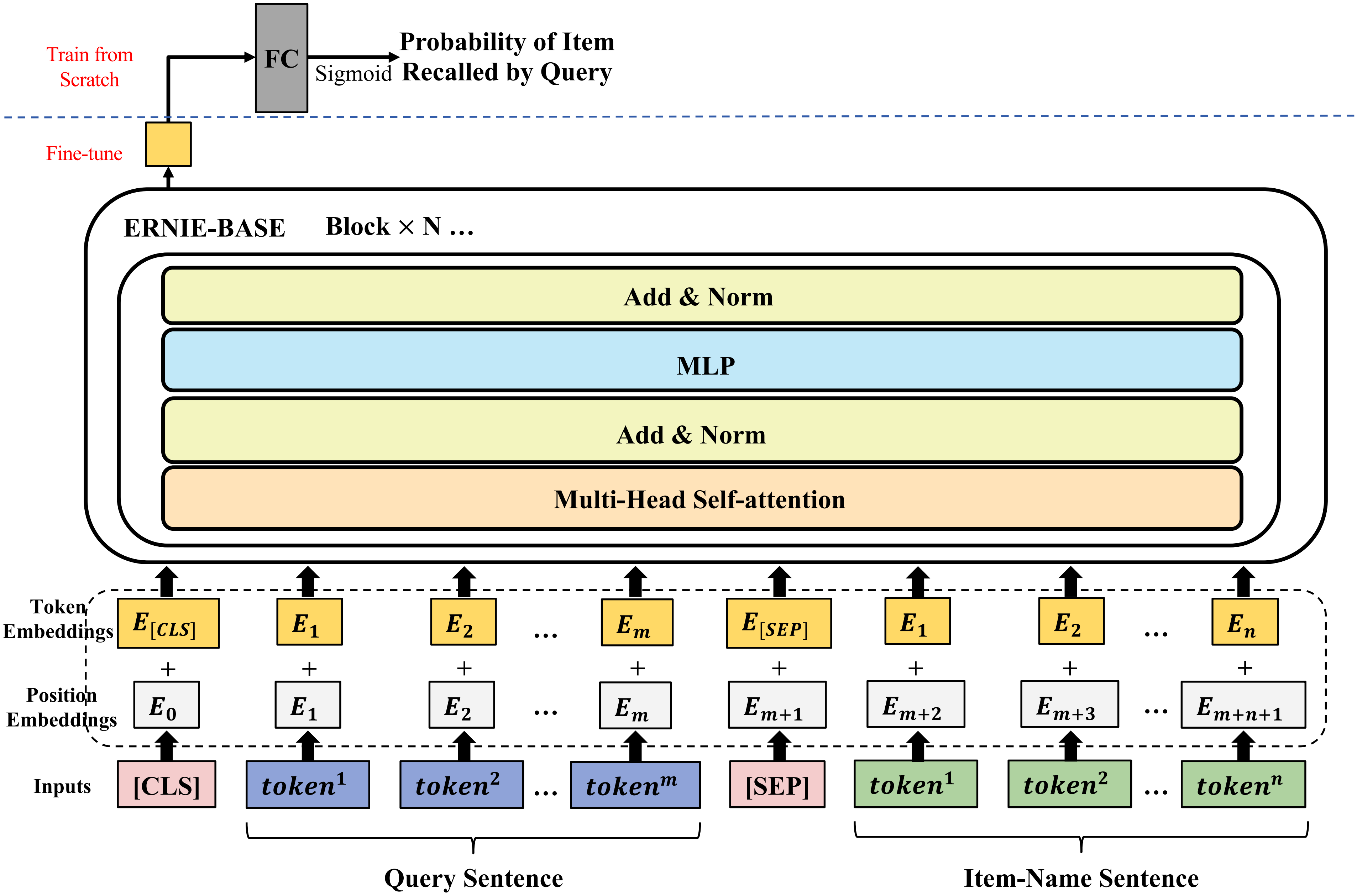
Ernie-Based Recalling

Xiaonan Wang

Search-Ads Algorithms Related Industry Project When Working as a Machine Learning Scientist

(During Period: 2020.09~2022.07)

Version 1: Ernie Single Tower Recalling



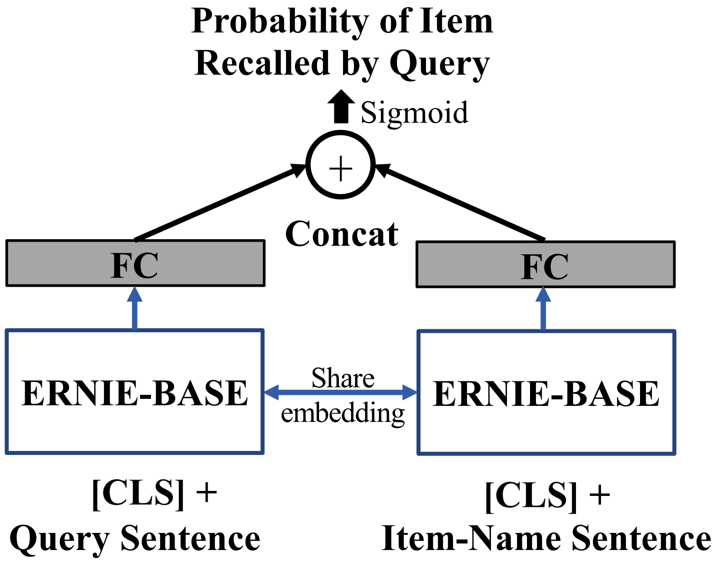
Version 1: Ernie Single Tower Recalling

Summarization

- Modeling recalling problem as a classification problem, and select Top-N items under a certain query when recalling. -> Do like ranking does.
- Construct samples using a pointwise approach, and cross-entropy loss.
- Sample optimization: using clicks and relevance as criteria, perform positive and negative sampling on the entire recall space to ensure that the distribution of the training space is consistent with that of the prediction space.

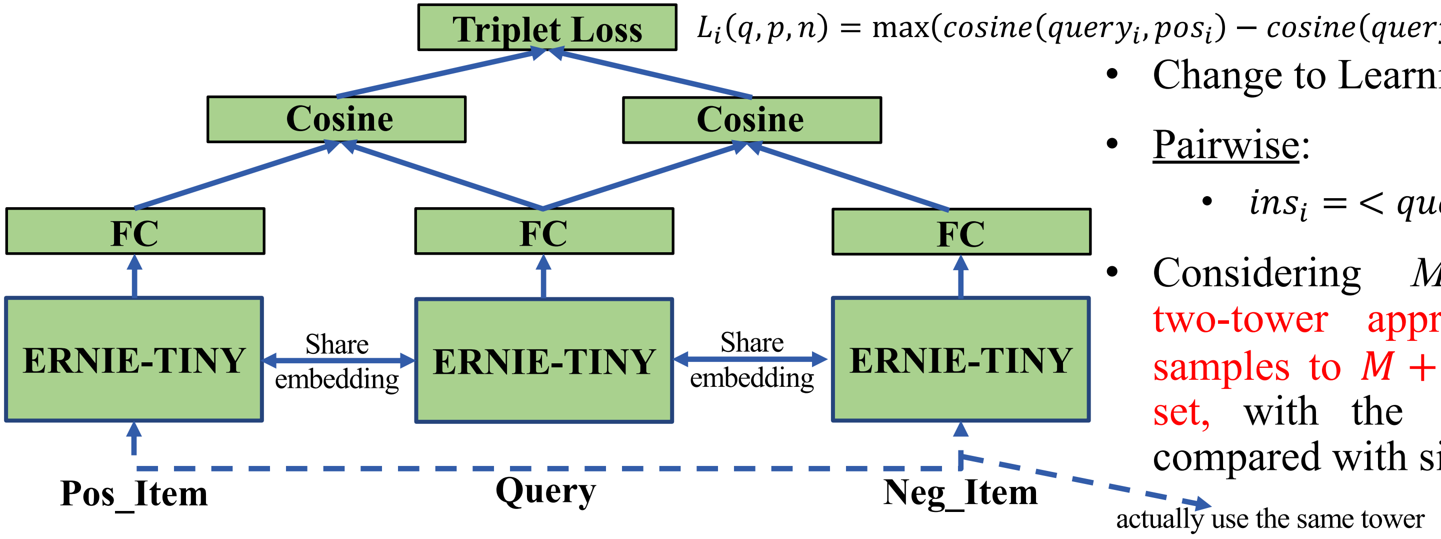
Version 2: Ernie Two Tower Recalling

Modeling as Classification Problem:



- Select Top-N items under a certain query when recalling. -> Do like ranking does.
- Pointwise and Cross-Entropy Loss
- Sample Optimization
- Considering there has M queries and N items in candidate set, two-tower approach decreases $M \times N$ samples to $M + N$ samples in inference set, with the sacrifice of accuracy compared with single tower approach.

Modeling as LTR Problem -- Training:

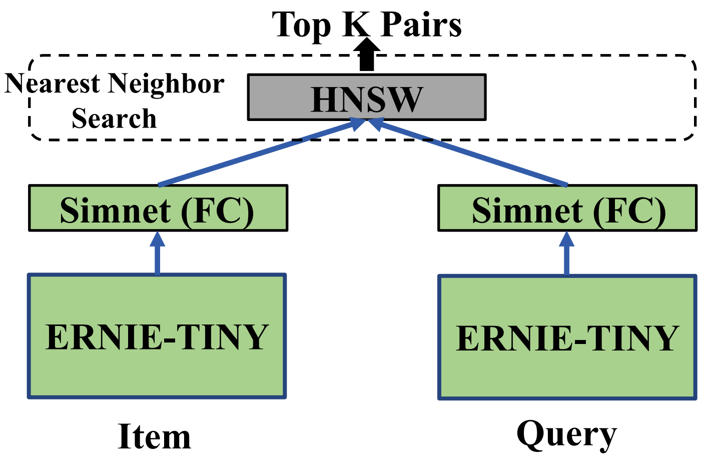


$L_i(q,p,n) = \max(cosine(query_i, pos_i) - cosine(query_i, neg_i) + \alpha, \ 0)$

- Change to Learning-To-Rank approach.
- Pairwise:
  - $ins_i = < query, pos\_item, neg\_item >$
- Considering M queries and N items: two-tower approach decreases $M \times N$ samples to $M + N$ samples in inference set, with the sacrifice of accuracy compared with single tower approach.

Version 2: Ernie Two Tower Recalling

Modeling as LTR Problem -- Inference:



- Faster Inference: Replace the 12-blocks ernie-base model with the extracted blocks ernie-tiny model
- Use the Learning-To-Rank idea and the pairwise paradigm to distinguish positive and negative samples at a fine-grained level, and describe the ranking relationship。