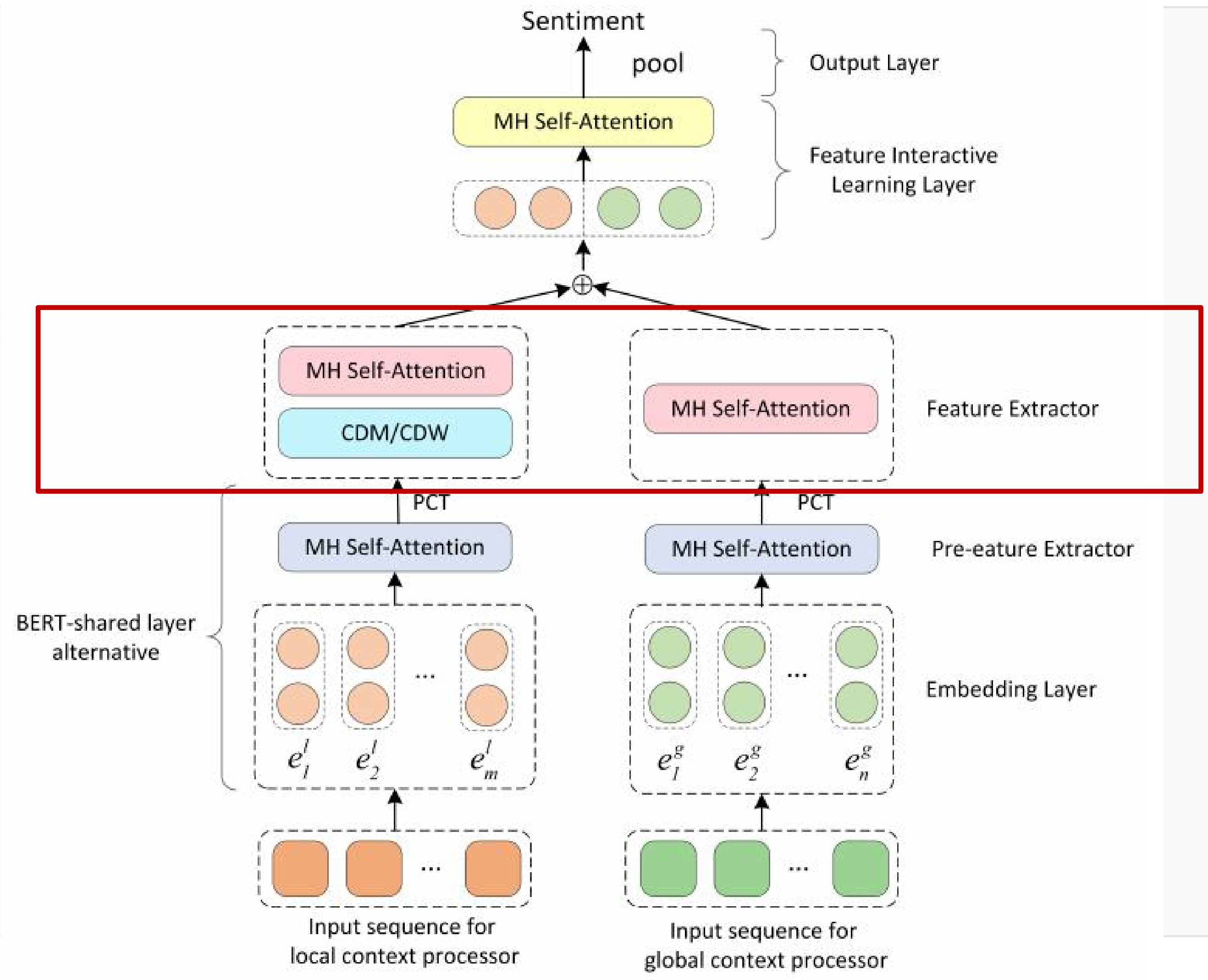


Figure 3. Overall architecture of LCF design. BERT-shared layer is alternative to substitute for embedding layer and Pre-Feature Extractor layer. MH Self-Attention: Multi-Head Self-Attention.

动机：如果只考虑局部上下文，会丢失掉语义相对较少的特征（全局上下文中所含特征和全局上下文与 aspect 间的关系）；

02 方法介绍—— Global Context Focused Layer

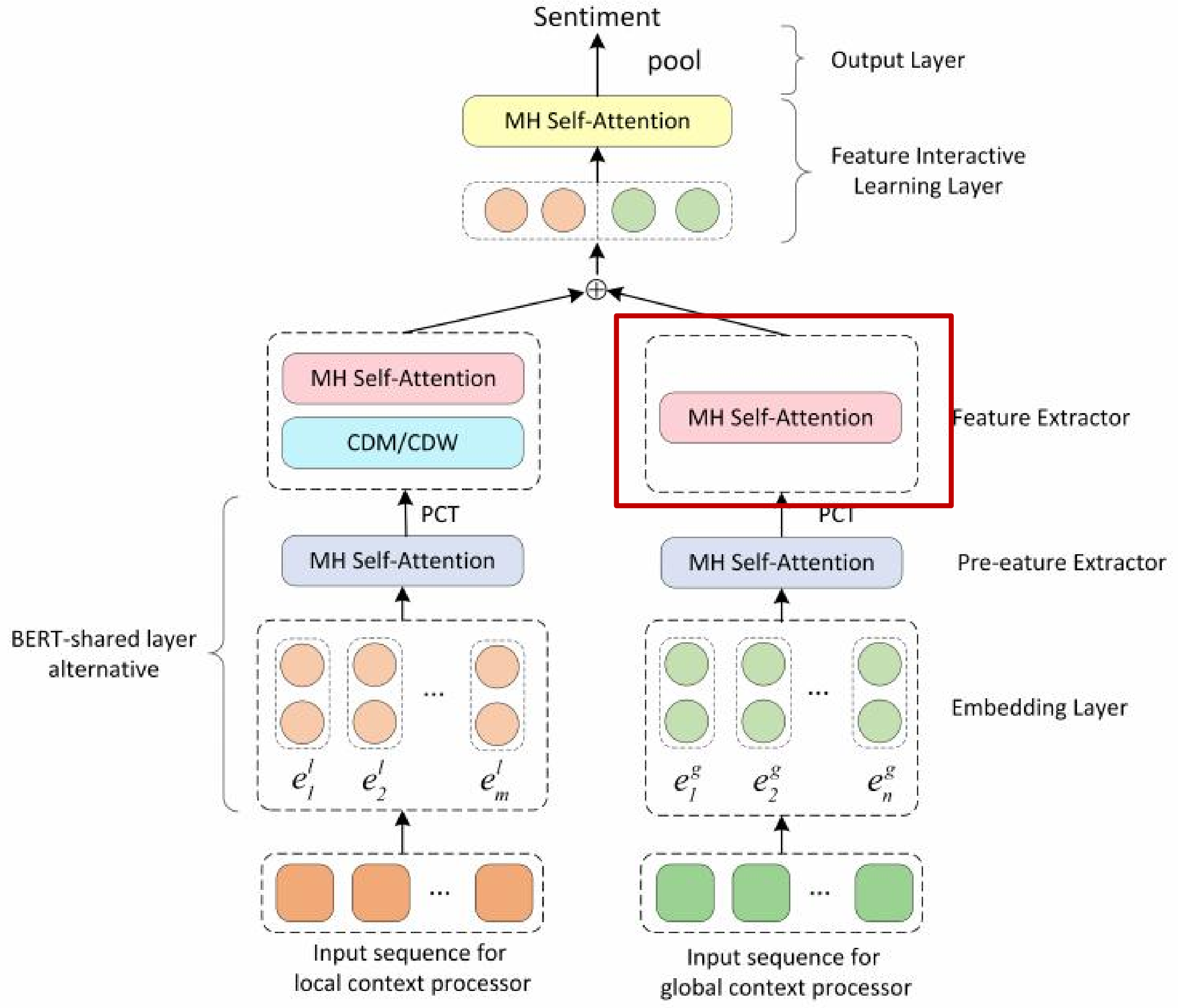


Figure 3. Overall architecture of LCF design. BERT-shared layer is alternative to substitute for embedding layer and Pre-Feature Extractor layer. MH Self-Attention: Multi-Head Self-Attention.

方法：全局上下文特征提取器直接用MHSA

$$O^g = MHSA\left(O^g_{PFE}\right)$$

02 方法介绍—— Local Context Focused Layer

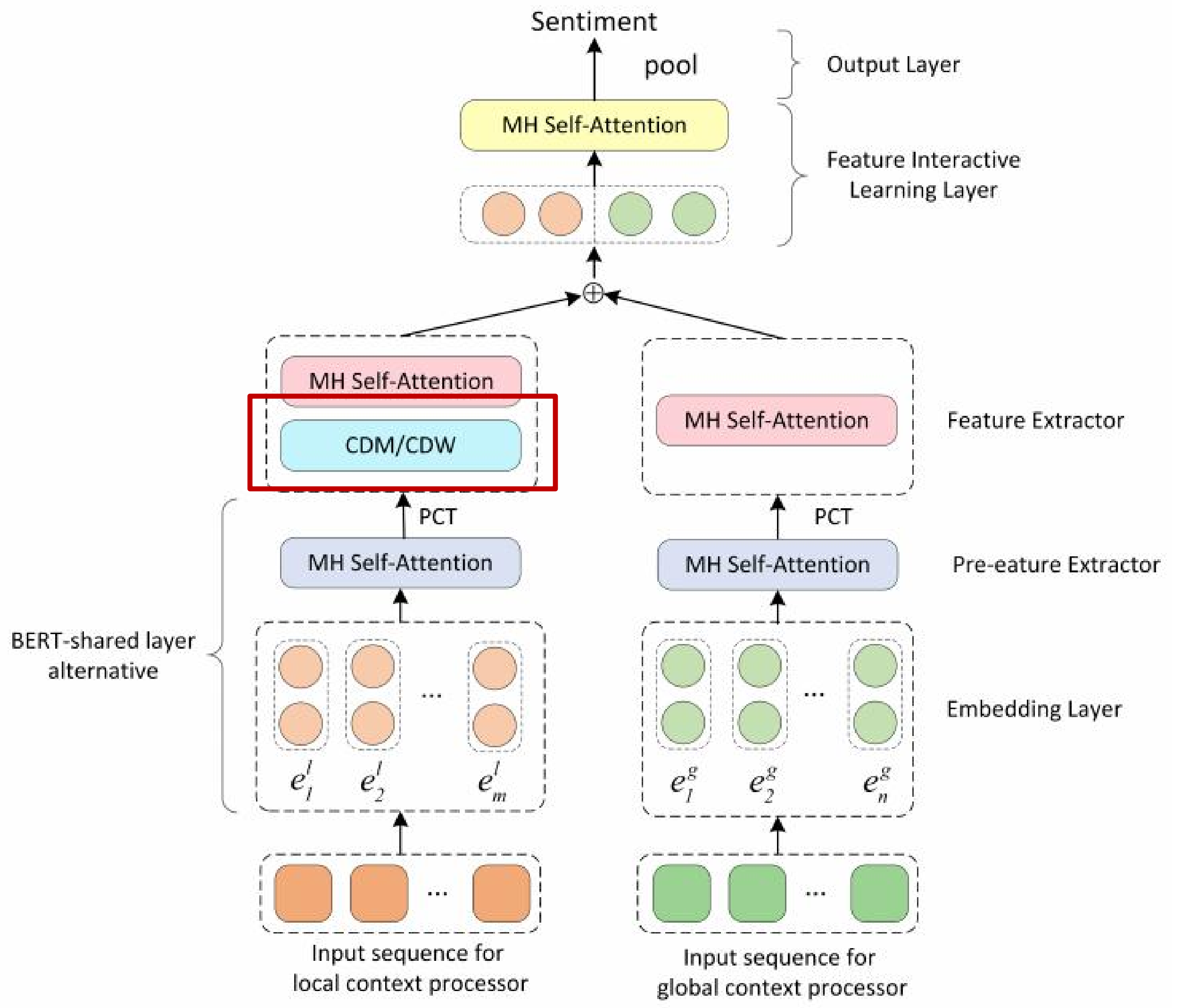


Figure 3. Overall architecture of LCF design. BERT-shared layer is alternative to substitute for embedding layer and Pre-Feature Extractor layer. MH Self-Attention: Multi-Head Self-Attention.

02 方法介绍——Local Context Focused Layer

目录

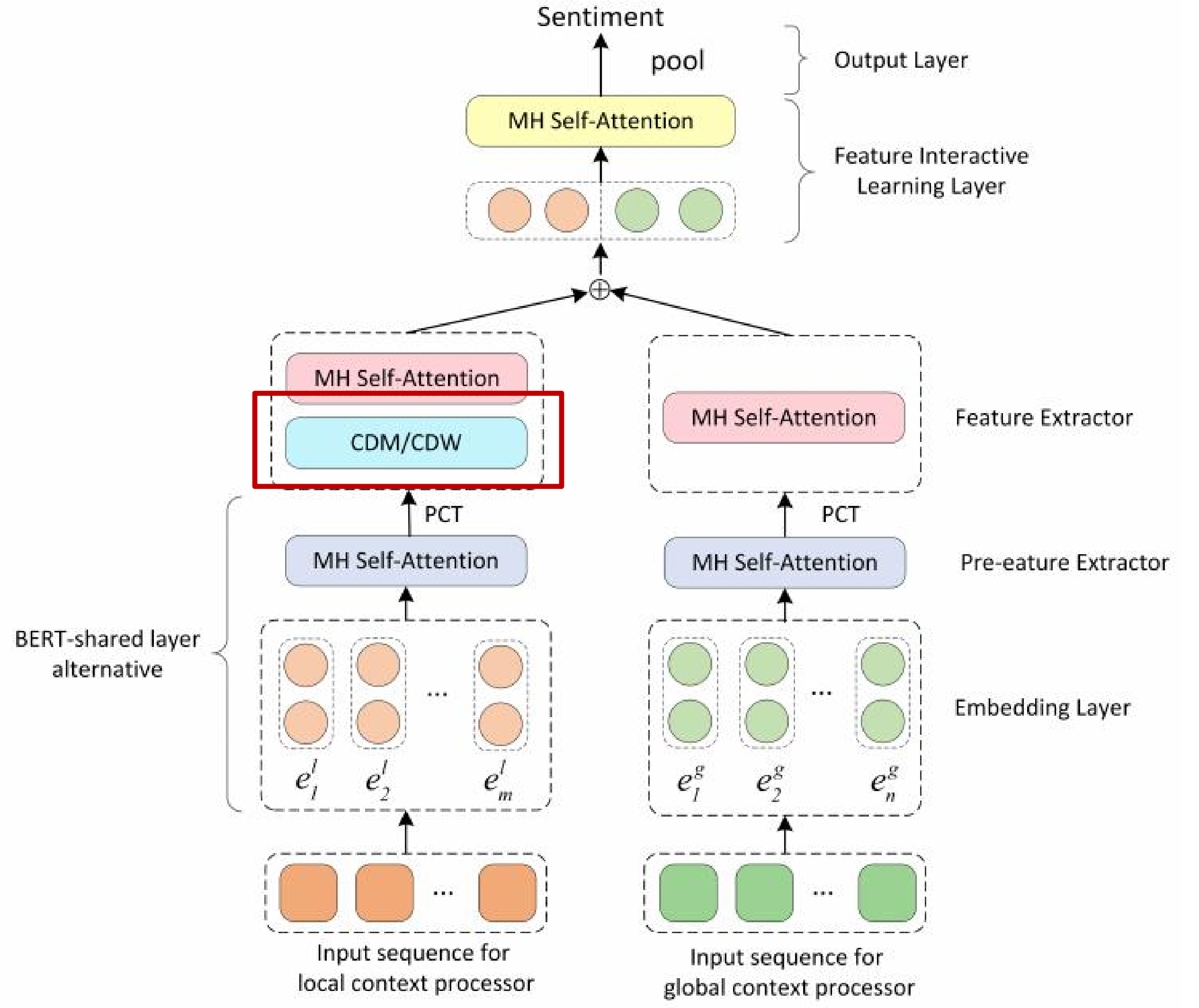


Figure 3. Overall architecture of LCF design. BERT-shared layer is alternative to substitute for embedding layer and Pre-Feature Extractor layer. MH Self-Attention: Multi-Head Self-Attention.

Local Context Focused Layer： CDM 和 CDW

02 方法介绍——Local Context Focused Layer

目录

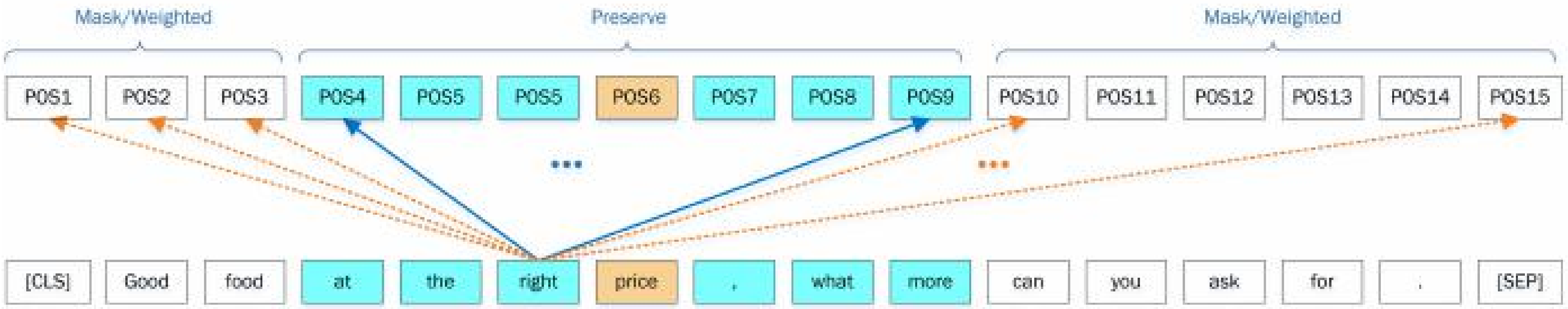


Figure 4. Diagram of the local context focus mechanism. The features of the output position (POS) that the dotted arrow points to will be masked or weighted down and the features of the output position that the solid arrow points to will be completely preserved. The example of context word in this picture is “right”.

方法：在计算完关注层中所有标记的输出后，高于或等于SRD阈值的每个输出位置上的输出特征将被屏蔽或减弱，而局部上下文单词的输出特征将被完全保留。

02 方法介绍——Local Context Focused Layer

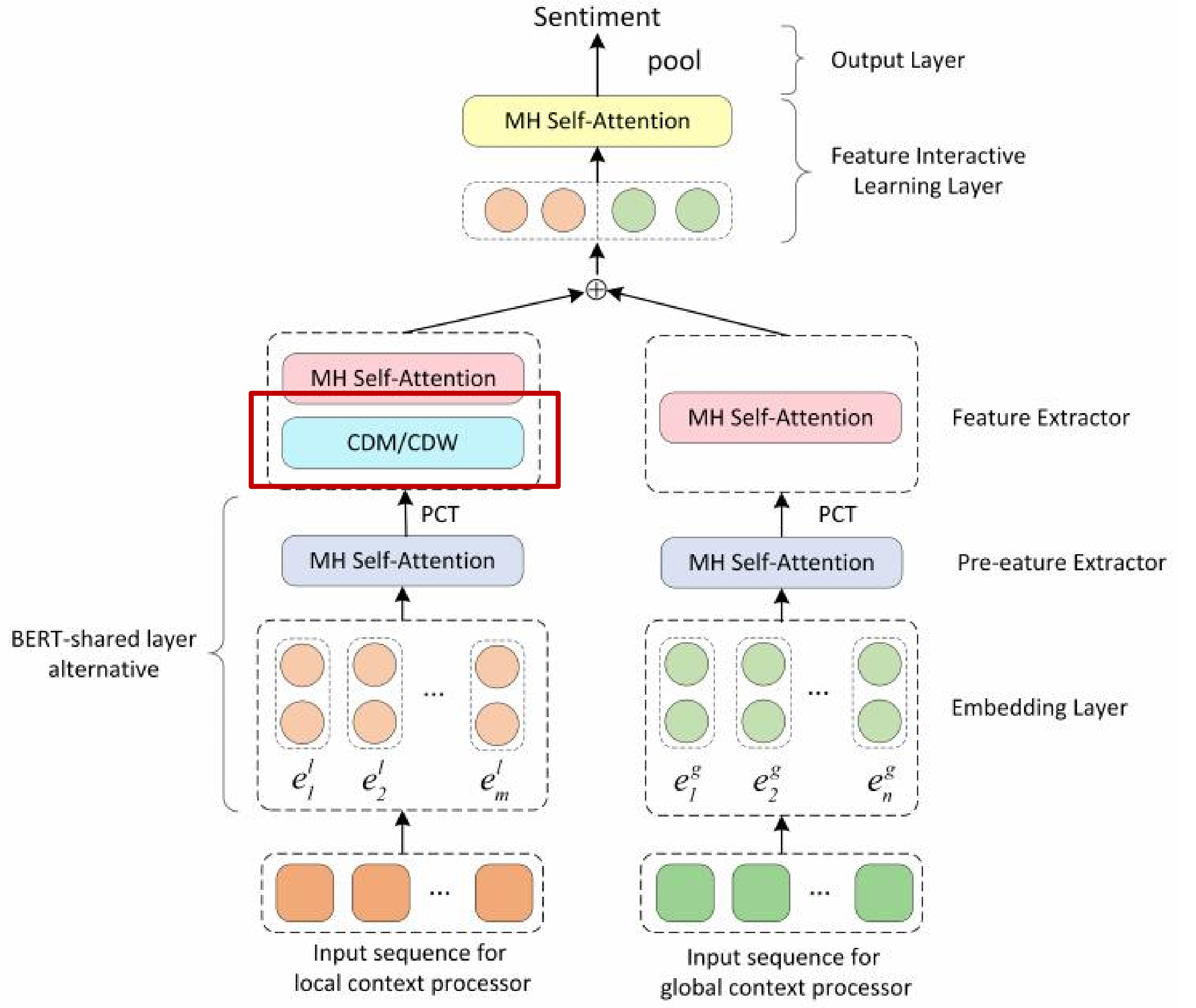


Figure 3. Overall architecture of LCF design. BERT-shared layer is alternative to substitute for embedding layer and Pre-Feature Extractor layer. MH Self-Attention: Multi-Head Self-Attention.

Local Context Focused Layer： CDM 和 CDW

Table 1. Algorithm flow of CDM and CDW mechanism.

1	Accepting the out representation O^l_{PFE} or O^l_{BERT} delivered from local context processor
2	Calculating the SRDs for each context word regarding to a specific aspect
3	For CDM mechanism: Constructing the mask matrices M for the input sequence according to SRDs For CDW mechanism: Constructing the weighting matrices W for the input sequence according to SRDs
4	Applying a matrix element-wise product operation for O^l_{PFE} and M or O^l_{BERT} and M
5	Output the representation of local context words O^l

02 方法介绍——Local Context Focused Layer

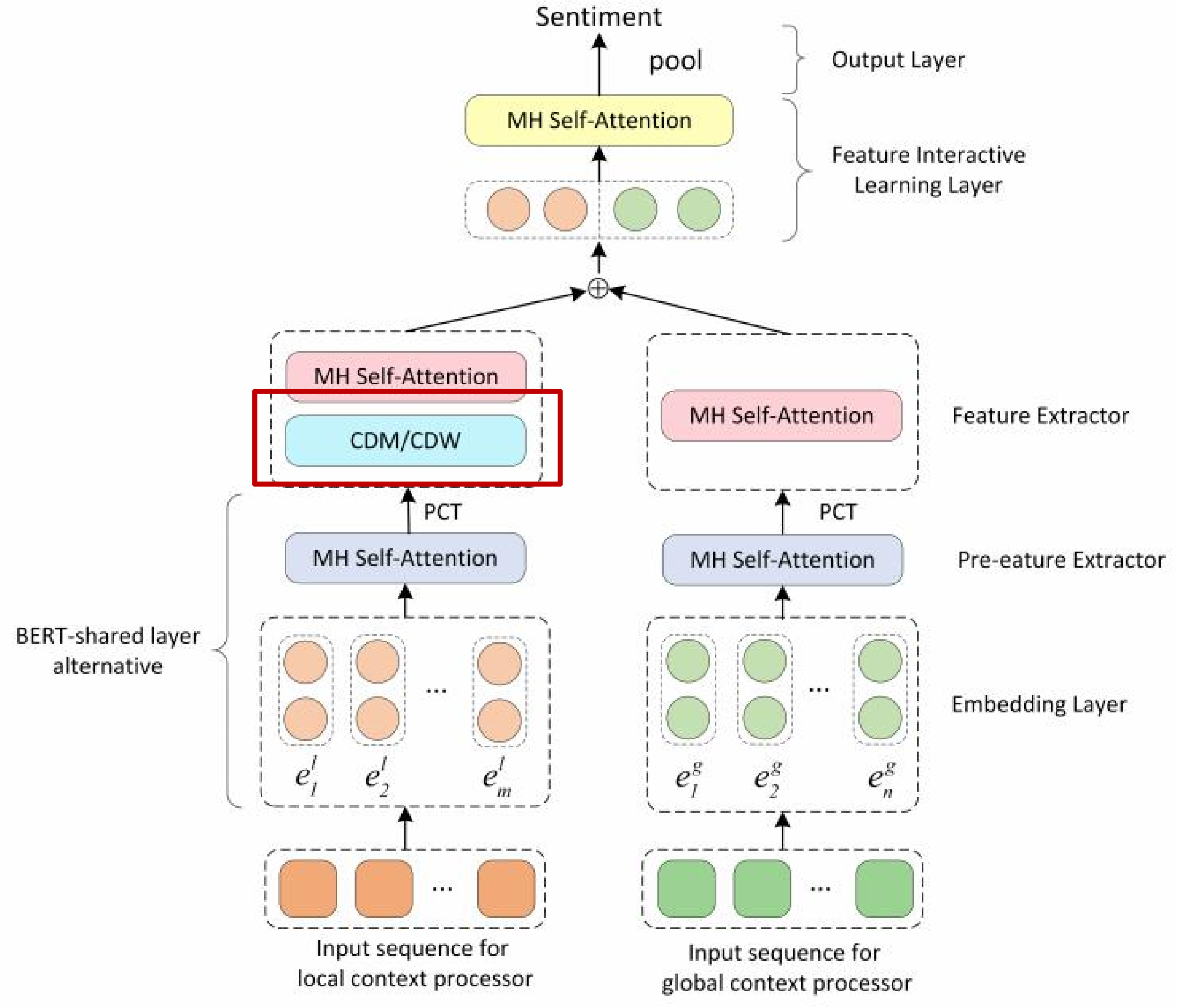


Figure 3. Overall architecture of LCF design. BERT-shared layer is alternative to substitute for embedding layer and Pre-Feature Extractor layer. MH Self-Attention: Multi-Head Self-Attention.

Local Context Focused Layer： CDM

方法：掩盖了PFE或BERT共享层所学习的语义相对较少的上下文特征(掩码相对语义少的上下文词在相对输出位置上的特征，与aspect 间关系将被保留)

$$V_i = \begin{cases} E & SRD_i \le \alpha \\ O & SRD_i > \alpha \end{cases}$$

$$M = [V_0^m, V_1^m, \ldots V_n^m]$$

$$O_{CDM}^l = O_{PFE}^l \cdot M$$

Local Context Focused Layer： CDM

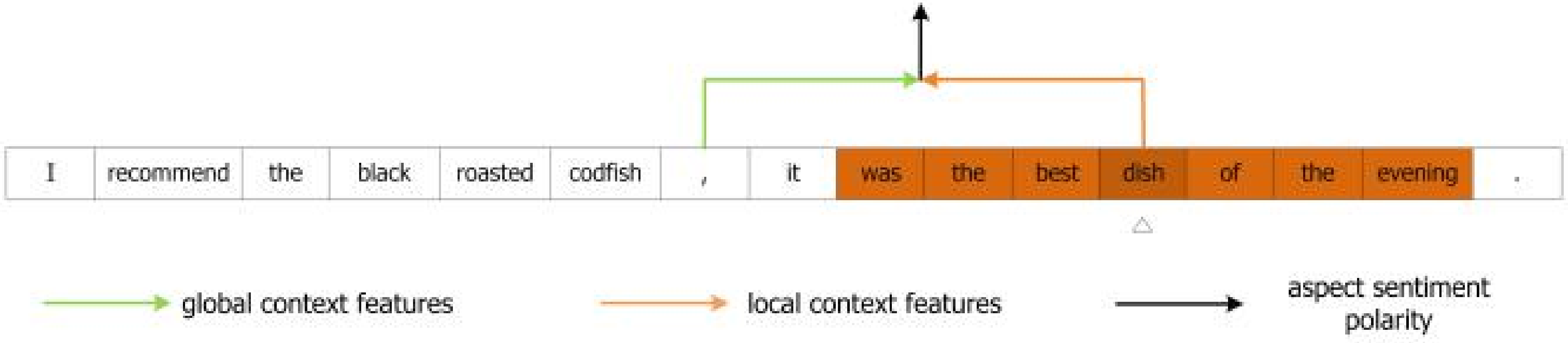


Figure 9. Dynamic mask for local context features during MHSA encoding process on sample-1. The features of the corresponding positions of local context words in the white boxes will be masked.

02 方法介绍——Local Context Focused Layer

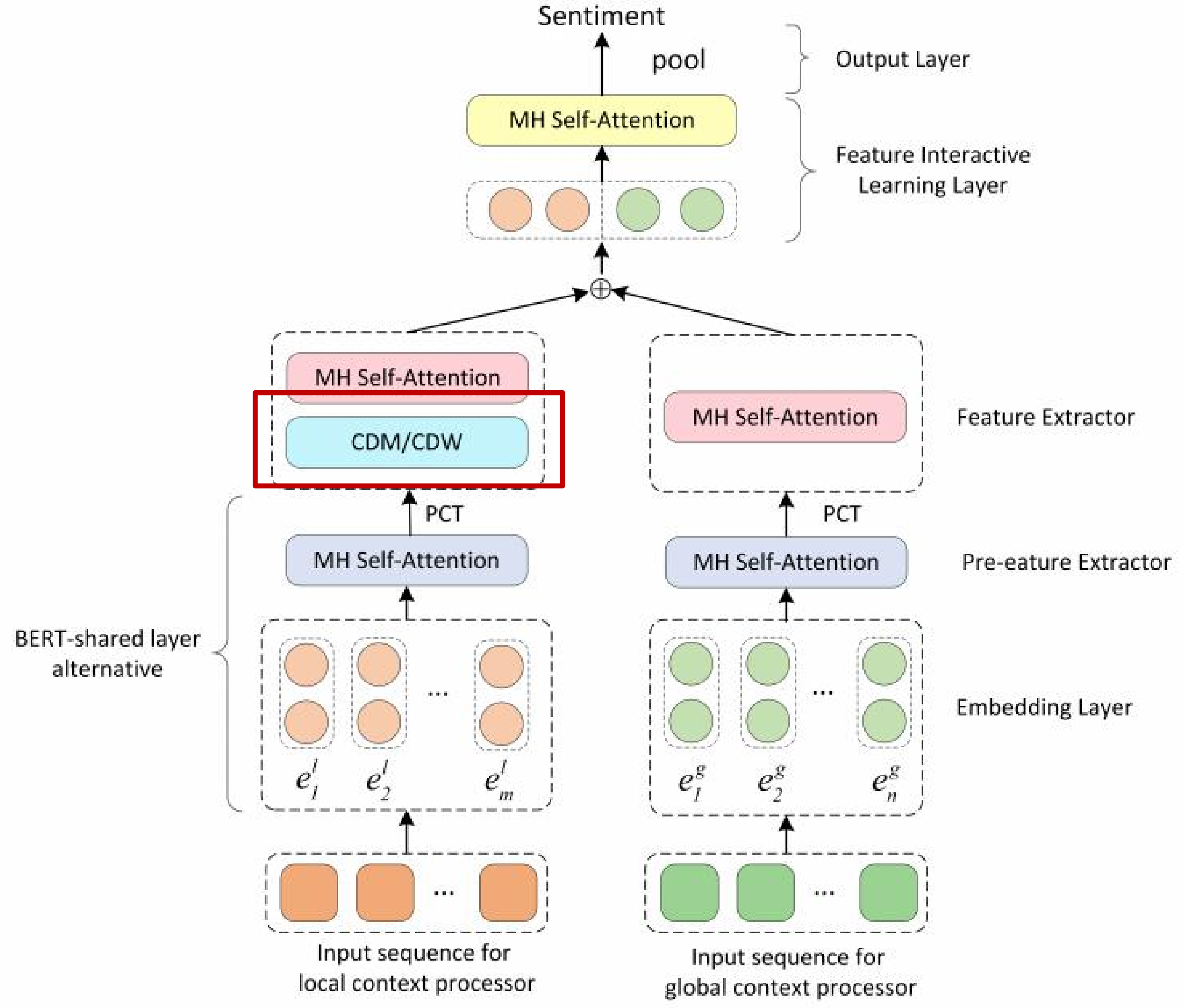


Figure 3. Overall architecture of LCF design. BERT-shared layer is alternative to substitute for embedding layer and Pre-Feature Extractor layer. MH Self-Attention: Multi-Head Self-Attention.

Local Context Focused Layer： CDW

方法：保留相对语义上下文词的特征，但会被加权衰减

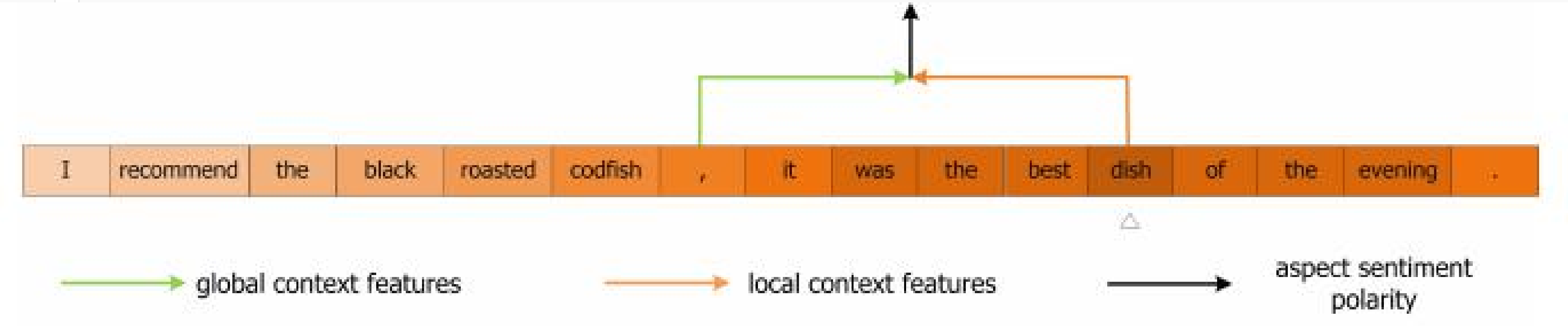
$$V_i = \begin{cases} E & SRD_i \le \alpha \\ \frac{SRD_i - \alpha}{n} \cdot E & SRD_i > \alpha \end{cases}$$

$$W = [V_0^w, V_1^w, \ldots V_n^w]$$

$$O_{CDW}^l = O_{PFE}^l \cdot W$$

方法介绍——Local Context Focused Layer

Local Context Focused Layer： CDW



02 方法介绍——Local Context Focused Layer

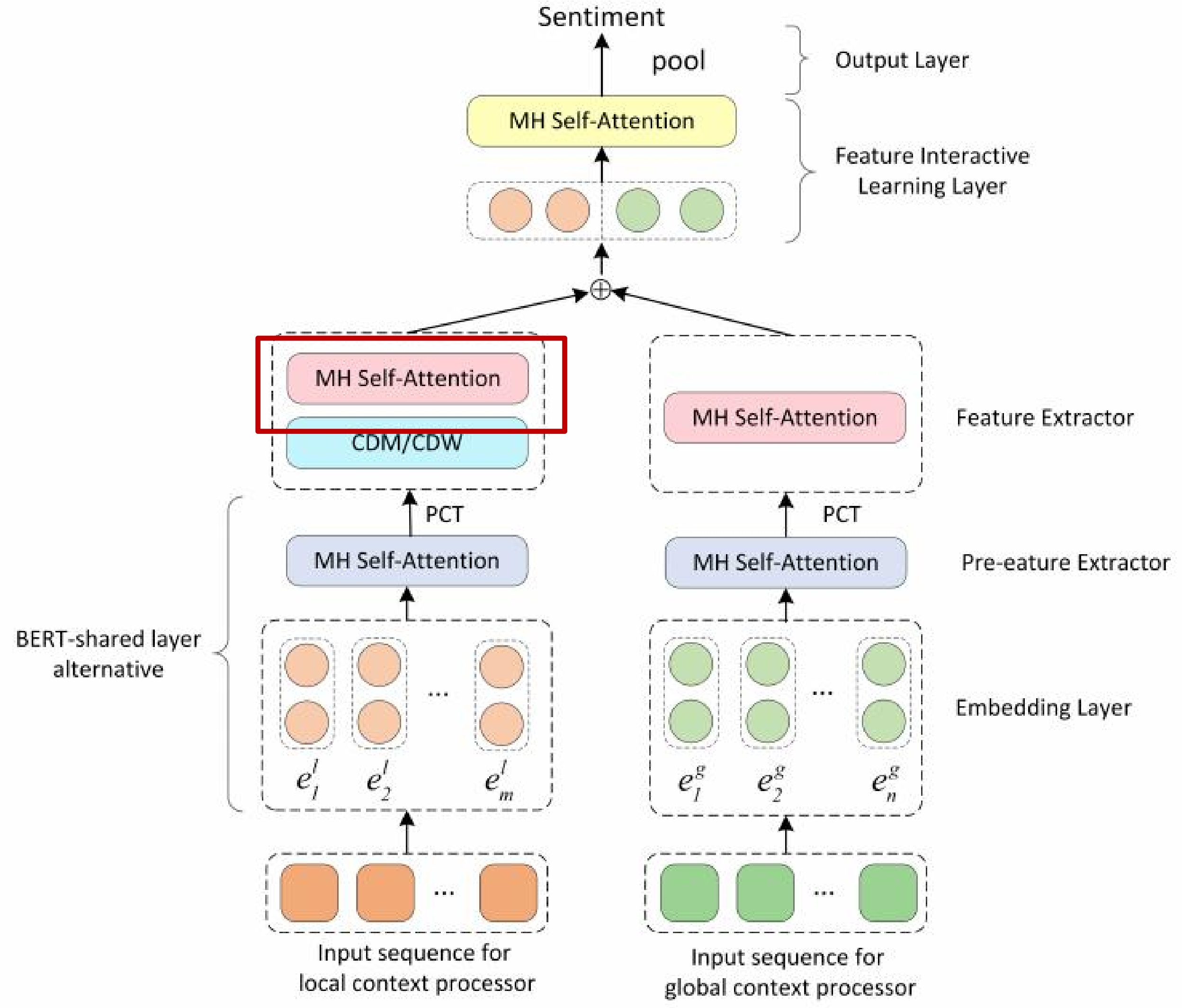


Figure 3. Overall architecture of LCF design. BERT-shared layer is alternative to substitute for embedding layer and Pre-Feature Extractor layer. MH Self-Attention: Multi-Head Self-Attention.

$$O^l = MHSA\left(O^l_{CDM}\right)$$

$$O^l = MHSA\left(O^l_{CDW}\right)$$

方法介绍—— Feature Interactive Learning Layer

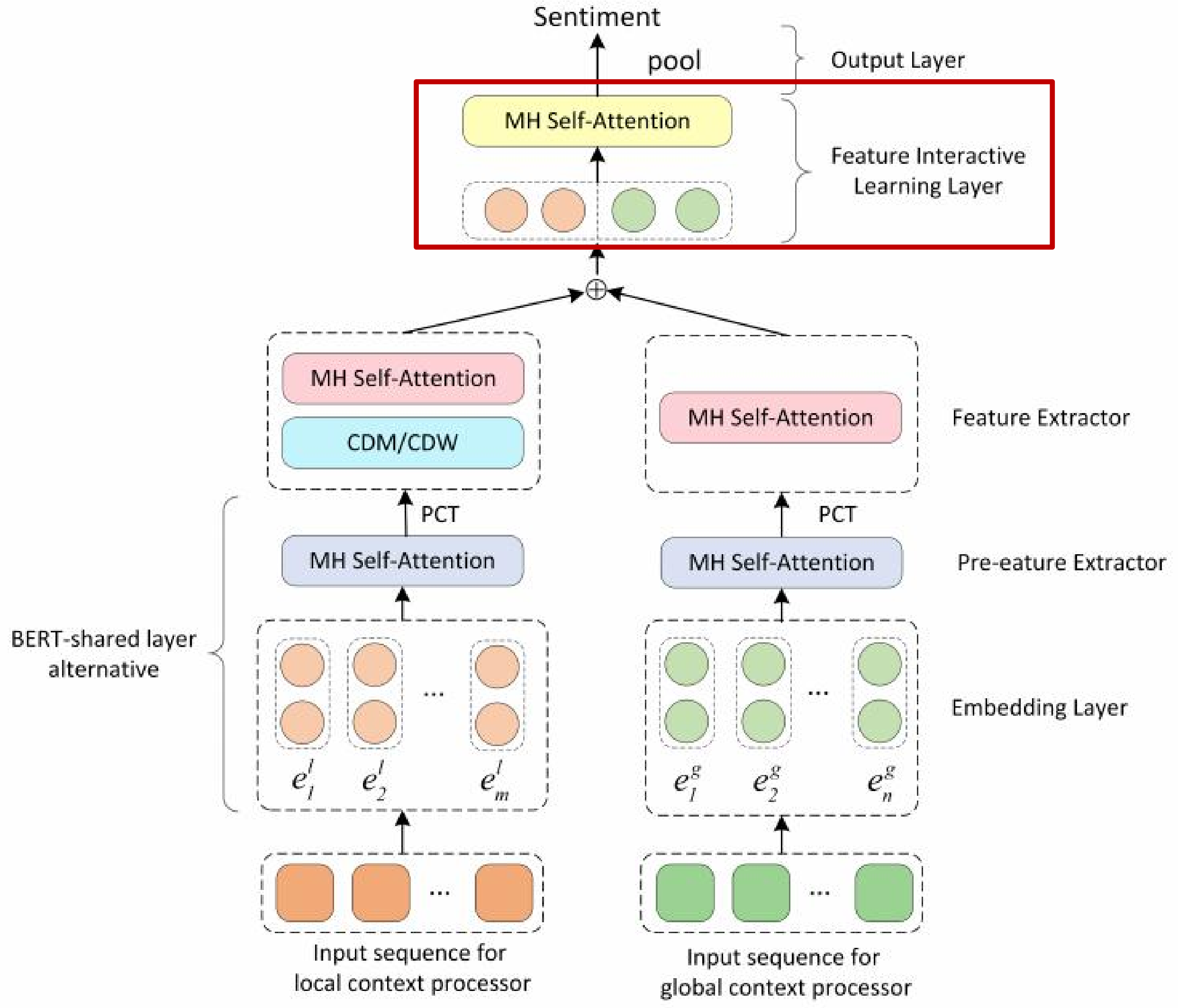


Figure 3. Overall architecture of LCF design. BERT-shared layer is alternative to substitute for embedding layer and Pre-Feature Extractor layer. MH Self-Attention: Multi-Head Self-Attention.

$$O^{lg} = \left[O^l; O^g\right]$$

$$O^{lg}_{dense} = W^{lg} \cdot O^{lg} + b^{lg}$$

$$O^{lg}_{FIL} = MHSA\left(O^{lg}_{dense}\right)$$

02 方法介绍—— Output Layer

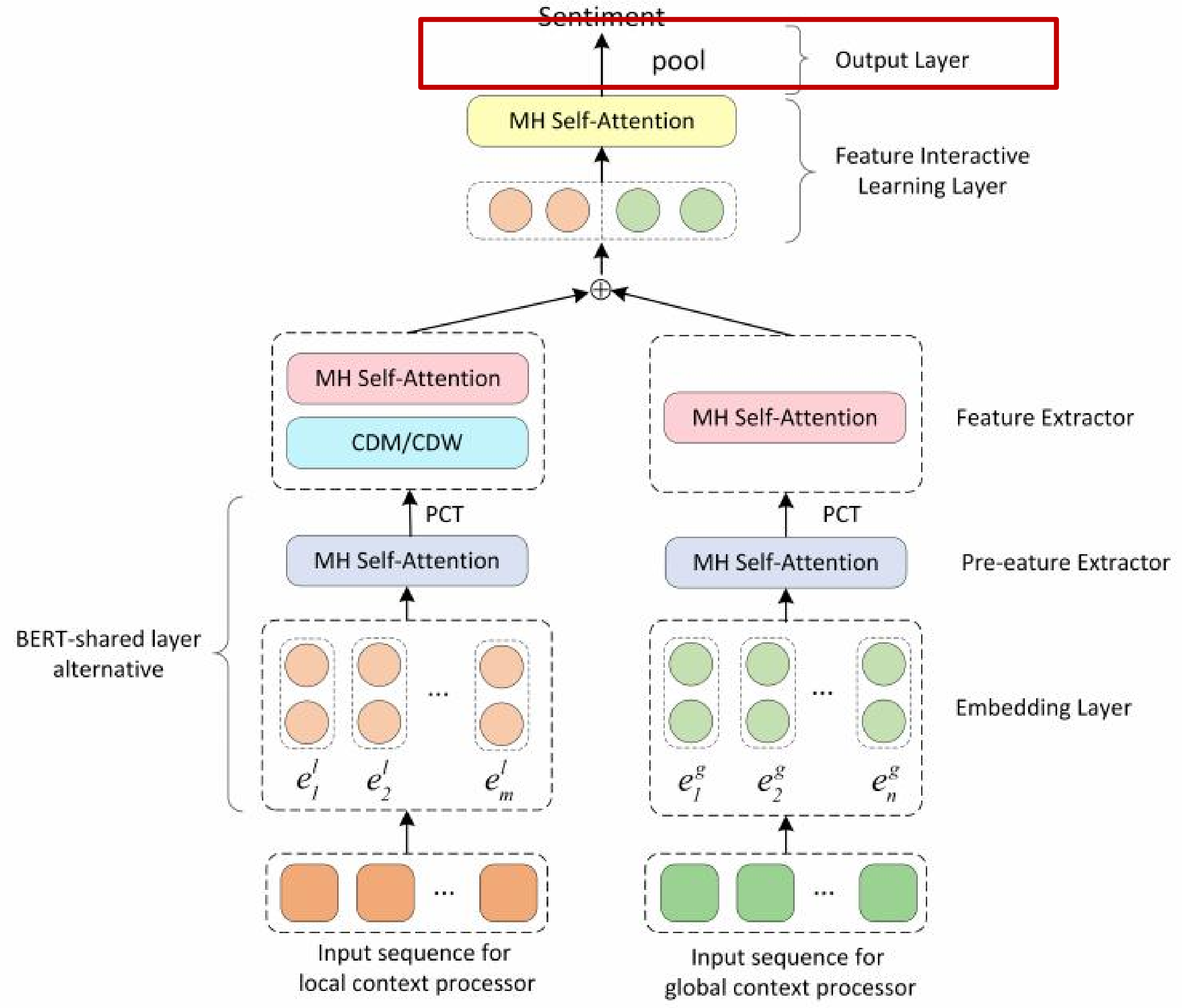


Figure 3. Overall architecture of LCF design. BERT-shared layer is alternative to substitute for embedding layer and Pre-Feature Extractor layer. MH Self-Attention: Multi-Head Self-Attention.

$$X^{lg}_{pool} = POOL\left(O^{lg}_{FIL}\right)$$

$$Y = Softmax\left(X^{lg}_{pool}\right) = \frac{\exp(X^{lg}_{pool})}{\sum_{k=1}^{C}\exp(X^{lg}_{pool})}$$

03 实验结果分析——数据集介绍

Table 2. Detail of benchmark datasets.

Datasets	Positive		Negative		Neural	
	Train	Test	Train	Test	Train	Test
Laptop	994	341	870	128	464	169
Restaurant	2164	728	807	196	637	196
Twitter	1561	173	1560	173	3127	346

03 实验结果分析——模型对比

	Models	Laptop		Restaurant		Twitter	
		Accuracy (%)	F1 (%)	Accuracy (%)	F1 (%)	Accuracy (%)	F1 (%)
Baselines	TD-LSTM	68.13	-	75.63	-	70.8	69
	ATAE-LSTM	68.7	-	77.2	-	-	-
	IAN	72.1	-	78.6	-	-	-
	RAM	74.49	71.35	80.23	70.8	69.36	67.30
	MGAN	75.39	72.47	81.25	71.94	72.54	70.81
BERT Models	BERT-PT	78.07	75.08	84.95	76.96	-	-
	BERT-SPC	80.25	77.41	85.98	78.79	75.29	73.63
LCF-GloVe	Glo-CDM	76.02	70.58	82.5	73.92	72.25	70.92
	Glo-CDW	75.24	71.46	81.61	72.26	71.82	69.83
LCF-BERT	Glo-CDM	**82.29**	**79.28**	**86.52**	**80.4**	**76.45**	**75.52**
	Glo-CDW	**82.45**	**79.59**	**87.14**	**81.74**	**77.31**	**75.78**

03

实验结果分析——LCF-GloVe Variations about SRD Threshold

目录

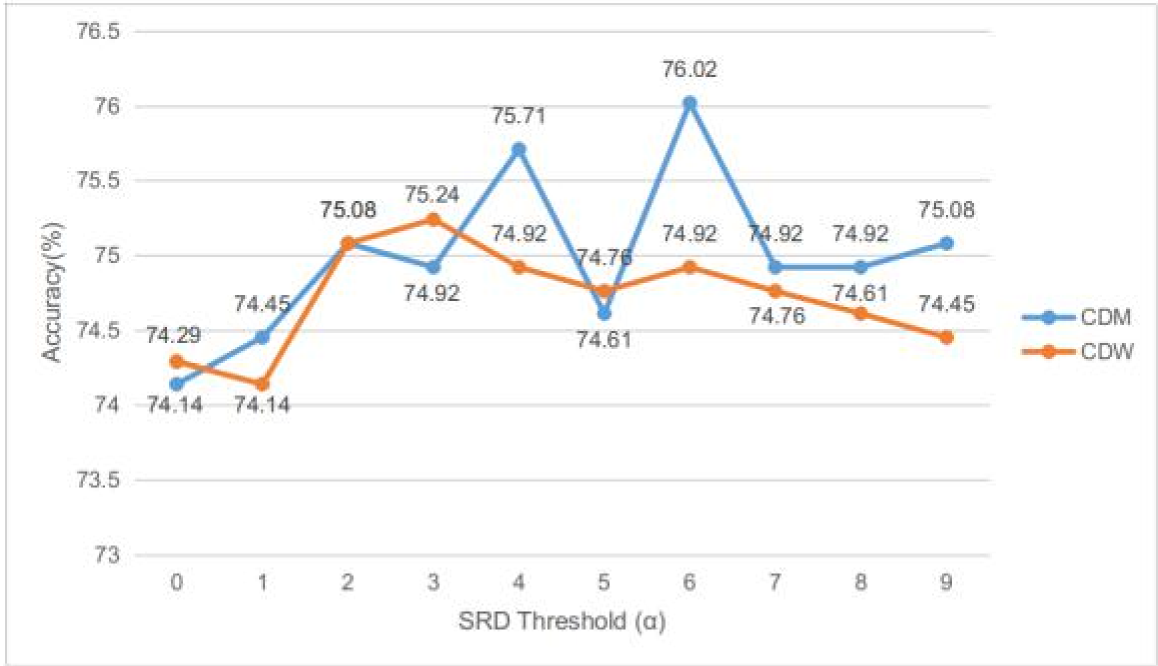


Figure 6. Accuracy on laptop dataset of LCF-GloVe model under different SRD thresholds (α) .

03 实验结果分析——LCF-GloVe Variations about SRD Threshold

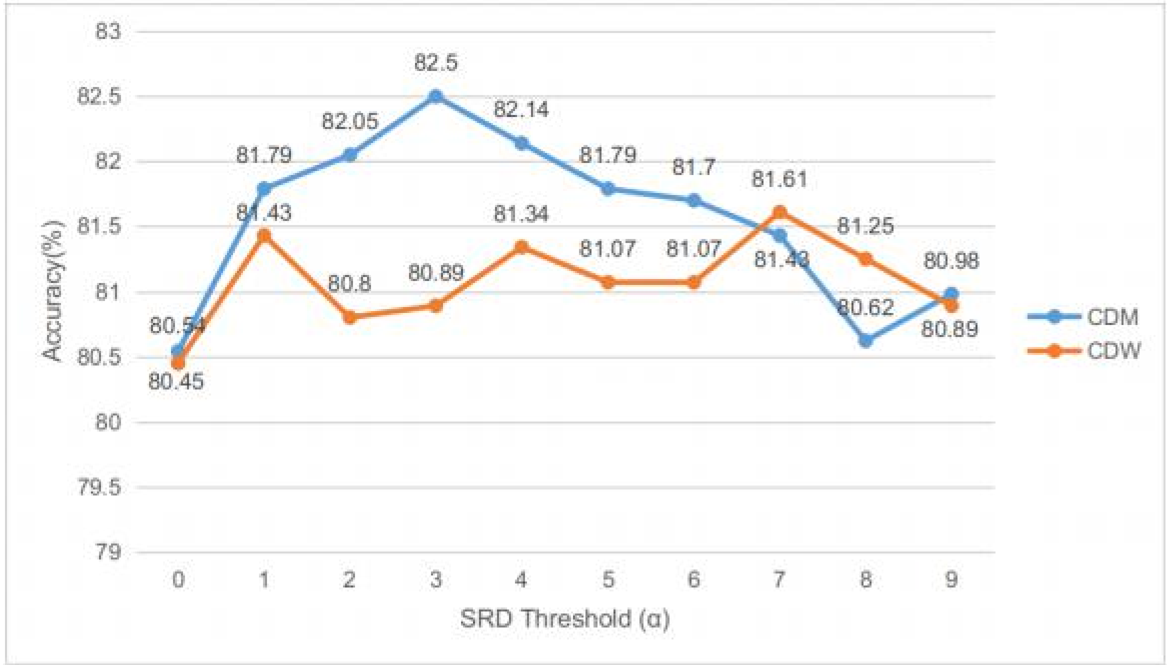


Figure 7. Accuracy on restaurant dataset of LCF-GloVe model under different SRD thresholds (α) .

03

实验结果分析——LCF-GloVe Variations about SRD Threshold

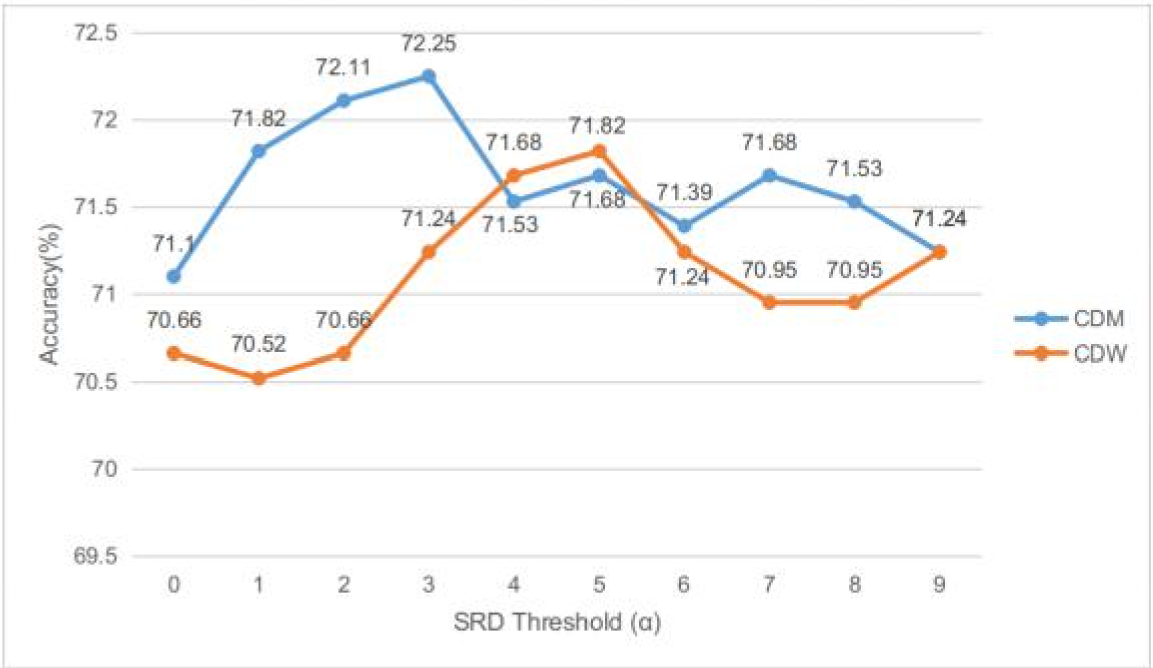


Figure 8. Accuracy on twitter dataset of LCF-GloVe model under different SRD thresholds (α) .

04 个人总结

感觉就是模型的组合，唯一眼前一亮的是对 CDM and CDW，以及实验分析。

敬请各位大佬批评指正