7.4. Pre-training Tasks for Embedding-based Large-scale Retrieval

논문 : https://arxiv.org/abs/2002.03932

참고자료

https://blog.pingpong.us/ml-seminar-season-4/

논문이 주장하는 문제

- 정보 검색은 많이 호출되고 빨라야됨
- 사용자 문장과 모든 문서를 pairwise해서 BERT에 연산시키면 너무 오래걸림
- 연산이 빠르고 학습이 필요없는 BM-25(token matching + TF-IDF weight) 방식을 선호해옴

논문이 주장하는 것

- Pre-train 과정에서 Paragraph-level의 task를 추가하자!
    - 3가지 Task를 제안
    - Inverse Cloze Task(ICT), Body First Selection(BFS), Wiki Link Prediction(WLP)
- Contrastive Learning 하자!
- 문서에 대한 정보를 미리 연산해놓은 뒤 사용자 문장만 BERT에 연산시키고 가장 유사한 문서 후보군들을 리턴하는 방식을 제안

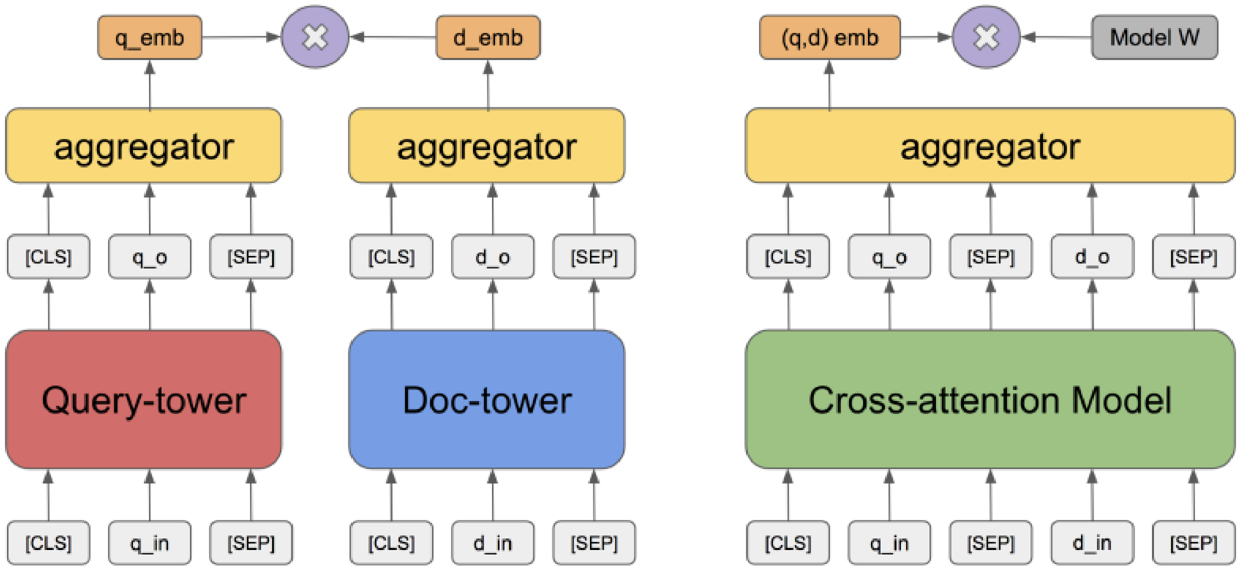

Figure 1: Difference between two-tower models and cross-attention models. Following previous works, we consider [CLS] embedding and average pooling as the aggregator's output for the two-tower Transformer model and the two-tower MLP model, respectively.

Contrastive Learning는 이전 논문과 동일

- Query와 전체 Document를 넣어서 Softmax 값 연산
- Query 별로 가능한 Document에 대한 Softmax값의 총합을 Loss로 사용

$$p_\theta(\boldsymbol{d}|\boldsymbol{q}) = \frac{\exp\left(f_\theta(\boldsymbol{q},\boldsymbol{d})\right)}{\sum_{\boldsymbol{d}'\in\mathcal{D}}\exp\left(f_\theta(\boldsymbol{q},\boldsymbol{d}')\right)},$$

Pre-train에서의 사용할 데이터 구조와 Loss - 설명이 빈약해서 최대한 이해한 방향으로 작성

- Inverse Cloze Task

    - 쿼리 : 랜덤한 문장
    - 문서 : 쿼리 문장들의 집합 (문서 내의 전체 문장)
    - 문제 : 랜덤한 문장의 앞뒤 문장을 맞추는 softmax 사용
- Body First Selection
    - 쿼리 : 위키피디아 문서중 첫번째 섹션(Summary)에서 랜덤한 문장
    - 문서 : Inverse Cloze Task와 동일
    - 문제 : 랜덤한 문장이 위키피디아 첫번째 섹션(대부분 Summary Section)이 어느 부분을 뜻하는지 맞추기
- Wiki Link Prediction
    - 쿼리 : 위키피디아 페이지의 첫번째 섹션의 랜덤문장
    - 문서 : 쿼리 문장 페이지에 있는 하이퍼링크로 연결된 다른 페이지
    - 문제 : 쿼리로 다른 페이지와의 연결성을 맞추기

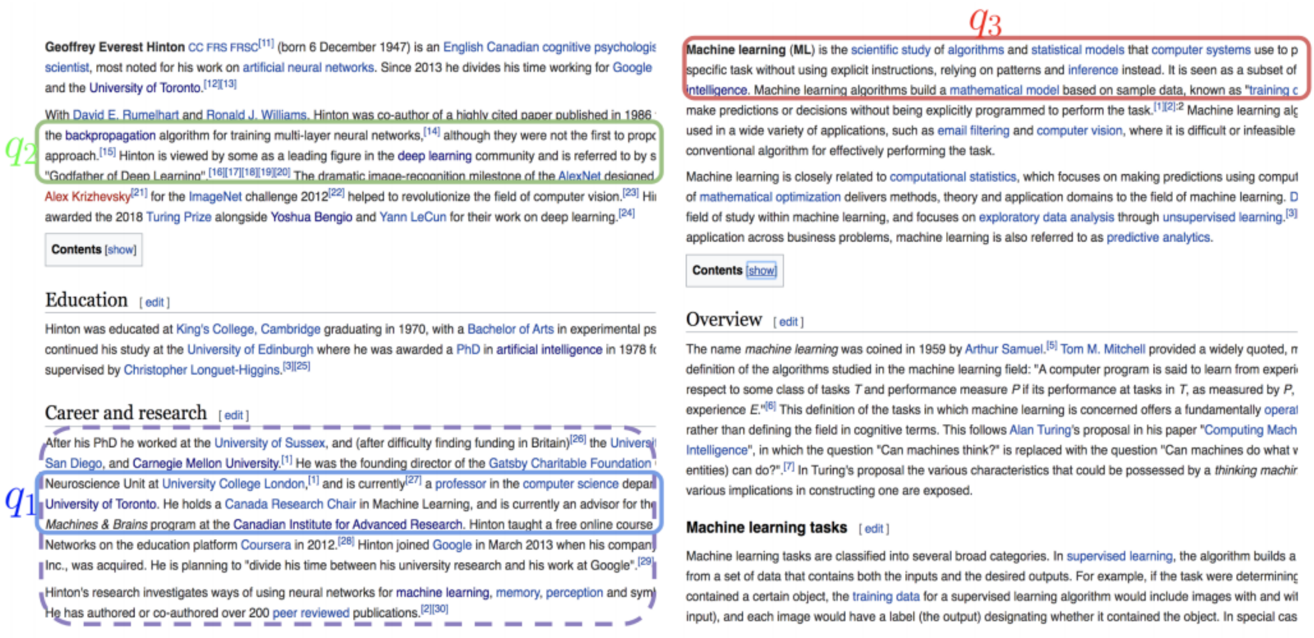

Figure 2: An illustrative example of the three pre-training tasks where each query $\boldsymbol{q}$ is highlighted in different colors. All queries are paired with the same text block $\boldsymbol{d}$. Concretely, $(\boldsymbol{q}_1,\boldsymbol{d})$ of ICT is defined locally within a paragraph; $(\boldsymbol{q}_2,\boldsymbol{d})$ of BFS is defined globally within an article; $(\boldsymbol{q}_3,\boldsymbol{d})$ of WLP is defined distantly across two related articles hyper-linked by the Wikipedia entity.

Inverse Cloze Task (ICT) Given a passage $\boldsymbol{p}$ consisting of n sentences, $\boldsymbol{p} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, the query $\boldsymbol{q}$ is a sentence randomly drawn from the passage, $\boldsymbol{q} = \boldsymbol{s}_i, i \sim [1, n]$, and the document $\boldsymbol{d}$ is the rest of sentences, $\boldsymbol{d} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{i-1}, \boldsymbol{s}_{i+1}, \ldots, \boldsymbol{s}_n\}$. See $(\boldsymbol{q}_1,\boldsymbol{d})$ in Figure 2 as an example. This task captures the semantic context of a sentence and was originally proposed by Lee et al. (2019).

Body First Selection (BFS) We propose BFS to capture semantic relationship outside of the local paragraph. Here, the query $\boldsymbol{q}_2$ is a random sentence in the first section of a Wikipedia page, and the document $\boldsymbol{d}$ is a random passage from the same page (Figure 2). Since the first section of a Wikipedia article is often the description or summary of the whole page, we expect it to contain information central to the topic.

Wiki Link Prediction (WLP) We propose WLP to capture inter-page semantic relation. The query $\boldsymbol{q}_3$ is a random sentence in the first section of a Wikipedia page, and the document $\boldsymbol{d}$ is a passage from another page where there is a hyperlink link to the page of $\boldsymbol{q}_3$ (Figure 2). Intuitively, a hyperlink link indicates relationship between the two Wikipedia pages. Again, we take a sentence from the first section because it is often the description or summary of the topic.

Masked LM (MLM) In addition to the above tasks, we also consider the classic masked language model (MLM) pre-training task as a baseline: predict the randomly masked tokens in a sentence. MLM is the primary pre-training task used in BERT (Devlin et al., 2019).

Pre-training tasks	#tokens	#pairs	avg. #query tokens	#doc tokens
ICT	11.2B	50.2M	30.41	193.89
BFS	3.3B	17.5M	28.02	160.46
WLP	2.7B	24.9M	29.42	82.14

Table 1: Data statistics of three pre-training tasks. #query tokens represent average number of tokens per query, and #doc tokens represent average number of tokens per passage.

ReQA Dataset	#query	#candidate	#tuples	#query tokens	#doc tokens
SQuAD	97,888	101,951	99,024	11.55	291.35
Natural Questions	74,097	239,008	74,097	9.29	352.67

Table 2: Data statistics of ReQA benchmark. candidate represents all (sentence, passage) pairs.

train/test ratio	Encoder	Pre-training task	R@1	R@5	R@10	R@50	R@100
1%/99%	BM-25	No Pretraining	**41.86**	58.00	63.64	74.15	77.91
	BoW-MLP	No Pretraining	0.14	0.35	0.49	1.13	1.72
	BoW-MLP	ICT+BFS+WLP	22.55	41.03	49.93	69.70	77.01
	Transformer	No Pretraining	0.02	0.06	0.08	0.31	0.54
	Transformer	MLM	0.18	0.51	0.82	2.46	3.93
	Transformer	ICT+BFS+WLP	37.43	**61.48**	**70.18**	**85.37**	**89.85**
5%/95%	BM-25	No Pretraining	41.87	57.98	63.63	74.17	77.91
	BoW-MLP	No Pretraining	1.13	2.68	3.62	7.16	9.55
	BoW-MLP	ICT+BFS+WLP	26.23	46.49	55.68	75.28	81.89
	Transformer	No Pretraining	0.17	0.36	0.54	1.43	2.17
	Transformer	MLM	1.19	3.59	5.40	12.52	17.41
	Transformer	ICT+BFS+WLP	**45.90**	**70.89**	**78.47**	**90.49**	**93.64**
80%/20%	BM-25	No Pretraining	41.77	57.95	63.55	73.94	77.49
	BoW-MLP	No Pretraining	19.65	36.31	44.19	62.40	69.19
	BoW-MLP	ICT+BFS+WLP	32.24	55.26	65.49	83.37	88.50
	Transformer	No Pretraining	12.32	26.88	34.46	53.74	61.53
	Transformer	MLM	27.34	49.59	58.17	74.89	80.33
	Transformer	ICT+BFS+WLP	**58.35**	**82.76**	**88.44**	**95.87**	**97.49**

Table 3: Recall@k on SQuAD. Numbers are in percentage (%).

train/test ratio	Encoder	Pre-training task	R@1	R@5	R@10	R@50	R@100
1%/99%	BM-25	No Pretraining	4.99	11.91	15.41	24.00	27.97
	BoW-MLP	No Pretraining	0.28	0.80	1.08	2.02	2.66
	BoW-MLP	ICT+BFS+WLP	9.22	24.98	33.36	53.67	61.30
	Transformer	No Pretraining	0.07	0.19	0.28	0.56	0.85
	Transformer	MLM	0.18	0.56	0.81	1.95	2.98
	Transformer	ICT+BFS+WLP	**17.31**	**43.62**	**55.00**	**76.59**	**82.84**
5%/95%	BM-25	No Pretraining	5.03	11.96	15.47	24.04	28.00
	BoW-MLP	No Pretraining	1.36	3.77	4.98	8.56	10.77
	BoW-MLP	ICT+BFS+WLP	11.40	30.64	40.63	62.95	70.85
	Transformer	No Pretraining	0.37	1.07	1.40	2.73	3.82
	Transformer	MLM	1.10	3.42	4.89	10.49	14.37
	Transformer	ICT+BFS+WLP	**21.46**	**51.03**	**62.99**	**83.04**	**88.05**
80%/20%	BM-25	No Pretraining	4.93	11.52	14.96	23.64	27.77
	BoW-MLP	No Pretraining	9.78	26.76	34.16	50.34	56.44
	BoW-MLP	ICT+BFS+WLP	13.58	37.78	50.40	76.11	82.98
	Transformer	No Pretraining	7.49	20.11	25.40	38.26	43.75
	Transformer	MLM	16.74	40.48	49.53	67.91	73.91
	Transformer	ICT+BFS+WLP	**30.27**	**63.97**	**75.85**	**91.84**	**94.60**

Table 4: Recall@k on Natural Questions. Numbers are in percentage (%).

Index	Ablation Configuration			R@100 on different train/test ratio			
	#layer	Pre-training task	emb-dim	1%	5%	10%	80%
1	4	ICT	128	77.13	82.03	84.22	91.88
2	4	BFS	128	72.99	78.34	80.47	89.82
3	4	WLP	128	56.94	68.08	72.51	86.15
4	12	No Pretraining	128	0.72	3.88	6.94	38.94
5	12	MLM	128	2.99	12.21	22.97	71.12
6	12	ICT	128	79.80	85.97	88.13	93.91
7	12	ICT+BFS+WLP	128	81.31	87.08	89.06	94.37
8	12	ICT+BFS+WLP	256	81.48	87.74	89.54	94.73
9	12	ICT+BFS+WLP	512	82.84	88.05	90.03	94.60

Table 5: Ablation study on Natural Questions based on Recall@100. Index 9 represents the proposed method in Table 4.

train/test ratio	Pre-training task	R@1	R@5	R@10	R@50	R@100
1%/99%	BM-25	3.70	9.58	12.69	20.27	23.83
	ICT	**14.18**	37.36	48.08	69.23	76.01
	ICT+BFS+WLP	13.19	**37.61**	**48.77**	**70.43**	**77.20**
5%/95%	BM-25	3.21	8.62	11.50	18.59	21.78
	ICT	**17.94**	45.65	57.11	76.87	82.60
	ICT+BFS+WLP	17.62	**45.92**	**57.75**	**78.14**	**83.78**
80%/20%	BM-25	3.12	8.45	11.18	18.05	21.30
	ICT	24.89	57.89	69.86	87.67	91.29
	ICT+BFS+WLP	**25.41**	**59.36**	**71.12**	**88.25**	**91.71**

Table 6: Open-domain retrieval results of Natural Questions dataset, where existing candidates are augmented with additional 1M retrieval candidates (i.e., 1M of $(\boldsymbol{s}, \boldsymbol{p})$ candidate pairs) extracted from open-domain Wikipedia articles.