

# ColBERTv2:

Effective and Efficient Retrieval via Lightweight Late Interaction

Monique Monteiro - [moniquelouise@gmail.com](mailto:moniquelouise@gmail.com)

# Conceitos Importantes

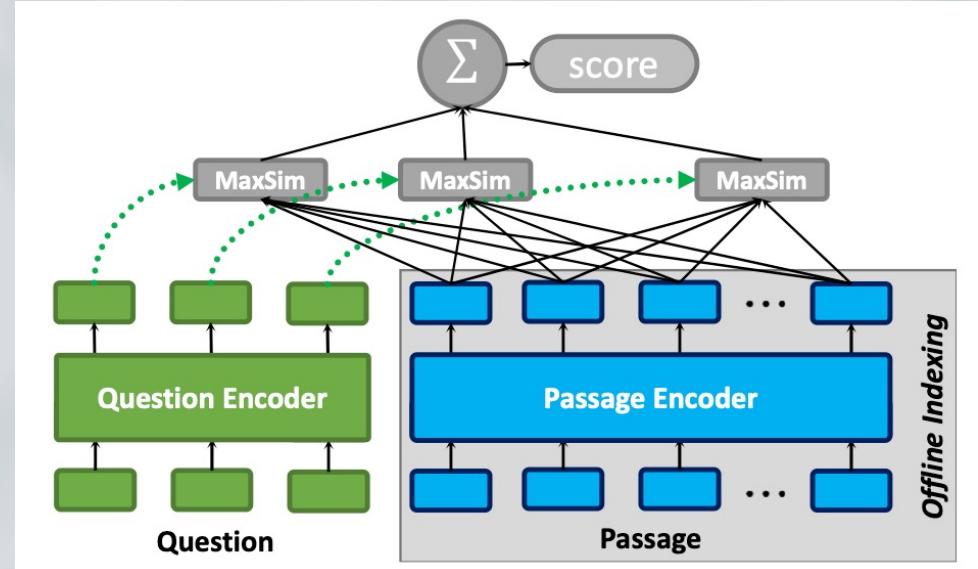
- *Late interaction models:*
  - Produzem representações multi-vetor na granularidade dos tokens
  - Decompõem a modelagem de relevância em computações escaláveis a nível dos tokens
  - Mais efetivos, porém com maior *footprint* de memória
  - Introduzidos pelo ColBERT
    - Um embedding para cada token da consulta
    - Relevância: soma das similaridades máximas entre cada vetor da consulta e todos os vetores do documento

# Contribuição

- ColBERTv2
  - Qualidade estado-da-arte dentro e fora do domínio de treinamento
  - Redução do footprint de memória em 6 a 10X em relação a modelos de interação tardia anteriores
  - Combina destilação de um *cross-encoder* com *hard-negative mining*
  - Mecanismo de compressão residual
    - Pode ser aplicada a modelos já existentes, sem treinamento adicional
  - LoTTE: um novo dataset para avaliação de buscadores fora do domínio (*long-tail topics*)

# Arquitetura

- Consultas e passagens são codificadas independentemente com BERT
- Embeddings resultantes de cada token são projetados em um espaço dimensional inferior
- Em tempo de inferência:
  - Operação “MaxSim”(maior similaridade entre cada token da query e todos os embeddings dos tokens da passagem)



$$S_{q,d} = \sum_{i=1}^N \max_{j=1}^M Q_i \cdot D_j^T$$

# Treinamento

1. ColBERT treinado com triplas como em Khatlab et al. (2021b) indexa as passagens (gera os vetores off-line)
2. Para cada *query* de treinamento, recupera as top-k passagens (do índice, por meio de MaxSim)
3. Usa um *cross-encoder* (MinILM de 22M parâmetros destilado) para rerankear
4. Coleta tuplas “w-way”(*query*, passagem com alto escore de rerankeamento, uma ou mais passagens de menor escore)
  - “w=64” seria o número de passagens (positiva + negativas)?
5. Destila os escores resultantes do *cross-encoder* na arquitetura ColBERT
  - Divergência KL (para alinhar as distribuições)
  - Exemplos negativos in-batch por GPU (entropia cruzada)
  - Função de perda final não é detalhada no artigo

# Representação

- Dado um conjunto de centroides  $C$ , ColBERTv2 codifica cada vetor  $v$  como o índice do centróide mais próximo e um vetor “quantizado”  $r$  que aproxima  $v - C_t$ 
  - KNN sobre uma “amostra de todas as passagens, proporcional à raiz quadrada do tamanho da coleção”
  - Para codificar  $r$ , cada dimensão é quantizada em 1 ou 2 bits.
- Em tempo de busca:
  - $V_{\text{aprox.}} = C_t + r$



# Indexação

1. Seleção de centroides:
  - Seleciona um conjunto de centroides  $C$
  - $|C|$  proporcional à raiz quadrada do número de *embeddings* no corpus
    - Qual é o critério para selecionar os centroides?
2. Codificação das passagens:
  - Tendo selecionado os centroides, codifica cada passagem no corpus
  - Invoca o BERT *encoder* e comprime os *embeddings*
3. Índice invertido:
  - Agrupa os IDs dos embeddings por centroide correspondente e salva no disco
  - Os centroides são as chaves do índice invertido?



# Recuperação

- Dada a representação  $Q$  de uma query:
  1. Para cada  $Q_i$ , os “ $n_{probe}$ ” centroides mais próximos são selecionados
  2. Usando o índice invertido, identifica os embeddings de passagens próximas aos centroides e os descompacta
  3. Computa as similaridades de cosseno com cada vetor da query
  4. Os escores são agrupados por ID da passagem para cada vetor da query
  5. Escores correspondentes à mesma passagem são “max-reduced”
  6. Escores máximos são somados ao longo dos tokens da query
  7. Top  $n$  passagens candidatas são selecionadas para ranqueamento
    - Carrega o conjunto completo de embeddings de cada passagem, e calcula a função de similaridade, reordenando as passagens.

# Dúvida básica

Por que modelos destilados são geralmente mais fortes do que os originais (*vanilla counterparts*)?

# Tópico avançado

Já foi feito algum trabalho de construção de dataset para RI em cima  
do Quora, em especial Quora PT?