Bing Sponsored Search Retriever

Review

윤정훈

September 18, 2023

Paper

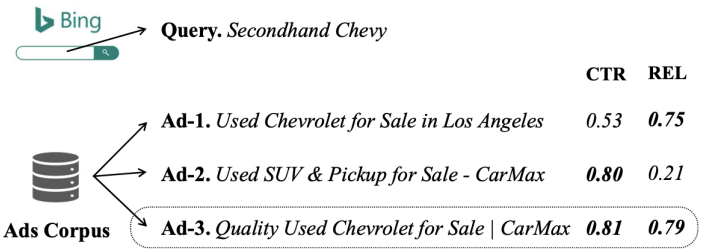
Uni-Retriever: Towards Learning The Unified Embedding Based Retriever in Bing Sponsored Search [J. Zhang et al.]

- KDD 2022
- J. Zhang et al. (Microsoft Research Asia)

Propose

검색광고에서 달성하려고 하는 2가지 목적을 동시에 만족하기 위한 framework 제안

- **High-relevance** ads : 사용자의 검색 의도(search intent)를 만족시키는 광고를 검색
- **High-CTR** ads : 사용자(overall user)의 클릭을 극대화 시키는 광고를 검색



Propose

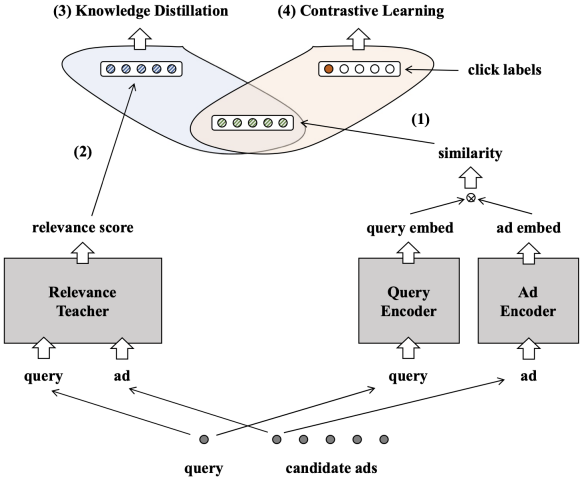
EBR in Sponsored Search

- 2가지 목적을 동시에 달성하기 위하여 multi-objective learning process 수행
  - Knowledge distillation : high-relevance
  - Contrastive learning : high-CTR

Serving EBR At Scale

- DiskANN 사용
  - Time consumption, Recall rate : the proximity-graph based algorithms
  - Memory Usage : the vector-quantization based algorithms

Architecture of Uni-Retriever



1. Query, ad는 각각 latent embedding vector로 인코딩되고, similarity는 inner product로 계산.
2. Query와 ad의 semantic closeness(relevance score)는 relevance teacher model에 의해 계산
3. Knowledge Distillation : Query-ad similarity와 relevance score의 차이 minimize
4. Contrastive learning : 클릭된 광고의 차별성을 구하기 위하여 (구별하기 위하여) 학습

Uni-Retriever

사용자 input query(q)에 대하여, 전체 광고 corpus에서(A), **semantically 가깝고 click을 받을 것 같은** ads(A_q)를 검색해주는 것

Objective function

$$\max.\ \Sigma_{A_q}\mathrm{CTR}(q, A_q) : s.t.\ \mathrm{REL}(q, a) \geq \epsilon, \forall a \in A_q$$

ϵ : relevance threshold.

Relaxation of objective function

$$\max.\ \Sigma_{A_q}\mathrm{CTR}(q, A_q) + \lambda \times \mathrm{REL}(q, a)$$

λ : CTR과 relevance score trade-off를 조절하는 positive value.

A_q : MIPS(maximum inner product search) 기반으로 검색되므로, query의 top-K embedding similarity ads.

Knowledge Distillation

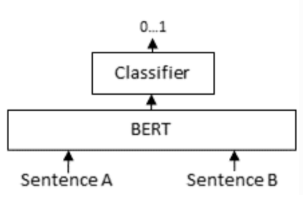
Relevance Teacher Model의 knowledge distillation를 이용하여 query, ad embedding을 학습

- BERT 기반의 binary classification model(BERT^{Tch})이며, 주어진 query와 ad의 semantically closeness를 예측한다.
- bi-encoder 구조가 아닌 **cross-encoder** 구조를 사용한다.
- $\text{Rel}_{q,a} = \sigma(W^T \text{BERT}^{Tch}([\text{CLS}, Query, \text{SEP}, Ad]))$
- Final layer의 CLS token에 대응하는 hidden state를 BERT output으로 사용
- $W \in R^{d \times 1}$은 linear projection
- $\sigma(\cdot)$은 sigmoid activation 함수
- Pretrained based on **manually labeled data from human experts**

Knowledge distillation

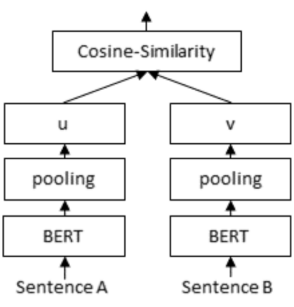
Cross-encoder

1. Sentence를 입력받아서 0과 1 사이의 output을 출력한다.
2. **Cross-Encoder achieve better performances than Bi-Encoders. However, for many application they are not pratical.**



Bi-encoder

1. Sentence를 입력받아서 embedding vector를 출력한다.
2. Embedding vector를 생성하므로, 더 효율적으로 서비스에 적용할 수 있음.



Knowledge distillation

Uni-Retriever

- Query, ad encoder는 같은 backbone BERT^{Uni} 공유
- Relevacne Teacher Model이 relevance score $\text{Rel}_{q,a}$를 예측하면, Uni-Retriever BERT^{Uni}는 teacher의 예측을 imitate

KD를 위한 Loss

$$\text{min. } \Sigma_q \Sigma_a || \text{Rel}_{q,a} - < \text{BERT}^{Uni}(q), \text{BERT}^{Uni}(a) > ||$$

$< \cdot >$은 inner product 연산.

Contrastive Learning

Contrastive Learning

- Corpus 내 clicked ads 즉, ground truth를 구별할 수 있도록 embedding을 학습
- Objective function은 InfoNCE loss를 사용함 [Mnih, 2012]

InfoNCE loss

$$\text{max. } \Sigma_q \frac{\exp(< \text{BERT}^{Uni}(q), \text{BERT}^{Uni}(a_q^+) >)}{\Sigma_{a^- \in N_q} \exp(< \text{BERT}^{Uni}(q), \text{BERT}^{Uni}(a^-) >)}$$

, where a_q^+ denotes the ground-truth, N_q are the negative samples to the query.

- Contrastive learning은 2가지 요소에 크게 영향을 받음.
  - **Negative samples의 scale**
  - **Negative samples의 hardness**

Contrastive Learning

Mini-batch B_q

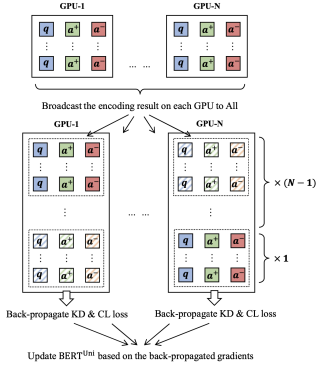
- $q \in B_q$의 ground truth ads가 있고, $q \neq q^{'} \in B_q$ 인 $q^{'}$ 의 ground truth ads가 q 의 negative samples로 사용된다.
- 최대 $|B_q| - 1$ 개의 negative samples이 존재 가능하다.
- q의 negative samples : $N_q = \{a^{+}_{q^{'} \neq q}\}_B$

Cross-device Negative Sampling

- Multiple distributed GPU devices에서 적용 가능하다.
- Mini-batches의 collection : $B = \{B_1, ..., B_N\}$
- $q^{'}(\neq q)$의 ground truth ads가 B_q 뿐만이 아닌 B에 존재한다.
- q의 negative samples : $N_q = \{a^{+}_{q^{'} \neq q}\}$

Contrastive Learning

- 다른 device에 있는 embedding이 detached 되면 미분 불가능하므로, CDNS는 좋은 성능을 낼 수 없다.
- Gradient compensation operation : cross-device embedding을 "virtually differentiable"한 상태로 만들어준다.



1. Query, ad embedding은 모든 device로 broadcasting 된다.
2. 각 device B_i에서, InfoNCE loss가 B_i 뿐이 아닌 broadcasting 된 query에 대해서 계산된다. 이 때 negative samples 또한 broadcasting 된 ads에서 추출된다.
3. 모든 device에서 back-propagation 수행 및 모델 업데이트

Contrastive Learning

ANN based hard negatives

- Hard negative를 얻기 위한 hueristic 기법이며 contrastive 학습시 효과적인 것으로 알려져 있음
- a : $\text{BERT}^{Uni} \in \text{ANN}(\text{BERT}^{Uni}(q))$에서 random하게 추출

Relevance Filtered Hard Negatives

- ANN 검색결과에서 TopK는 제거하여, low relevance ads에서 hard negative samples를 추출하는 기법

Disentangle and Multi-objective Learning

- Uni-Retriever는 knowledge distillation과 contrastive learning이 함께 학습.
- 학습이 진행될때, backbone은 BERT^{Uni}를 사용하지만, 각 목적별로 다른 pooling head를 각각 사용한다.
  - $W_{\text{rel}}\text{BERT}^{Uni}(\cdot)$: knowledge distillation의 output embedding vector
  - $W_{\text{ctr}}\text{BERT}^{Uni}(\cdot)$: contrastive learning의 output embedding vector
- 학습 input : {query : q, positive ad : a_q^+, hard-negative ad : a_q^-}
- Both losses(from relevance teacher model & contrastive learning)는 summation & back-propagation되며 BERT^{Uni} backbone과 각각의 pooling heads W_{rel}, W_{ctr} update.

References

John Smith (2012)

Learning word embeddings efficiently with noise-contrastive estimation

NIPS2013

J. Zhang et al. (2022)

Uni-Retriever: Towards Learning The Unified Embedding Based Retriever in Bing Sponsored Search

KDD2022