知乎搜索

文本相关性和知识蒸馏

申站@知乎搜索团队

2020-11-21

知乎

大纲

- 知乎搜索文本相关性的演进
- BERT 的应用和问题
- 知识蒸馏及常用方案
- 知乎搜索在 BERT 蒸馏上的实践

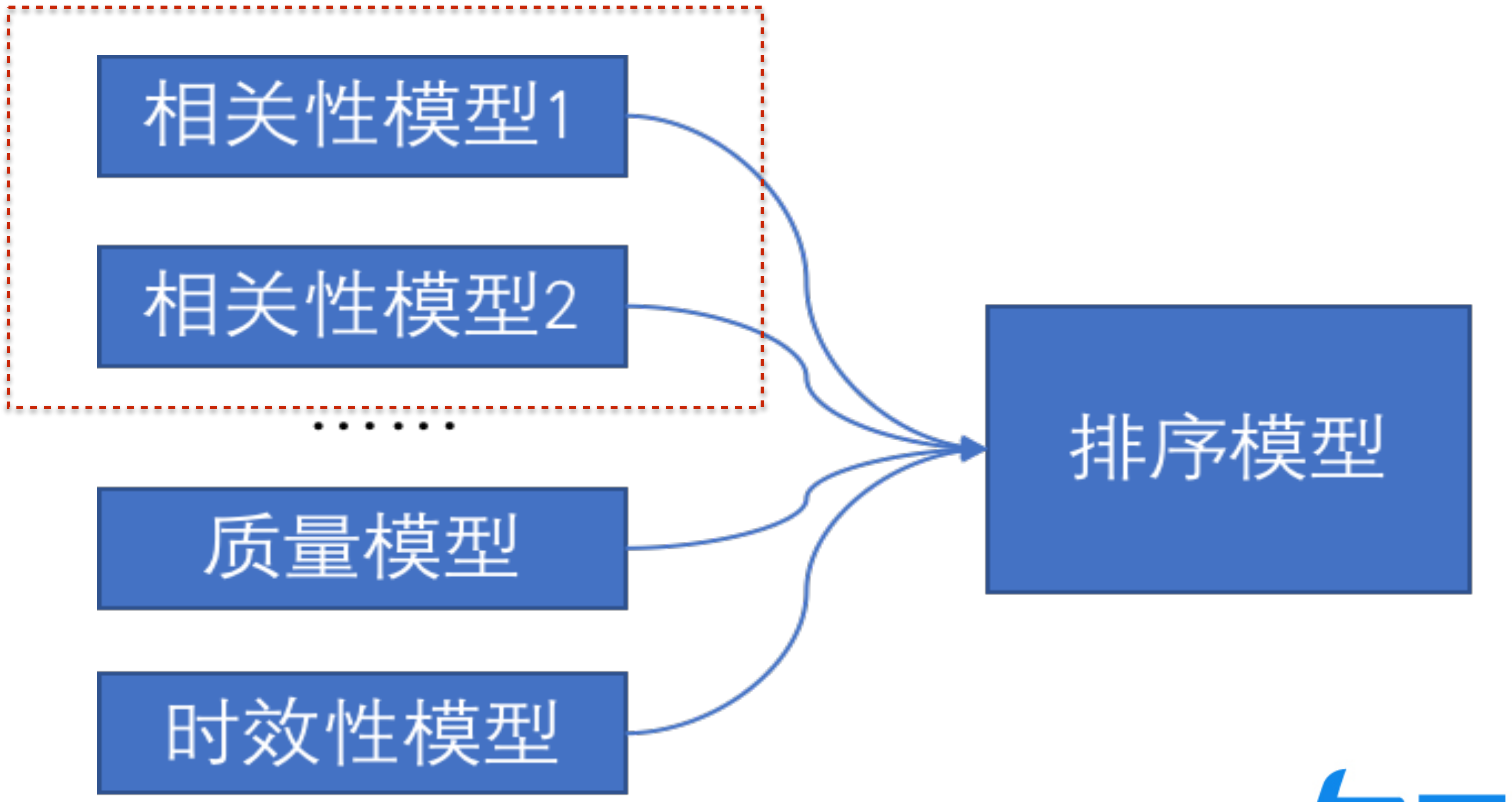
知乎

文本相关性的演进

定义：用户 query 意图和 doc 内容的相关程度

相关性两个维度：

- 字面匹配
- 语义相关



知乎

文本相关性的演进

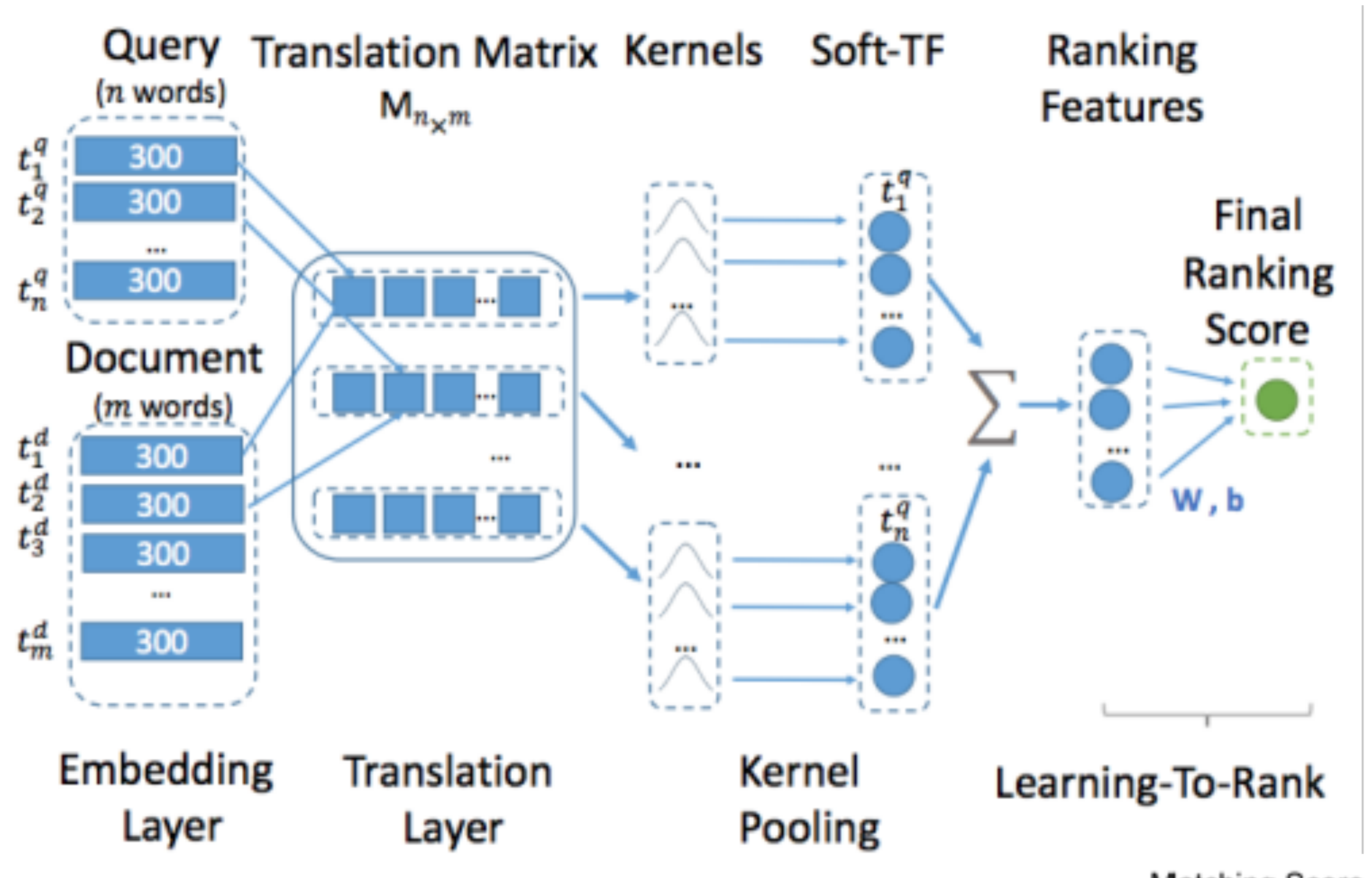
- **Before NN**
  - TF-IDF/BM25
  - 词频/权重/覆盖率
  - 紧密度/同义词
- Before BERT
- BERT

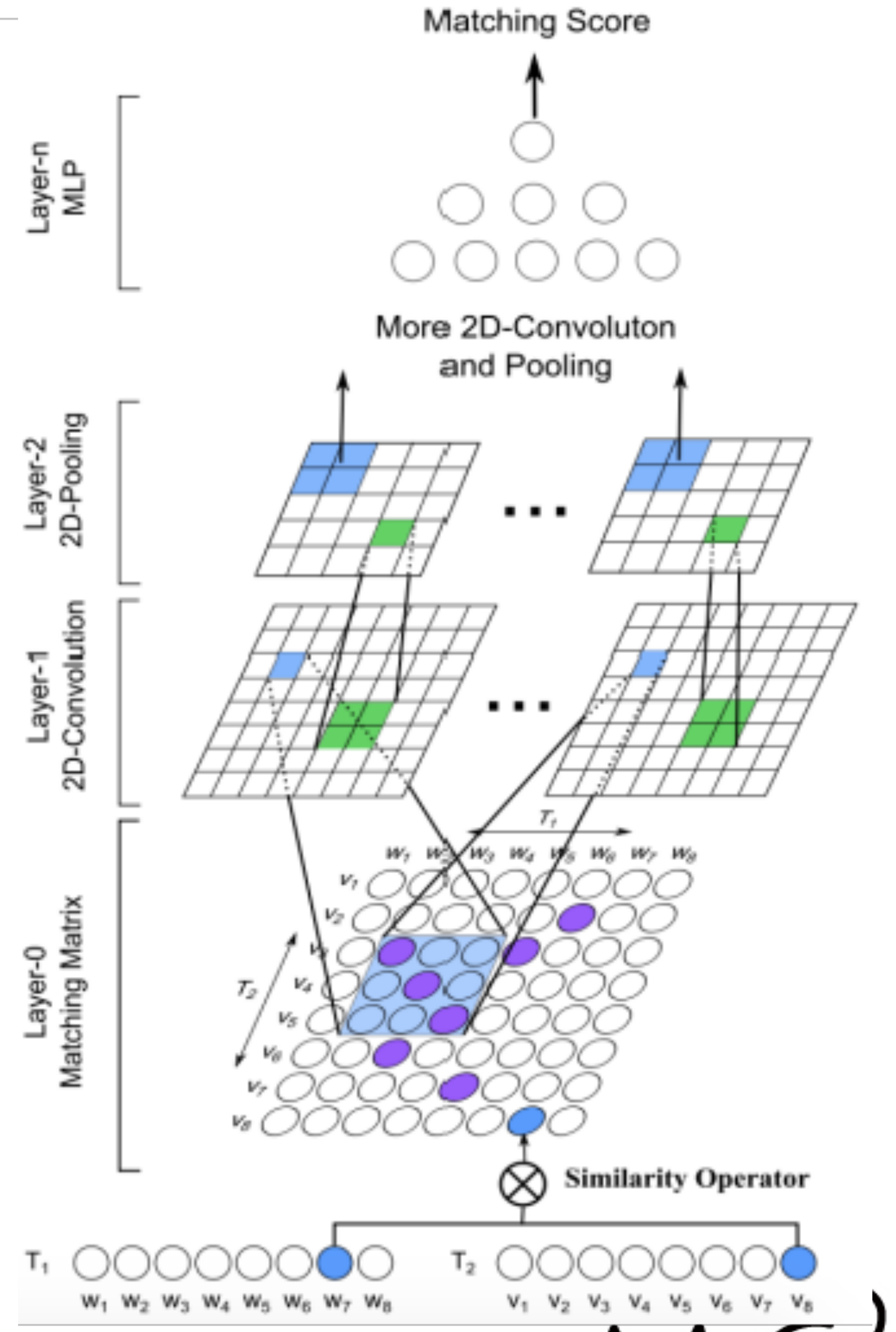
$$\text{score}(D,Q)=\sum_{i=1}^{n}\text{IDF}(q_i)\cdot\frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)},$$

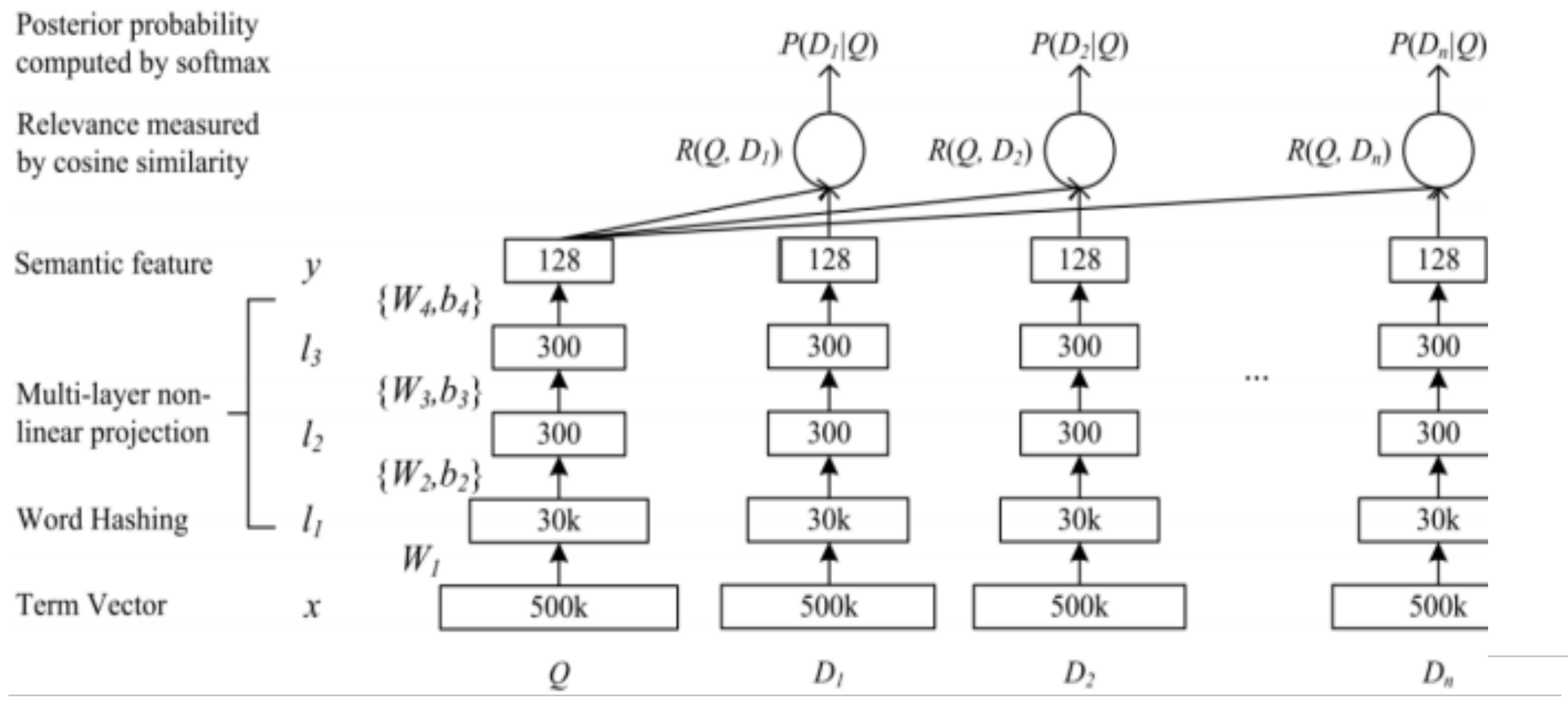
ID	Feature description	Category
1	$\sum_{q_i \in q \cap d} c(q_i,d)$ in body	Q-D
2	$\sum_{q_i \in q \cap d} c(q_i,d)$ in anchor	Q-D
3	$\sum_{q_i \in q \cap d} c(q_i,d)$ in title	Q-D
4	$\sum_{q_i \in q \cap d} c(q_i,d)$ in URL	Q-D
5	$\sum_{q_i \in q \cap d} c(q_i,d)$ in whole document	Q-D
6	$\sum_{q_i \in q} idf(q_i)$ in body	Q
7	$\sum_{q_i \in q} idf(q_i)$ in anchor	Q
8	$\sum_{q_i \in q} idf(q_i)$ in title	Q
9	$\sum_{q_i \in q} idf(q_i)$ in URL	Q
10	$\sum_{q_i \in q} idf(q_i)$ in whole document	Q
11	$\sum_{q_i \in q \cap d} c(q_i,d)\cdot idf(q_i)$ in body	Q-D
12	$\sum_{q_i \in q \cap d} c(q_i,d)\cdot idf(q_i)$ in anchor	Q-D
13	$\sum_{q_i \in q \cap d} c(q_i,d)\cdot idf(q_i)$ in title	Q-D
14	$\sum_{q_i \in q \cap d} c(q_i,d)\cdot idf(q_i)$ in URL	Q-D
15	$\sum_{q_i \in q \cap d} c(q_i,d)\cdot idf(q_i)$ in whole document	Q-D
16	$|d|$ of body	D
17	$|d|$ of anchor	D
18	$|d|$ of title	D
19	$|d|$ of URL	D
20	$|d|$ of whole document	D
21	BM25 of body	Q-D
22	BM25 of anchor	Q-D
23	BM25 of title	Q-D
24	BM25 of URL	Q-D
25	BM25 of whole document	Q-D
26	LMIR.ABS of body	Q-D
27	LMIR.ABS of anchor	Q-D
28	LMIR.ABS of title	Q-D
29	LMIR.ABS of URL	Q-D
30	LMIR.ABS of whole document	Q-D
31	LMIR.DIR of body	Q-D
32	LMIR.DIR of anchor	Q-D
33	LMIR.DIR of title	Q-D
34	LMIR.DIR of URL	Q-D
35	LMIR.DIR of whole document	Q-D
36	LMIR.JM of body	Q-D
37	LMIR.JM of anchor	Q-D
38	LMIR.JM of title	Q-D
39	LMIR.JM of URL	Q-D
40	LMIR.JM of whole document	Q-D
41	Sitemap based term propagation	Q-D
42	Sitemap based score propagation	Q-D
43	Hyperlink based score propagation: weighted in-link	Q-D
44	Hyperlink based score propagation: weighted out-link	Q-D
45	Hyperlink based score propagation: uniform out-link	Q-D
46	Hyperlink based propagation: weighted in-link	Q-D
47	Hyperlink based feature propagation: weighted out-link	Q-D
48	Hyperlink based feature propagation: uniform out-link	Q-D
49	HITS authority	Q-D
50	HITS hub	Q-D
51	PageRank	D
52	HostRank	D
53	Topical PageRank	Q-D
54	Topical HITS authority	Q-D
55	Topical HITS hub	Q-D
56	Inlink number	D
57	Outlink number	D
58	Number of slash in URL	D
59	Length of URL	D
60	Number of child page	D
61	BM25 of extracted title	Q-D
62	LMIR.ABS of extracted title	Q-D
63	LMIR.DIR of extracted title	Q-D
64	LMIR.JM of extracted title	Q-D

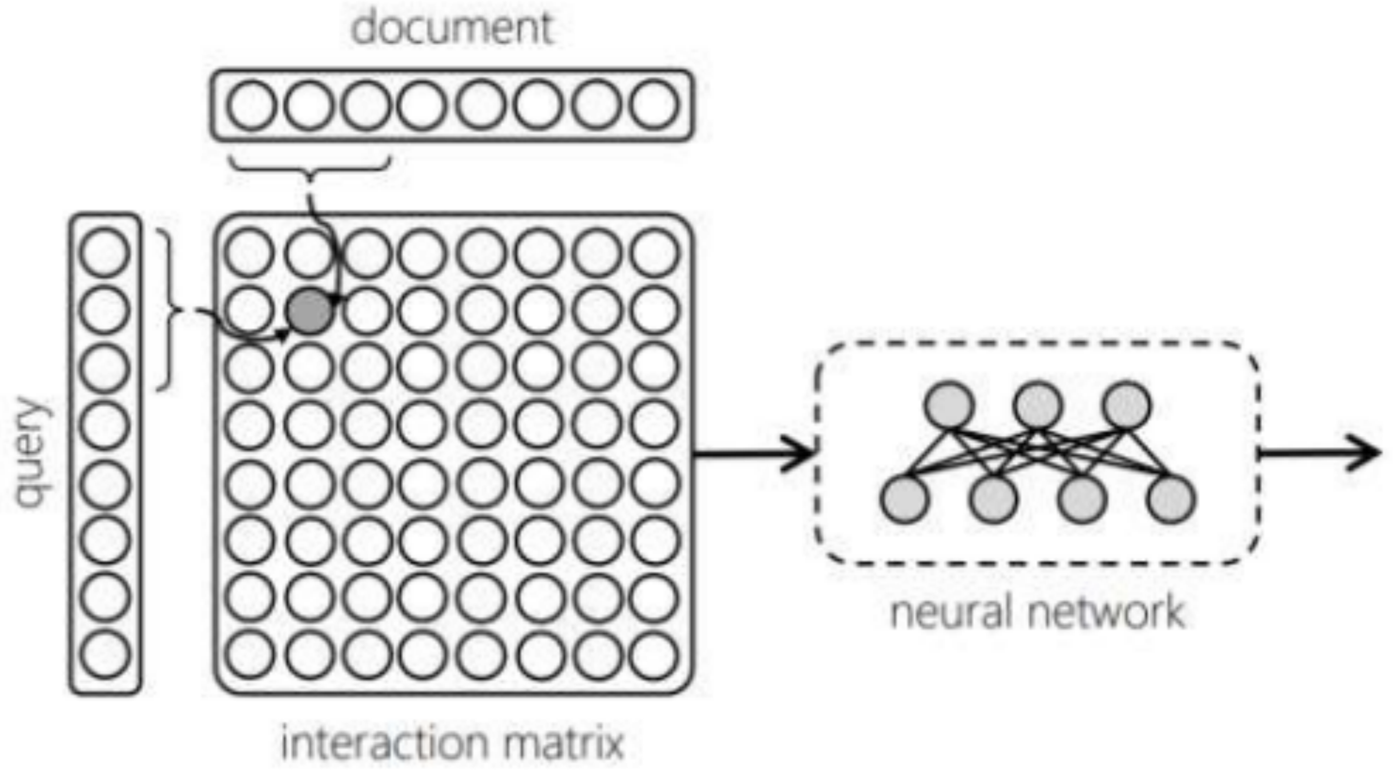
文本相关性的演进

- Before NN
- **Before BERT**
  - Embedding: word/char level
  - 表示模型: (C)DSSM
  - 交互模型: MatchPyramid, (Conv-)KNRM
- BERT

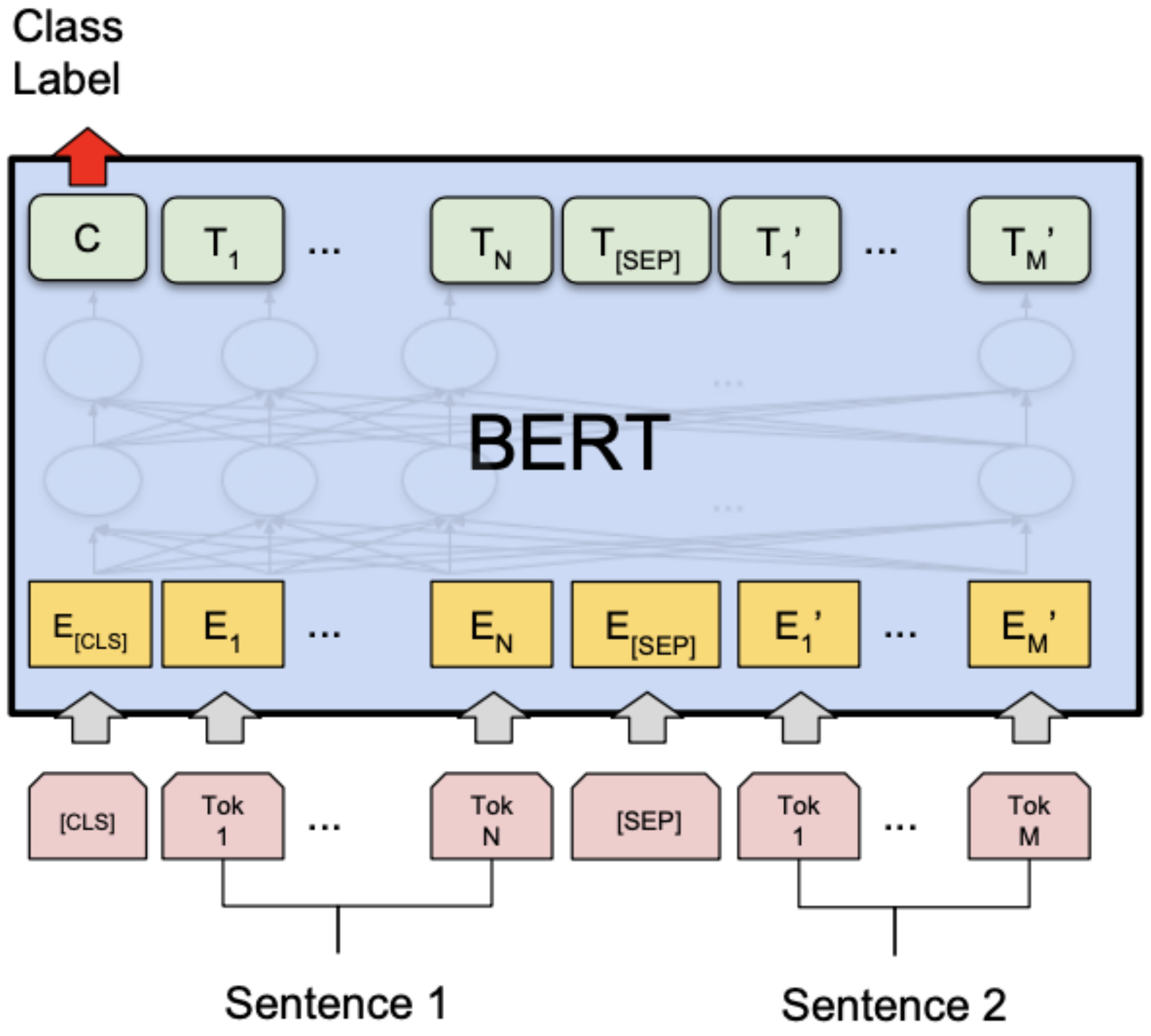


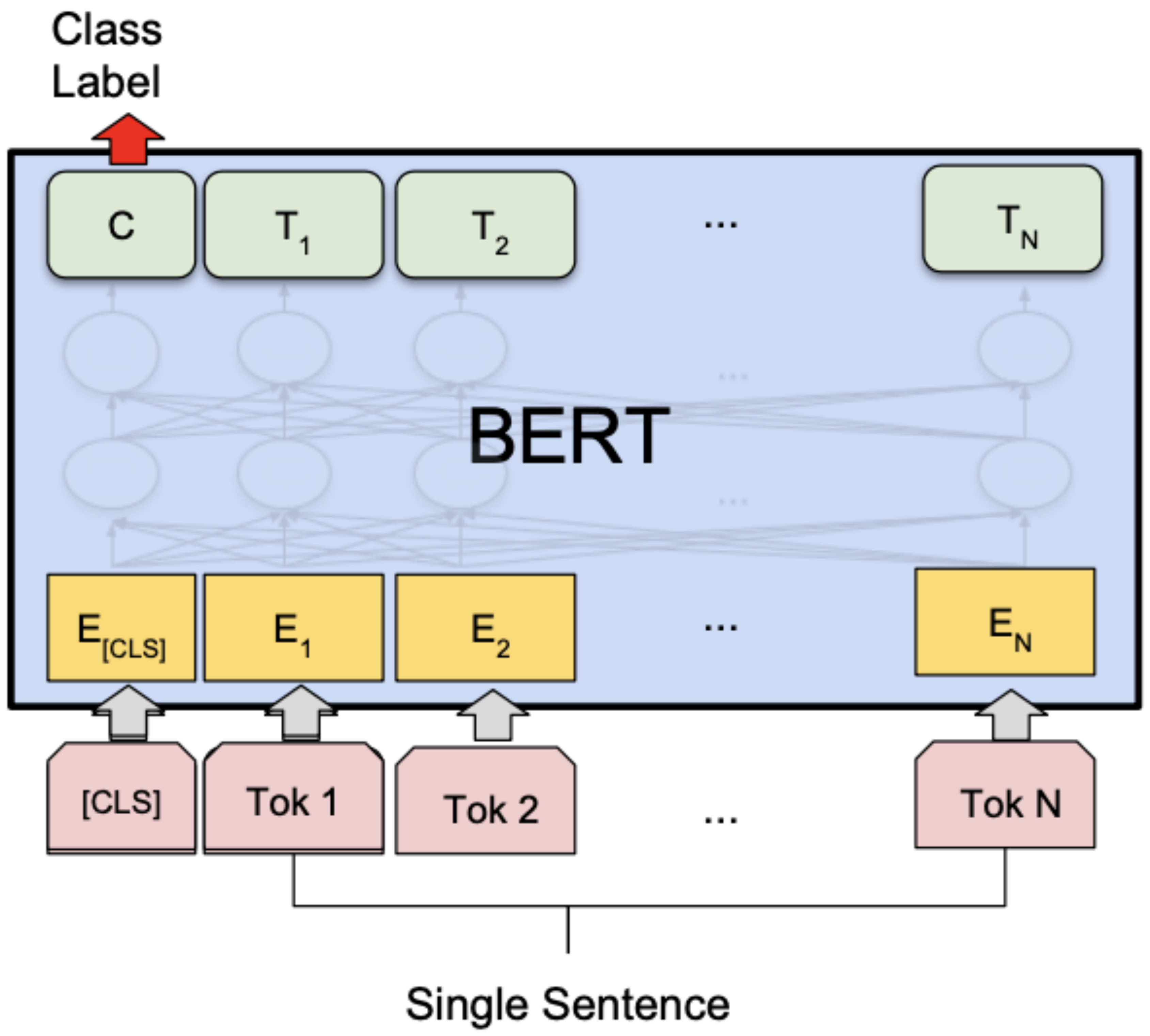






BERT相关性训练：交互模型 vs 表示模型

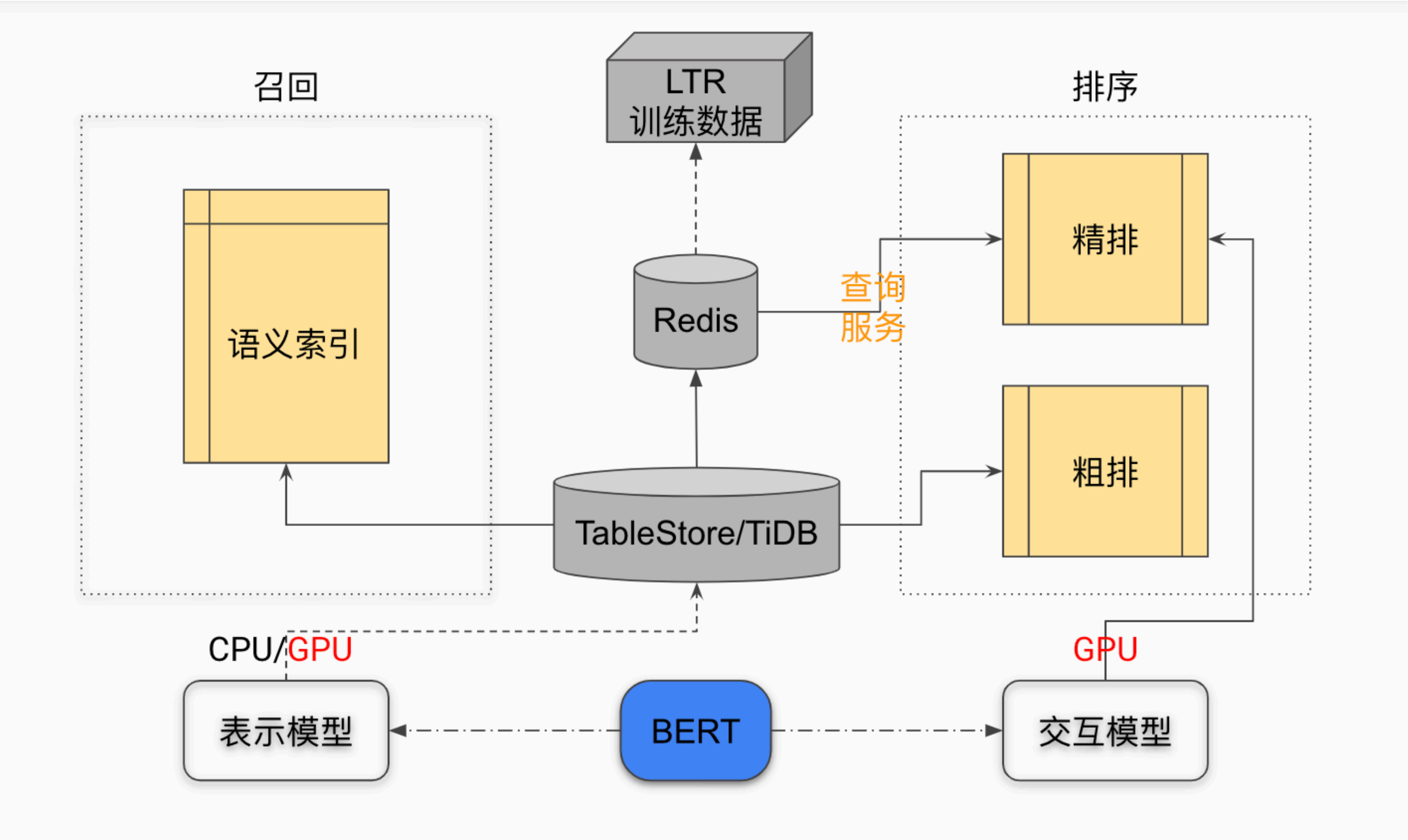




交互模型： Score(q, d) = Dense(Bert(q, d))

表示模型： Score(q, d) = Cosine(Bert(q), Bert(d))

搜索业务架构中的BERT



知乎

BERT表示模型语义召回

- 相关性任务 fine-tune
- BERT as Encoder
- Doc 向量构建语义索引(faiss)
- Query 向量召回

BERT带来的问题

- 交互模型服务 latency 过高
- 交互模型显存占用过大，精排排序 doc 量受限
- 向量查询服务带宽消耗过大、 latency 高

- 语义索引规模过大，latency 过高，离线构建慢
- 在线服务 GPU 机器需求大，预算压力
- 离线存储 TableStore/TiDB 资源消耗

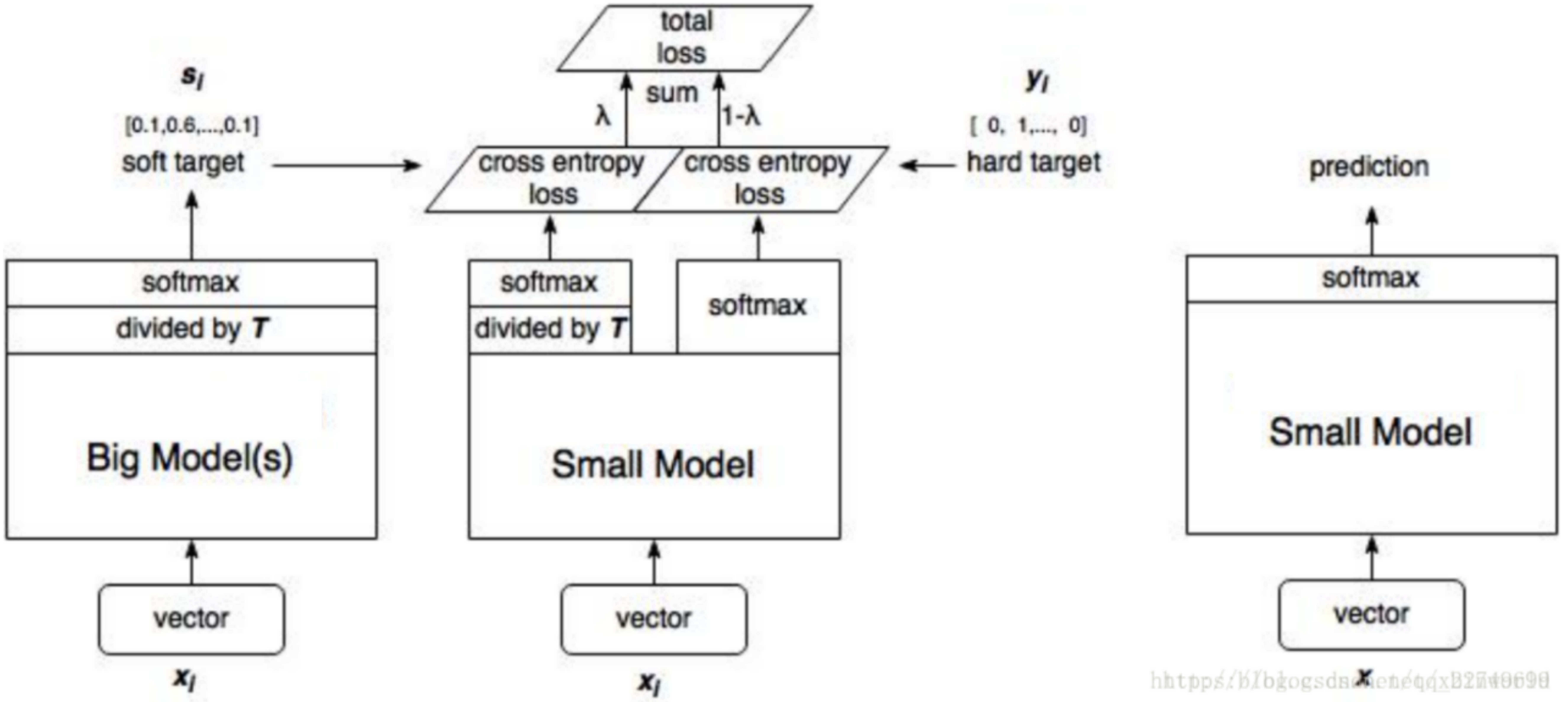
- 离线训练日志规模过大，日更 LTR 训练慢
- BERT 向量维度太大，无法引入二轮排序特征
- 无法建立全量正文语义索引/正文特征缺失
- ……

蒸馏前的尝试：

- cuBERT[1] (1.5x faster)
  - 混合精度 (Nvidia Tensor Core)
- Cache (2x faster)
- 减小 max_seq_length
- 直接训练小模型/减少层数fine-tune
- 直接对 BERT 做维度压缩
- 规则过滤部分 content 做语义召回/特征
- Poly-encoder [2]

1. https://github.com/zhihu/cuBERT
2. Humeau S, Shuster K, Lachaux M A, et al. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring[J]. arXiv preprint arXiv:1905.01969, 2019.

知识蒸馏



Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

知识蒸馏

- **Soft target *vs* Hard target**
  - Label Smoothing
  - Label Augmentation
- **Temperature**

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

-

Hard target

cow	Dog	cat	Car
0	1	0	0

Soft target

cow	Dog	cat	Car
0.001	0.9	0.009	1E-06

$\downarrow$ *Zi/T*

cow	Dog	cat	Car
0.05	0.6	0.035	0.005

Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

BERT蒸馏方案

基于任务分类:

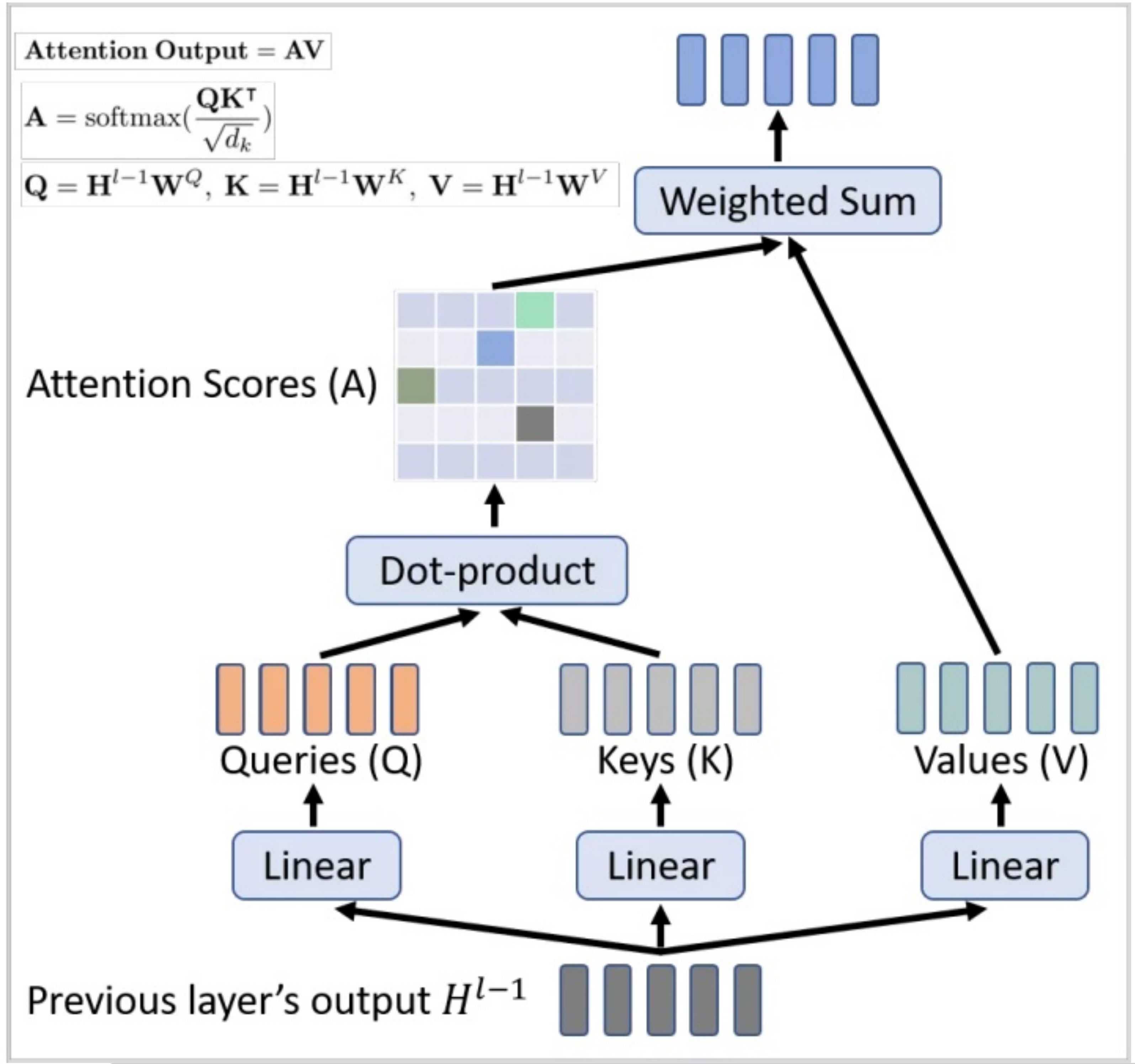
- 预训练任务蒸馏
  - DistilBERT
  - MiniLM
  - MobileBERT
- 下游任务蒸馏
  - Patient-KD
  - Bert to Simple NN
  - Pre-train Distill
  - Bert-of-theuseus
- 两段式
  - TinyBERT

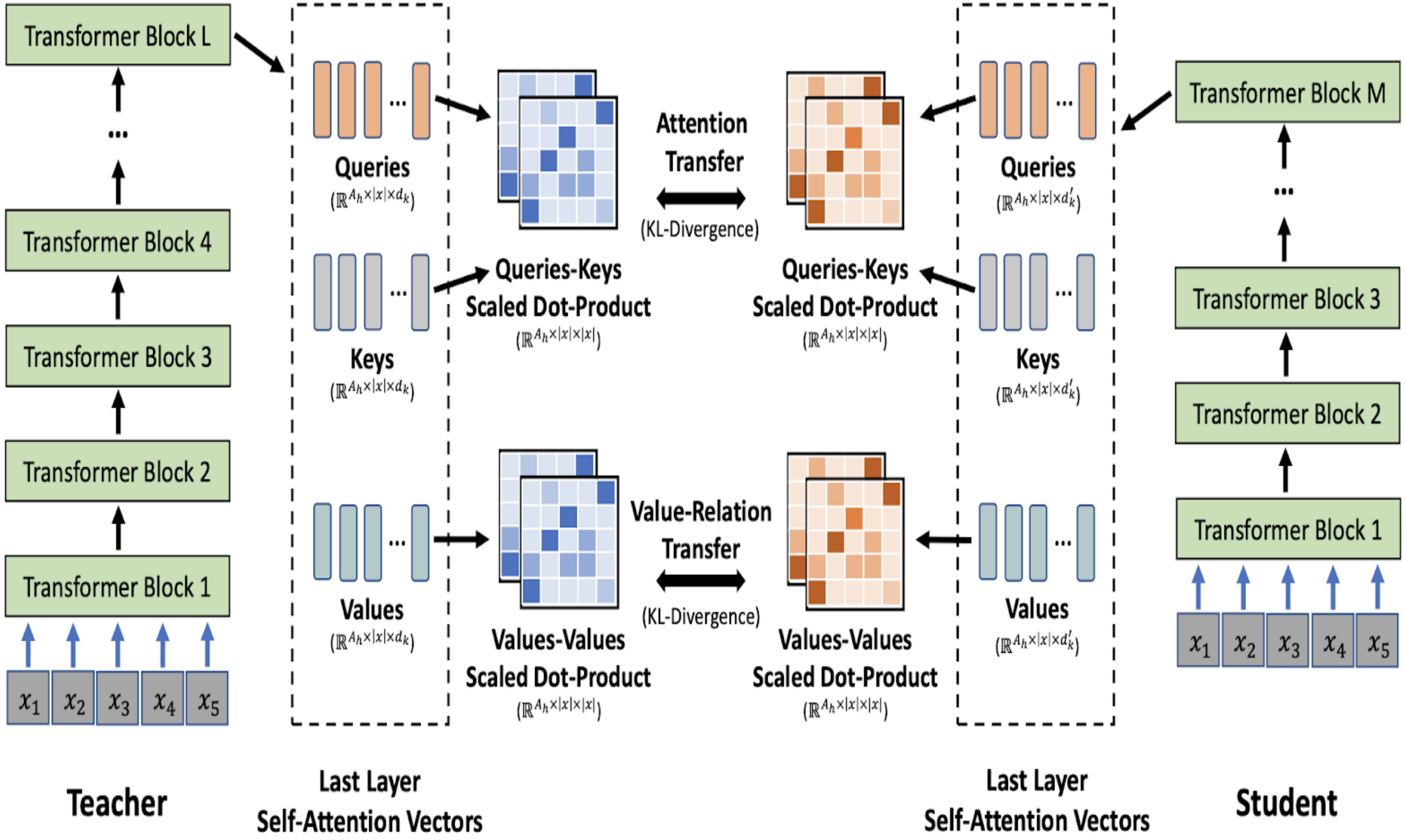
基于技巧分类:

- 迁移知识
  - Predict label
  - Attention score
  - Hidden output
- 模型结构
  - Width & Depth
  - Transformer block alter
  - Loss design
  - Layer initialization
  - Simple NN

蒸馏-MiniLM

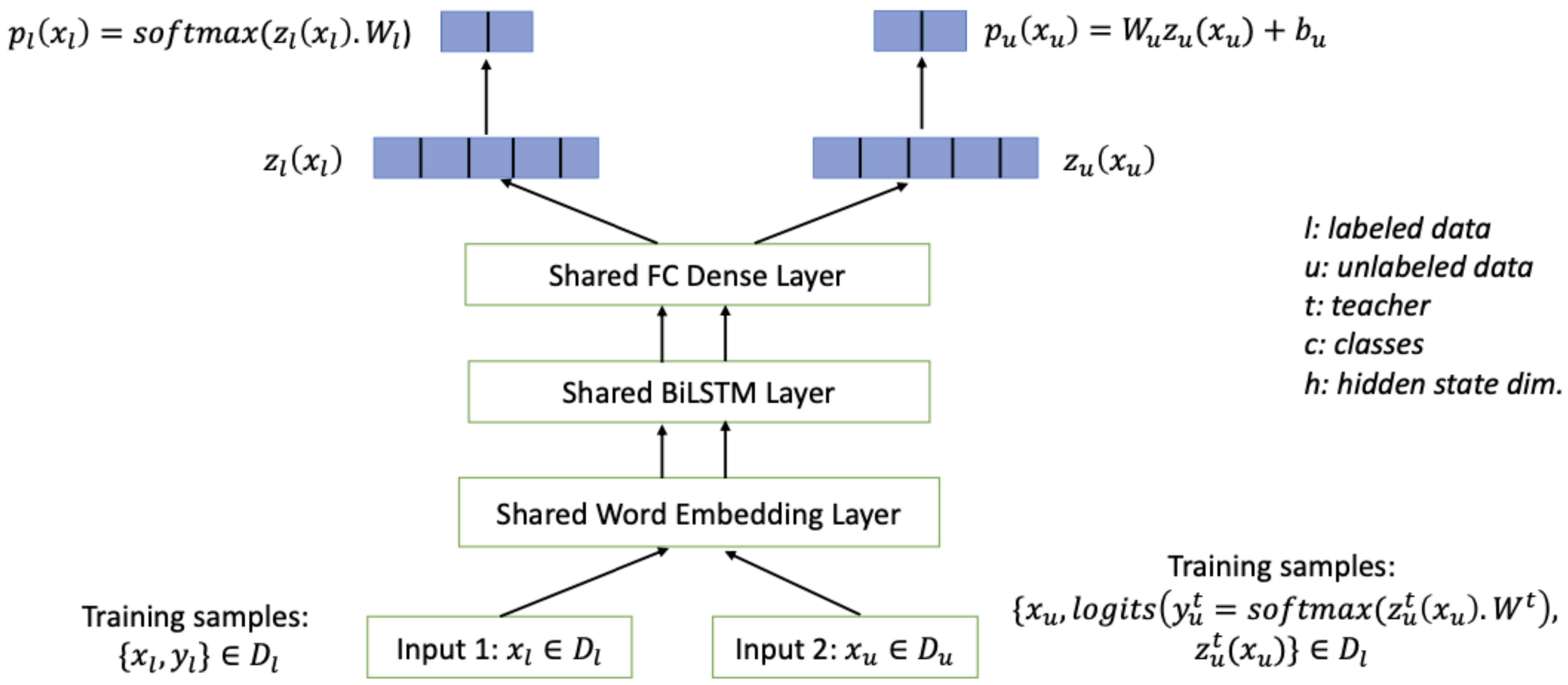
- 引入 Attention values 关系矩阵迁移
- Last layer Attention Distribution 迁移
- 使用 assistant 网络





Wang W, Wei F, Dong L, et al. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers[J]. arXiv preprint arXiv:2002.10957, 2020.

蒸馏-BERT to Simple NN



$$multitask\ loss = \sum_{x_l \in D_l} \sum_{c} \underbrace{y_{l,c} \log(p_{l,c}(x_l))}_{\text{Cross-entropy loss (labeled)}} + \sum_{x_u \in D_u} \left(\underbrace{\sum_{c} \left(logit(y_{u,c}^t) - p_{u,c}(x_u) \right)^2}_{\text{Mean-squared logit loss (unlabeled)}} + \sum_{h} \underbrace{\left(z_{u,h}^t(x_u) - z_{u,h}(x_u) \right)^2}_{\text{Mean-squared representation loss (unlabeled)}} \right) \Big/ |D_u|$$

Mukherjee S, Awadallah A H. Distilling transformers into simple neural networks with unlabeled transfer data[J]. arXiv preprint arXiv:1910.01769, 2019.

BERT蒸馏上的实践和收益

- 蒸馏目标：离线精度对比线上 BERT 无损
  - BERT base 直接蒸馏无法避免精度损失
  - 更大 teacher 模型选择（BERT-large/Robert-large/XLNET）

知乎

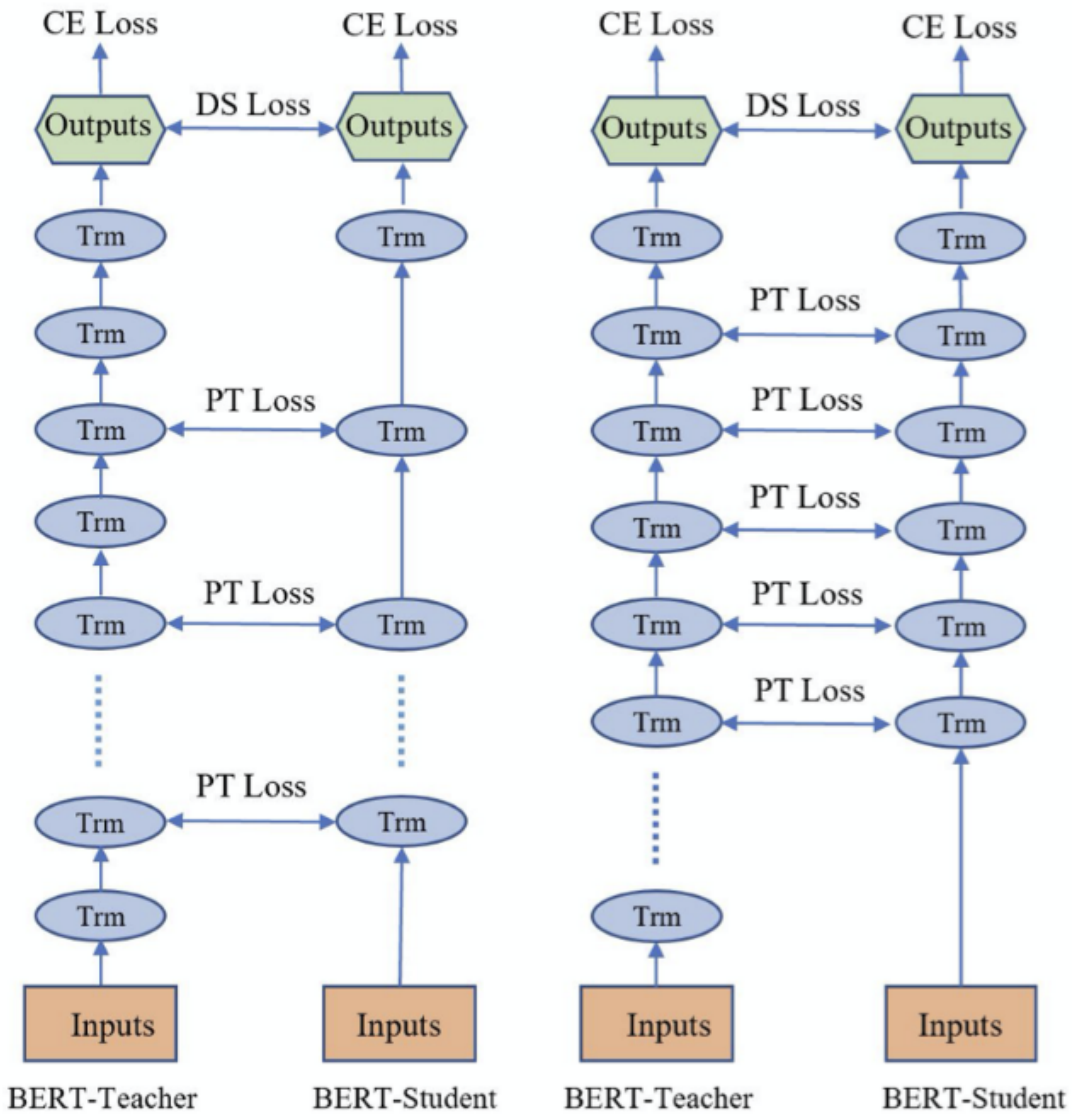
BERT交互模型蒸馏

- 基于 Patient-KD 方法，直接蒸馏 24 => 6（BERT base 隔层初始化）
- 实验助教 24 => 6 => 3 better than 24 => 3
  - （助教为 BERT base last 6 层初始化）
- 训练数据：标注数据 & 随机采样无标注数据
- 迁移知识: hidden layer logits + final logits
- Point-wise loss: RMSE/Cross entropy/Cosine

Teacher	**Student**	**nDCG@10**
Robert-large	-	0.914121
-	BERT-base	0.907743
-	BERT-6L	0.903115
BERT-base	BERT-6L	0.905856
Robert-large	BERT-6L	**0.911133**
Robert-large	BERT-3L	0.904888

蒸馏–PatientKD

- 策略:
  - 下游任务蒸馏
  - PKD-skip / PKD-last
  - BERT base 初始化
- Loss 设计:
  - LCE : student 的预测与真实标签的交叉熵
  - LDS : student 与 teacher 的预测的交叉熵
  - LPT : 隐藏层 normalized MSE



(a) PKD-*Skip* (b) PKD-*Last*

$$L_{PKD} = (1-\alpha)L^{s}_{CE} + \alpha L_{DS} + \beta L_{PT}$$

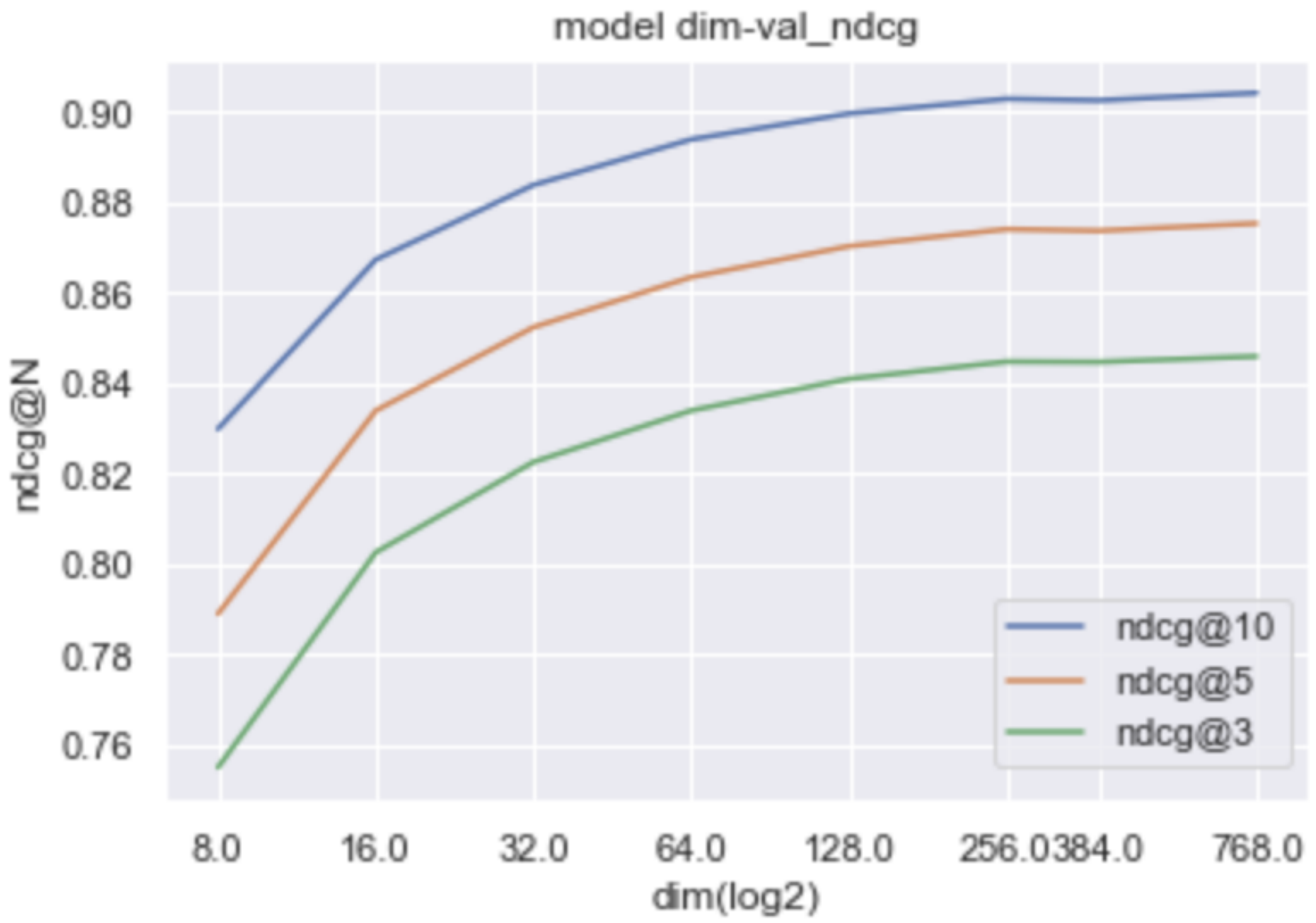
Sun S, Cheng Y, Gan Z, et al. Patient knowledge distillation for bert model compression[J]. arXiv preprint arXiv:1908.09355, 2019.

BERT表示模型蒸馏

dim	val_acc	ndcg@5
base	0.8516	0.8651
64	0.8749	0.8635
128	0.8910	**0.8703**
768	0.9011	0.8753

- 蒸馏的同时维度压缩
- 交互模型作为 teacher 蒸馏
- Pairwise loss: teacher 差值拟合

$$\sum_{ij}(P(S_i - S_j) - P(T_i - T_j))$$



维度压缩指标趋势图

蒸馏的收益

Online

交互模型

- 排序相关性特征 P95 减少为 1/2，搜索入口下降 40ms
- 服务 RTX 2080Ti 8 卡 GPU 机器数减少一半

表示模型

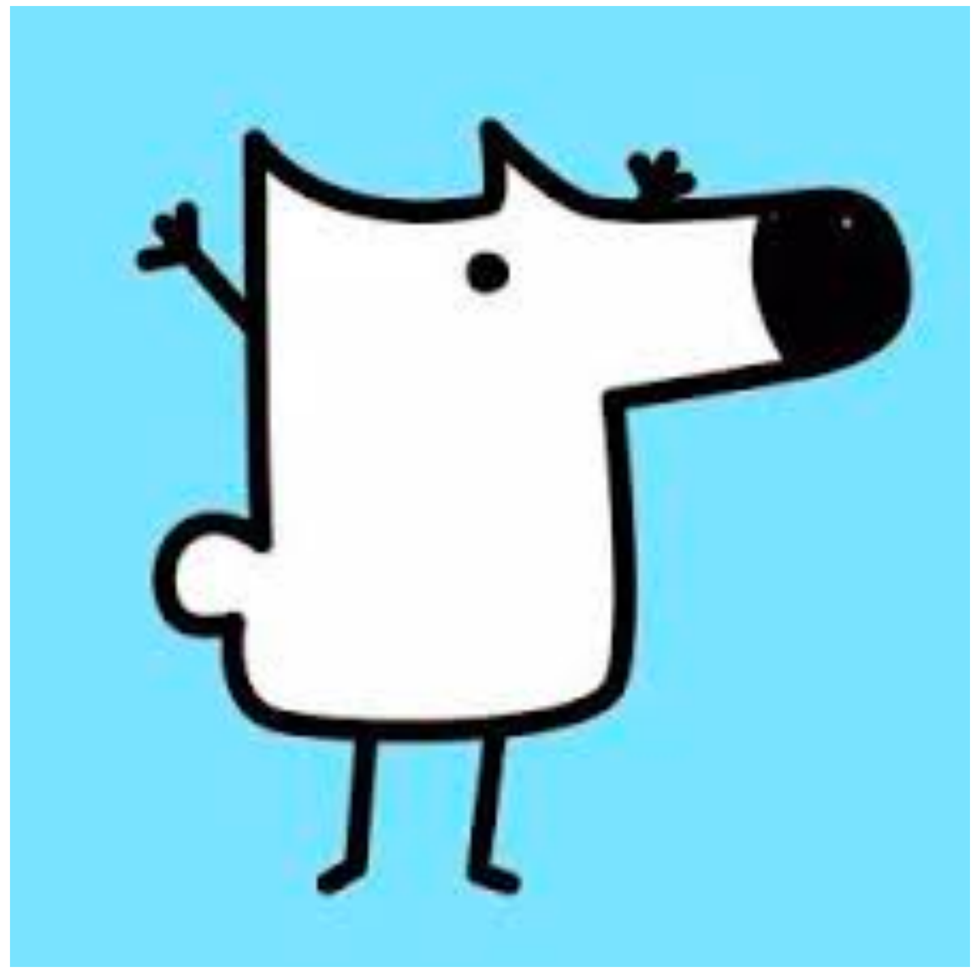
- 语义索引存储规模 title 减少为 1/4、content 较少为 1/6
- 语义索引召回 P99 title 减少为 1/3，content 减少为 1/2
- 向量查询服务 P95 约降为 1/4, Redis 存储约较少为 1/5
- 扩充全量 content 数据语义索引和特征服务，次日留存 +0.17%
- con完全替换掉content KNRM/Pyramid，节省 GPU 资源

Offline

表示模型

- 日常语义索引构建时间减少为 1/4
- TableStore/TiDB 存储变为原来的 1/6
- LTR 训练数据减少为原来的 1/4
- LTR 日常训练时间 10h => 5h
- 粗排模型引入 32d 向量特征，提高粗排精度
- 精排引入 BERT 向量 End2End 训练，满意点击比 +0.16%

THANKS



持续招人中~

大家一起来构建更好的知乎搜索~

欢迎投递: fangkuan@zhihu.com