ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction

...

11 de maio de 2023

Conceitos Importantes

- Retriever que acopla um mecanismo agressivo de compressão residual com uma estratégia de supervisão denoizada para melhorar simultaneamente a qualidade e a pegada espacial da interação tardia.
- Codificador cruzado e mineração de hard-negative para aumentar a qualidade além de qualquer método existente.
- Usa um mecanismo de compressão residual para reduzir a pegada de espaço da interação tardia em 6–10 ×, preservando a qualidade.

Conceitos Importantes

- Indexação
  - **Seleção do Centróide**. No primeiro estágio, ColBERTv2 seleciona um conjunto de centróides de cluster C.
  - **Codificação de passagem**. Tendo selecionado os centróides, cada passagem é codificada usando somente os centróides.
- **Inversão de índice**. Para suportar a busca rápida do vizinho mais próximo, agrupamos os IDs de incorporação que correspondem a cada centróide e salvamos essa lista invertida no disco

Contribuições do Artigo

- Focaram na recuperação de interação tardia e foi investigada a compressão usando uma abordagem de compressão residual que pode ser aplicada pronta para uso em modelos de interação tardia, sem treinamento especial.
- A abordagem de compactação residual do ColBERTv2 reduz significativamente os tamanhos dos índices em comparação com o ColBERT padrão.
- Enquanto o ColBERT requer 154 GiB para armazenar o índice para MS MARCO, o ColBERTv2 requer apenas 16 GiB ou 25 GiB ao compactar embeddings para 1 ou 2 bit(s) por dimensão, respectivamente, resultando em taxas de compactação de 6 a 10 ×.

Contribuições do Artigo

- Focaram na recuperação de interação tardia e foi investigada a compressão usando uma abordagem de compressão residual que pode ser aplicada pronta para uso em modelos de interação tardia, sem treinamento especial.
- A abordagem de compactação residual do ColBERTv2 reduz significativamente os tamanhos dos índices em comparação com o ColBERT padrão.
- Enquanto o ColBERT requer 154 GiB para armazenar o índice para MS MARCO, o ColBERTv2 requer apenas 16 GiB ou 25 GiB ao compactar embeddings para 1 ou 2 bit(s) por dimensão, respectivamente, resultando em taxas de compactação de 6 a 10 ×.

Obrigado

Manoel Veríssimo