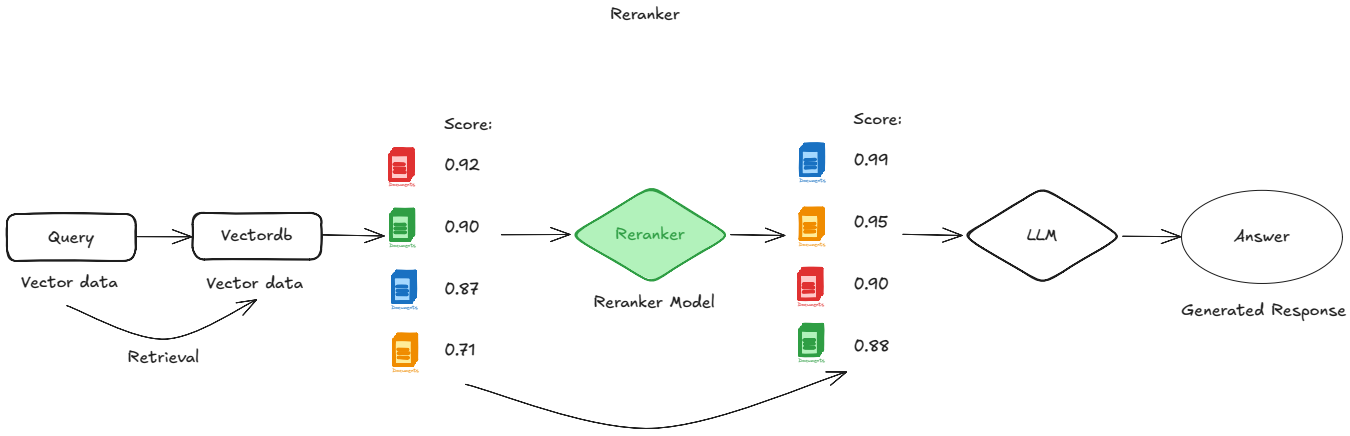
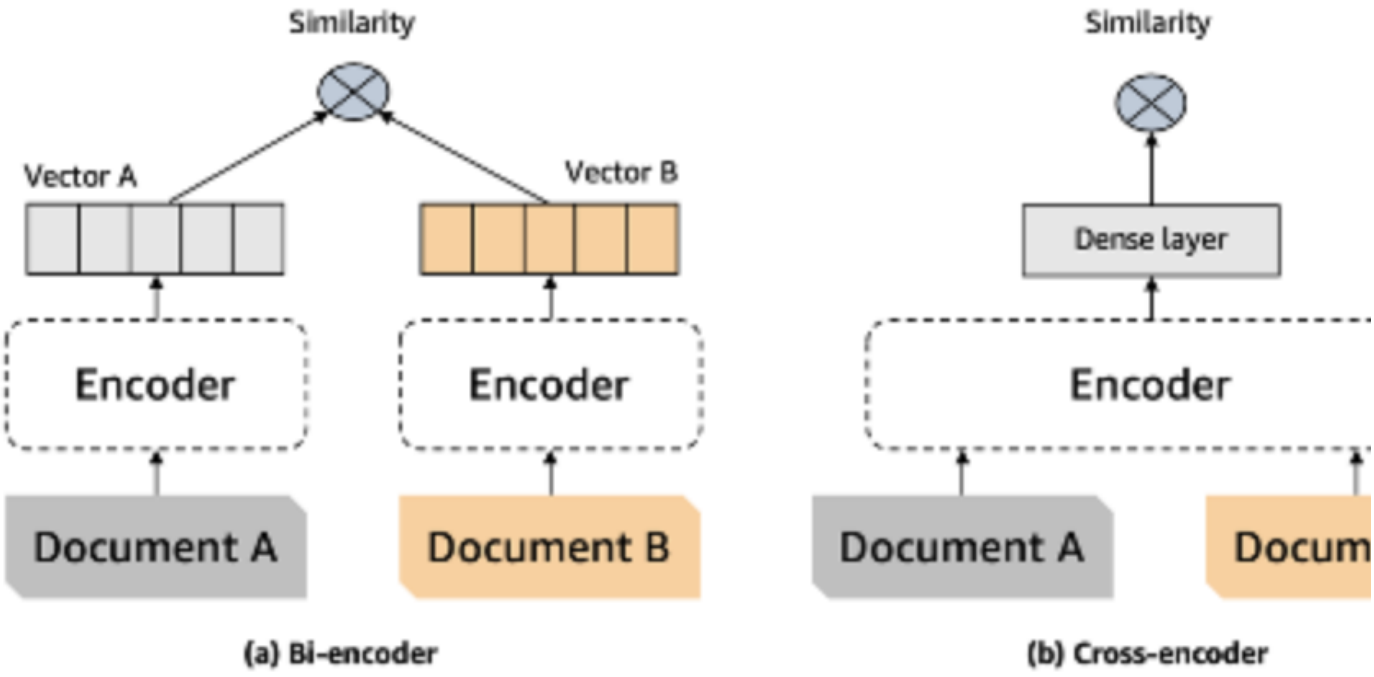
리랭커(Reranker)



개요

Reranker(리랭커)는 현대적인 두 단계 검색 시스템(Two-Stage Retrieval System)에서 사용되는 핵심 컴포넌트 입니다. 대규모 데이터셋에서 효율적이고 정확한 검색을 수행하기 위해 설계되었으며, 주로 첫번째 단계에서 Retriever 가 찾아낸 문서들의 순위를 재조정 하는 역할을 합니다. Reranker 는 검색 시스템의 두 번째 단계에서 작동하며, 초기 **검색 결과의 정확도를 향상시키는 것을 목표**로 합니다. Retriever 단계에서 대규모 문서 집합에서 관령성 있는 후보 문서들을 빠르게 추출한 후 **Reranker 단계에서 이 후보 문서들을 더 정교하게 분석하여 최종적으로 순위를 결정**합니다.

작동 원리



Reranker와 Retrieval의 차이점 및 장단점

Retrieval은 주어진 질문에 대해 관련 있는 문서를 빠르게 찾아내는 데 중점을 둡니다. 반면 **Reranker**는 Retrieval 단계에서 찾은 문서들을 LLM을 이용하여 질문과의 관련성을 재평가하고 순위를 조정하여 더 정확한 결과를 제공하는 데 집중합니다.

1. Retrieval

- **장점:**
  - 빠른 속도: 일반적으로 벡터 검색이나 BM25와 같은 효율적인 알고리즘을 사용하여 빠르게 문서를 검색합니다.
  - 광범위한 검색: 대규모 데이터셋에서도 효과적으로 작동합니다.
- **단점:**
  - 정확도: 질문과 문서 간의 의미론적 유사성을 완벽하게 파악하지 못할 수 있습니다.
  - 맥락 이해 부족: 질문의 맥락을 고려하지 않고 단순히 키워드 매칭에 의존하는 경향이 있습니다.

2. Reranker

- **장점:**
  - 높은 정확도: LLM을 사용하여 질문과 문서의 의미론적 유사성을 정확하게 평가합니다.
  - 맥락 이해: 질문의 맥락을 고려하여 문서의 순위를 조정합니다.
- **단점:**
  - 속도: LLM을 사용하기 때문에 Retrieval에 비해 속도가 느립니다.
  - 계산 비용: LLM 사용에 따른 계산 비용이 발생합니다.

bi-encoder(Retrieval)와 cross-encoder (Reranker)의 주요 차이점

1. 입력 방식:

- **bi-encoder:** 질문과 문서를 **각각 별도로** 인코딩하여 두 개의 벡터를 생성합니다.
- **cross-encoder:** 질문과 문서 쌍을 **동시에** 입력으로 받아 하나의 벡터를 생성하거나 관련성 점수를 출력합니다.

2. 작동 방식:

- **bi-encoder:** 질문과 문서를 각각 인코딩한 후, 두 벡터 사이의 유사도 (cosine similarity 등)를 계산하여 관련성을 평가합니다.
- **cross-encoder:** 질문과 문서를 함께 분석하여 둘 사이의 의미론적 유사성을 직접적으로 평가합니다.

3. 장단점:

특징	bi-encoder	cross-encoder
정확도	상대적으로 낮음	상대적으로 높음
속도	빠름	느림
계산 비용	낮음	높음
맥락 이해	제한적	뛰어남
활용	초기 검색, 대규모 데이터셋	Reranker, 정밀한 검색

4. 핵심 차이:

- bi-encoder는 질문과 문서를 독립적으로 처리하여 각각의 의미를 파악합니다.
- cross-encoder는 질문과 문서를 함께 처리하여 둘 사이의 상호 작용과 맥락을 고려합니다.

비유를 통해 이해해 보세요:

- **bi-encoder:** 두 사람의 외모를 각각 사진으로 보고 닮은 정도를 판단하는 것과 같습니다.
- **cross-encoder:** 두 사람을 직접 만나 대화를 나눠보고 서로 얼마나 잘 맞는지 판단하는 것과 같습니다.

결론적으로, bi-encoder는 빠른 속도와 낮은 계산 비용이 장점이지만, cross-encoder는 더 정확하고 맥락을 잘 이해하는 장점이 있습니다. 따라서, 검색 시스템에서는 bi-encoder를 사용하여 빠르게 후보 문서를 찾고, cross-encoder (reranker)를 사용하여 정확도를 높이는 방식을 결합하여 사용하는 경우가 많습니다.

Two-Stage Retrieval System 방식 (Retrieval + Reranker)의 장점

Retrieval과 Reranker를 함께 사용하는 2-Stage 방식은 각각의 장점을 결합하여 더욱 효과적인 검색 시스템을 구축할 수 있습니다.

- **높은 정확도와 효율성:** Retrieval을 통해 빠르게 후보 문서를 찾고, Reranker를 통해 정확도를 높일 수 있습니다.
- **유연성:** Retrieval과 Reranker를 독립적으로 구성하고 최적화할 수 있습니다.
- **향상된 사용자 경험:** 더욱 정확하고 관련성 높은 검색 결과를 제공하여 사용자 만족도를 높일 수 있습니다.

2-Stage 방식의 작동 예시:

1. 사용자가 질문을 입력합니다.
2. Retrieval 단계에서 벡터 검색을 사용하여 질문과 관련 있는 후보 문서 10개를 빠르게 찾습니다.
3. Reranker 단계에서 LLM을 사용하여 10개의 후보 문서를 질문과의 관련성에 따라 재정렬합니다.
4. 재정렬된 문서 목록을 사용자에게 제공합니다.

이처럼 2-Stage 방식은 Retrieval의 효율성과 Reranker의 정확성을 결합하여 최적의 검색 결과를 제공하는 데 효과적인 방법입니다.