
HARD NEGATIVE MINING STRATEGIES FOR RETRIEVAL

Sionic AI 리서치 인턴 김태은

EMBEDDING MODELS

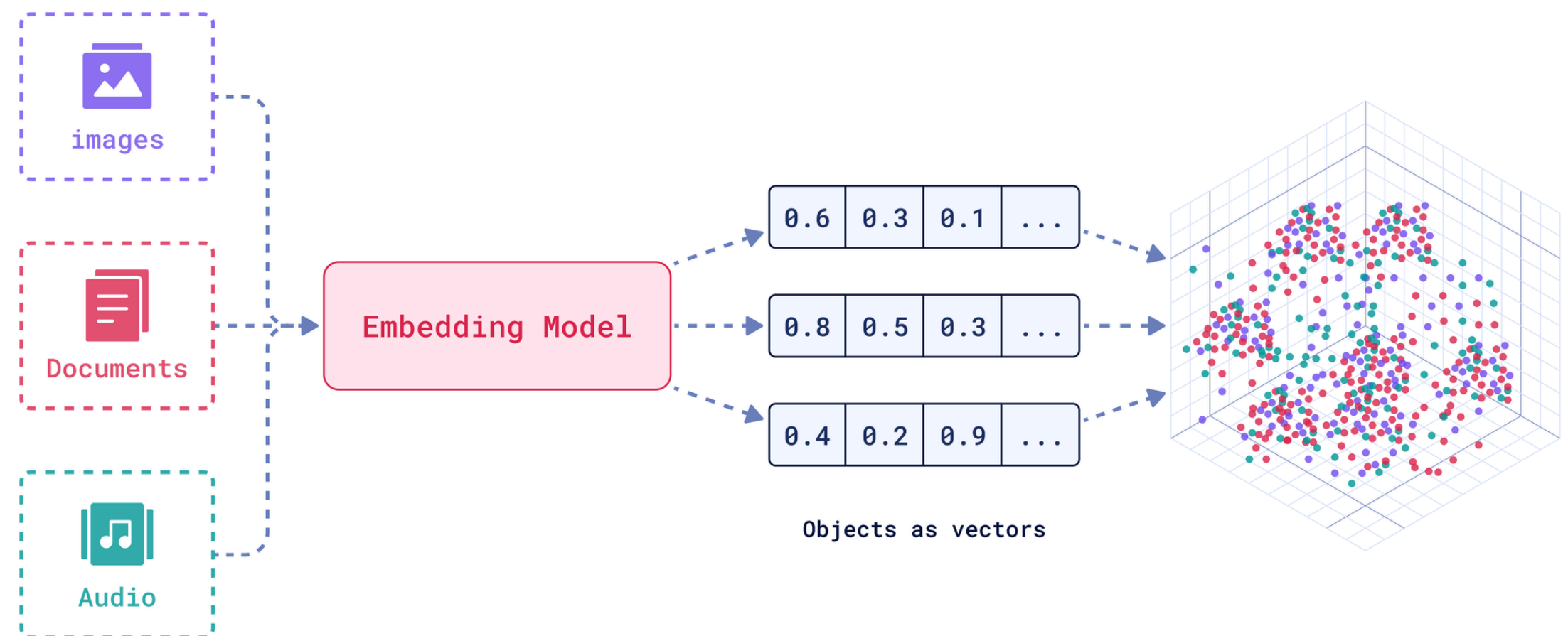
텍스트, 오디오, 또는 시각적 데이터를 고유한 벡터 공간 내에서 벡터 형태로 변환하여 의미 정보를 인코딩하는 모델

tasks:

- Information retrieval
- question answering
- semantic textual similarity

Application

- RAG



NEGATIVES AND CONTRASTIVE LEARNING

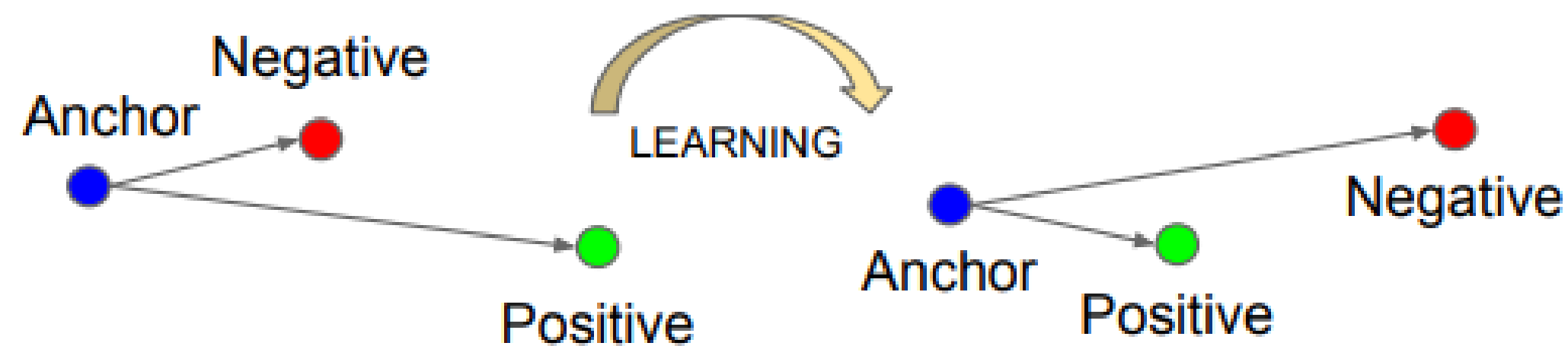


Figure 3. The **Triplet Loss** minimizes the distance between an *anchor* and a *positive*, both of which have the same identity, and maximizes the distance between the *anchor* and a *negative* of a different identity.

일반적으로, 임베딩 모델은 대조 학습 (contrastive learning)을 통해 훈련됩니다. 대조 학습은 데이터 포인트 간의 차이를 학습하는 자가 지도 학습(self-supervised learning)의 한 종류입니다.

Anchor: query

Positive: answer

Negative: chunks that are irrelevant to a query or insufficient to answer it

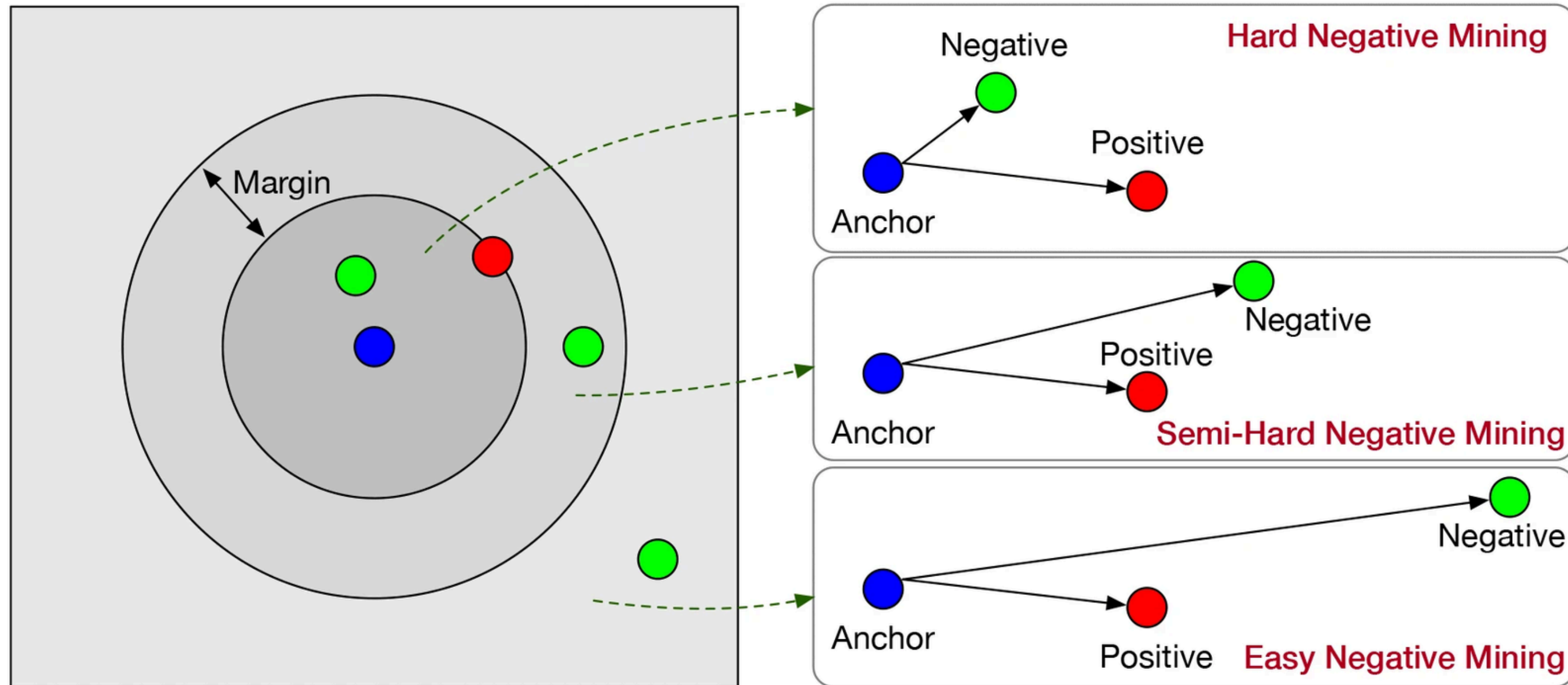
INFONCE LOSS FUNCTION

$$\begin{aligned} & L(q_i, p_i^+, p_{i,1}^-, \dots, p_{i,n}^-) \\ &= -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^n e^{\text{sim}(q_i, p_{i,j}^-)}}. \end{aligned} \tag{2}$$

유사한 데이터 포인트는 가깝게 만들고, 유사하지 않은 데이터 포인트 간의 거리를 늘리는 것을 목표로 합니다.

HARD NEGATIVES

임베딩 공간에서 쿼리와 의미적으로 가깝지만 답을 하기엔 부족한 네거티브 데이터



쉬운 네거티브는 모델 성능 향상에 큰 기여를 하지 않을 수 있습니다

NEGATIVE MINING STRATEGIES

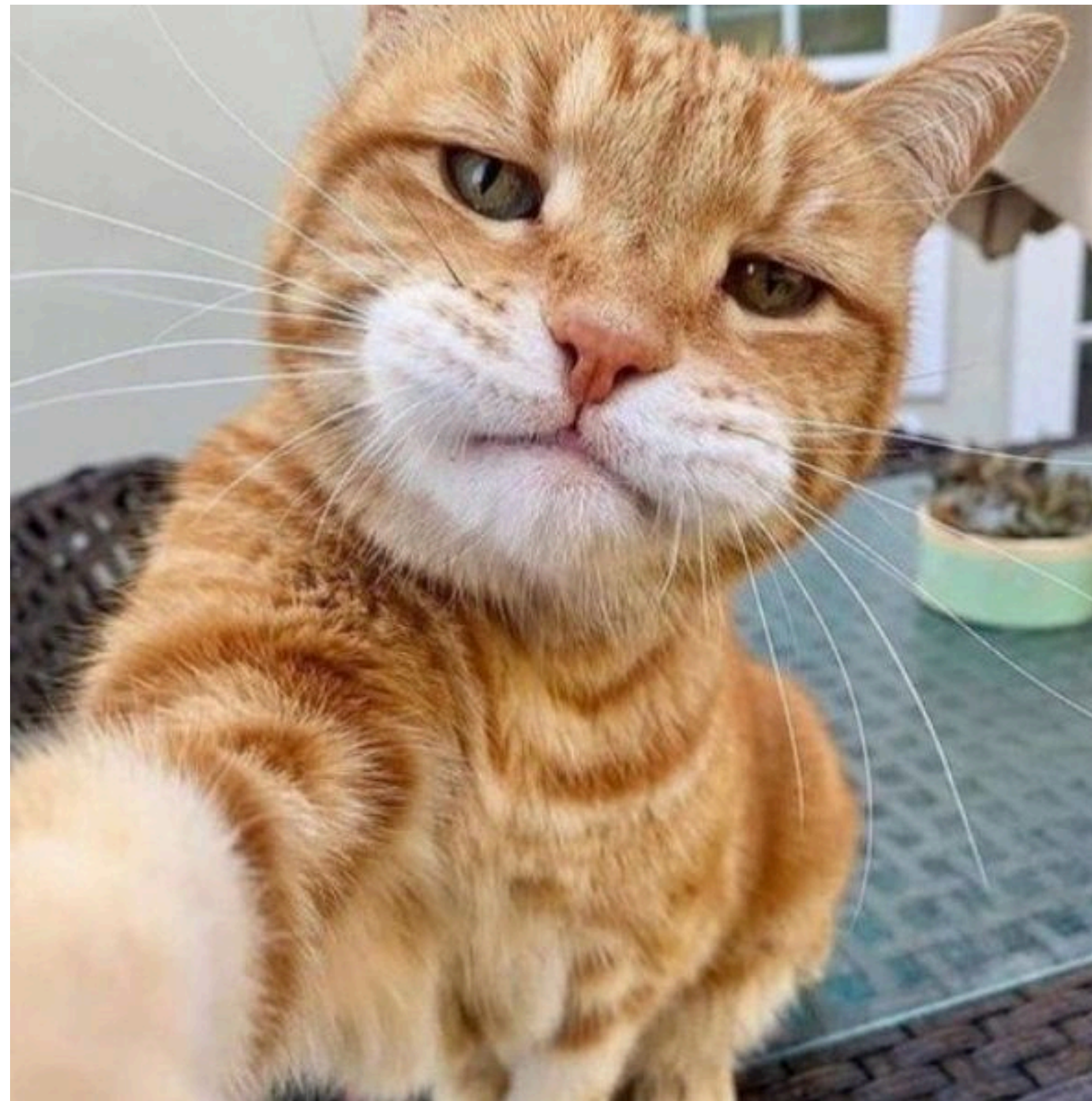
RANDOM SAMPLING

Definition

각 쿼리에 대해 랜덤 쿼리에서 랜덤 네거티브 샘플 추출

Note

- 심플
- 너무 쉬운 네거티브가 나올 수 있음



IN-BATCH NEGATIVE

학습방법: In-Batch Negative


학습 데이터셋

Q1 — D1

Q2 — D2

Q3 — D3

Q4 — D4



Q: (4, 512)

D: (4, 512)

$$Q \cdot D^T \rightarrow (4,4)$$

similarity = $\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$

	D1	D2	D3	D4
Q1	○	×	×	×
Q2	×	○	×	×
Q3	×	×	○	×
Q4	×	×	×	○

Definition

배치 내에서 다른 쿼리에 대응하는 포지티브를 네거티브로 사용

$$B \times d * d * B = B \times B$$

Note

- 계산 효율적
- 랜덤 샘플링의 일종. 어려운 negative일 가능성이 적음

BEST MATCH 25: BM25

TF-IDF

$$TF(t, d) = \frac{\text{문서 } d \text{ 에서 단어 } t \text{ 가 등장한 횟수}}{\text{문서 } d \text{ 에 등장한 모든 단어의 수}}$$

$$IDF(t, D) = \log \left(\frac{\text{총 문서의 개수}}{\text{단어 } t \text{ 를 포함하는 문서의 수}} \right)$$

$$TF-IDF(t, d, D) = TF(t, d) * IDF(t, D)$$

BM25

$$score(D, Q) = \sum_{i=1}^n IDF(q_i) * \frac{\text{문서 } D \text{ 에서 } q_i \text{ 의 term frequency}}{f(q_i, D) + \text{문서 } D \text{ 의 길이} * \frac{f(q_i, D) * (k_1 + 1)}{f(q_i, D) + k_1 * (1 - b + b * \frac{|D|}{avgdl})}$$

파라미터 k_1 문서 집합의 평균 문서 길이 b

Definition

TF-IDF의 한 형태로, 문서에서 단어의 중요도를 측정하며, 문서 길이를 정규화함

Note

- 효과적인 관련성 순위 부여
- lexical 초점

TOP-K RETRIEVAL

쿼리와 가장 유사한 상위 K개 후보를 선택
하며, 포지티브 문장은 제외함

Top-k shifted by N

- 위양성(false negatives)을 고려하여 상위 N순위 이후의 탑-K 네거티브를 선택합니다.
- doesn't account for relevance score; may lose valuable hn or keep false neg

Top-k with absolute threshold

- 유사도 점수가 일정 임계값을 초과하는 네거티브를 무시함
- TopK-Abs thresholds negative scores with respect only to the query, regardless the positive passage relevance.



HARD NEGATIVE MINING STRATEGIES

POSITIVE AWARE HARD NEGATIVE MINING

쿼리와의 관련성을 고려하여 포지티브 문장을 반영한 하드 네거티브 마이닝; false negatives 방지

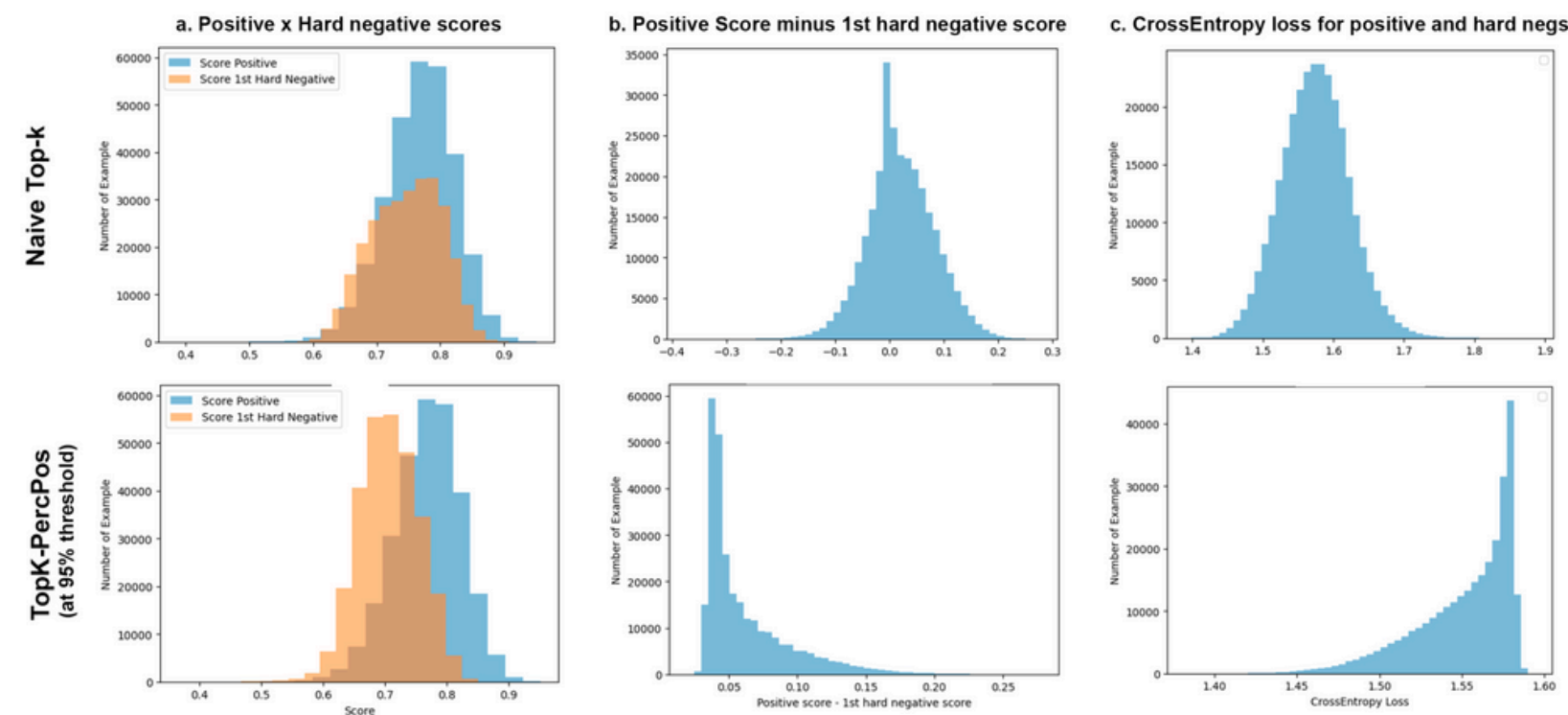


Figure 1: Histograms comparing *Naive Top-k* and *TopK-PercPos* mining methods

Top-k Margin Positive Threshold

- 0.05 in NV-Retriever
- $\text{max.neg.threshold} = \text{pos.score} - \text{absolute.margin}$

Top-k Percentage Positive Threshold

- 95% of positive score
- $\text{max.neg.threshold} = \text{pos.score} * \text{percentage.margin}$

MINED HARD NEGATIVES + DISTILLATION FROM CROSS ENCODER

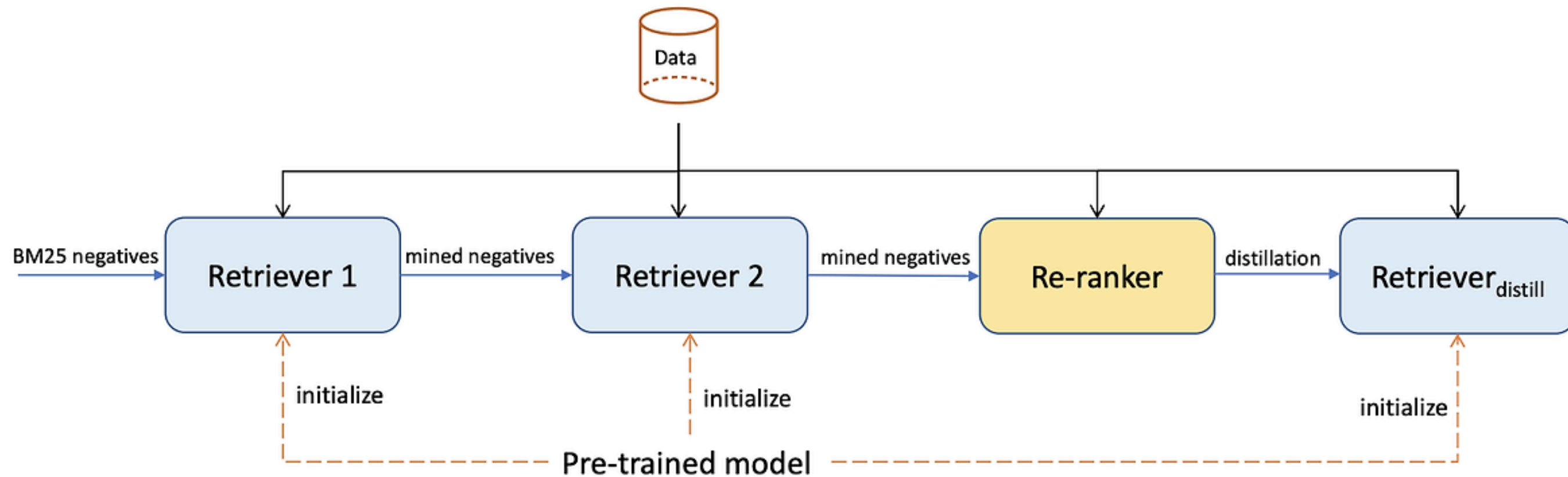


Figure 2: Illustration of our supervised fine-tuning pipeline. Note that we only use SimLM to initialize the biencoder-based retrievers. For cross-encoder based re-ranker, we use off-the-shelf pre-trained models such as ELECTRA_{base}.

BM25 negatives + In-batch negatives => retriever 1 => mined-negatives-1 => retriever 2
=> mined-negatives-2 => Re-ranker => re-ranked negatives-2 => Distilled Retriever

SYNTHETIC HARD NEGATIVE MINING + IN-BATCH

You have been assigned a retrieval task **{task}**
Your mission is to write one text retrieval example for this task in JSON format. The JSON object must contain the following keys:

- **"user_query"**: a string, a random user search query specified by the retrieval task.
- **"positive_document"**: a string, a relevant document for the user query.
- **"hard_negative_document"**: a string, a hard negative document that only appears relevant to the query.

Please adhere to the following guidelines:

- The "user_query" should be **{query_type}, {query_length}, {clarity}**, and diverse in topic.
- All documents should be at least **{num_words}** words long.
- Both the query and documents should be in **{language}**.

... (omitted some for space)

Your output must always be a JSON object only, do not explain yourself or output anything else. Be creative!



```
{"user_query": "How to use Microsoft Power BI for data analysis",  
"positive_document": "Microsoft Power BI is a sophisticated tool that requires time and practice to master. In this tutorial, we'll show you how to navigate Power BI ... (omitted) ",  
"hard_negative_document": "Excel is an incredibly powerful tool for managing and analyzing large amounts of data. Our tutorial series focuses on how you...(omitted)" }
```


ONLINE HARD NEGATIVE MINING: ANCE

- 근접 근사(ANN)를 활용한 검색으로 하드 네거티브를 학습
- Faiss IndexFlatIP를 통해 상위 200개 중 하나를 무작위로 샘플링
- 매 m 번째 배치마다 체크포인트를 만들고 새로운 인덱스 표현을 추론하여 비동기적으로 수행함

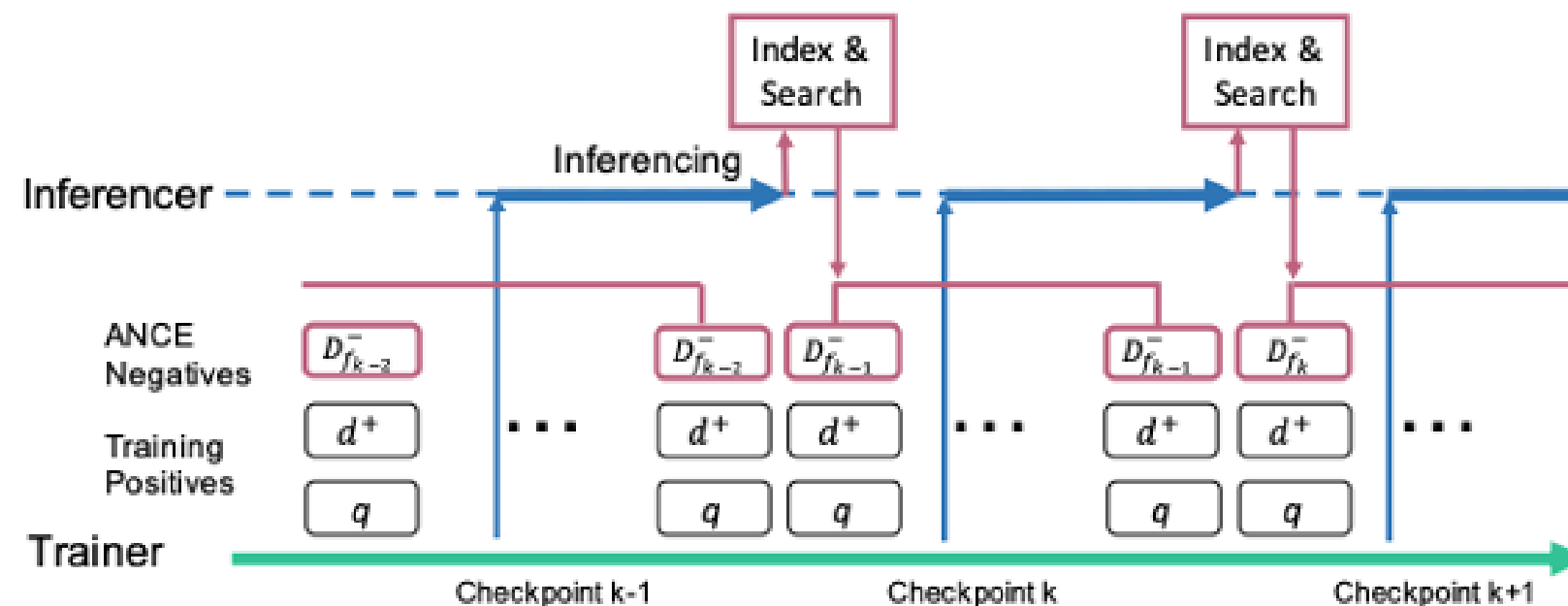


Figure 2: ANCE Asynchronous Training. The Trainer learns the representation using negatives from the ANN index. The Inferencer uses a recent checkpoint to update the representation of documents in the corpus and once finished, refreshes the ANN index with most up-to-date encodings.

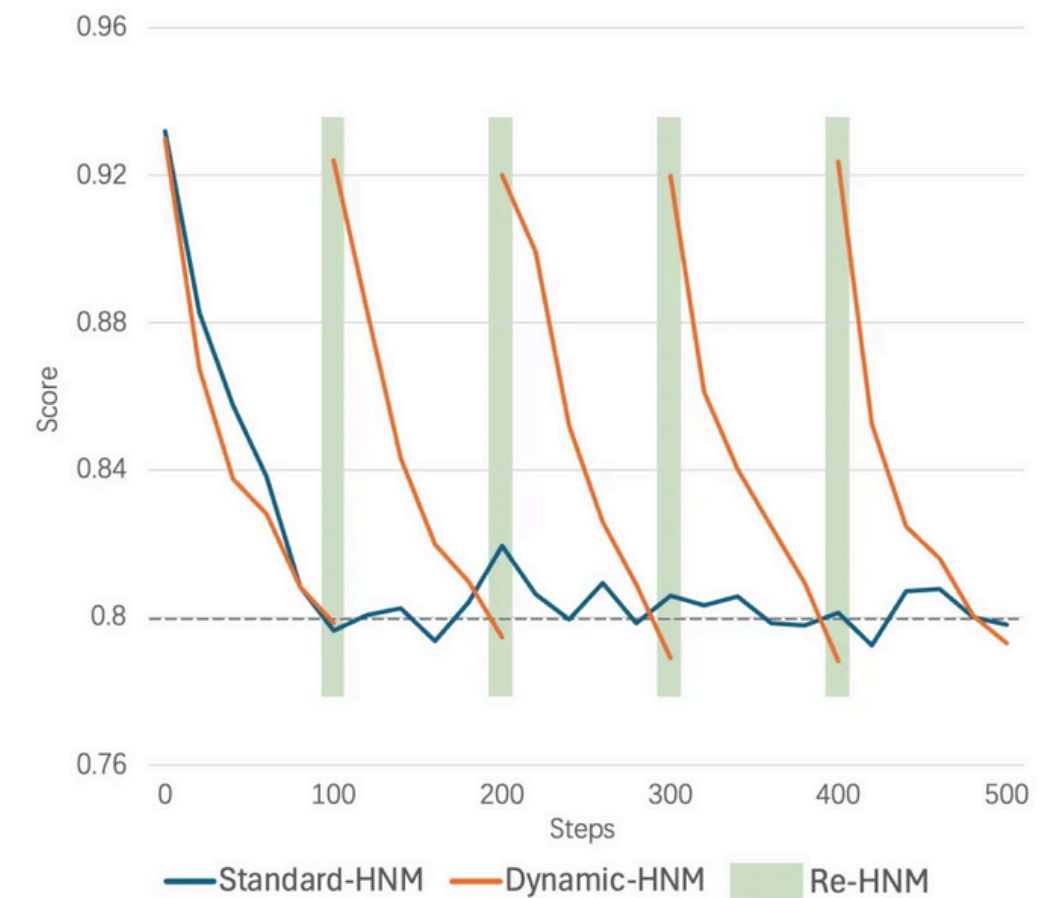


Figure 2: Dynamic Hard Negative Mining vs. Standard Hard Negative Mining: Score-Steps Curves. Hard negatives are checked every 100 steps. When the score multiplied by 1.15 is less than the initial score and the absolute value of the score is less than 0.8, we consider the negative example no longer difficult and replace it with a new hard negative.

**FEEL FREE TO APPROACH US IF
YOU HAVE ANY QUESTIONS.**

Thank you for listening!