고려대학교 KOREA UNIVERSITY

Natural Language Processing & Artificial Intelligence

겨울방학 세미나

고려대학교 NLP&AI 연구실
발표자: 손준영

◆ 논문 선정 기준과 선정한 논문: (1)

1. Matryoshka Representation Learning (NeurIPS 2022)

- MTEB English Leaderboard 상위 10개 모델 중 8개가 LLMs 기반의 백본을 활용하여 학습되었음
- 모델 크기와 비례하여 Embedding Dimensions가 대체로 매우 큰데 (3584~8192), 효율성 관점에서 어떤 기술을 고려/적용하고 있을지에 대한 의문.

  - 이러한 industry 관점을 고려하여 이들을 teacher models로 활용, distillation하여 구축한 smaller 모델이 jasper 모델

- Voyage-3 모델이나 Stella and Jasper 모델은 이러한 효율성을 고려하기 위하여 vector dimensions를 축소하기 위한 Matryoshka Representation Learning (MRL) 기반의 방법론을 활용하였음

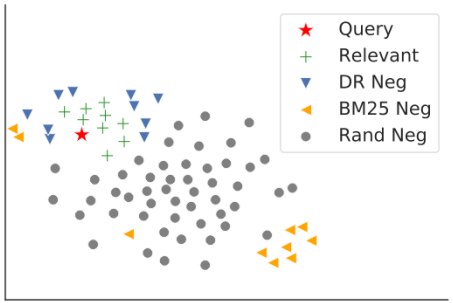
Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)
1	voyage-3-m-exp					74.03
2	NV-Embed-v2	7851	29.25	4096	32768	72.31
3	jasper_en_vision_language_v1					72.02
4	bge-en-icl	7111	26.49	4096	32768	71.67
5	LENS-d8000	7111	26.49	4096	32768	71.62
6	LENS-d4000	7111	26.49	4096	32768	71.21
7	stella_en_1.5B_v5	1543	5.75	8192	131072	71.19
8	SFR-Embedding-2_R	7111	26.49	4096	32768	70.31
9	gte-Qwen2-7B-instruct	7613	28.36	3584	131072	70.24
10	stella_en_400M_v5	435	1.62	8192	8192	70.11

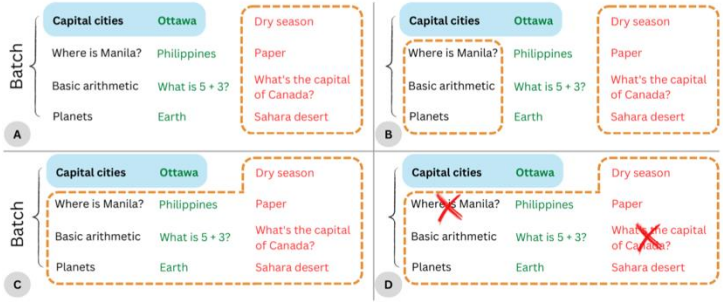
◆ 논문 선정 기준과 선정한 논문: (2)

2. Contextual Document Embeddings (ICLR 2025 8/6/6)

- Contrastive Learning의 "In-batch negatives" 품질/난이도가 성능에 주는 영향이 큰 특성을 고려하기 위한 다양한 접근 방법들이 있음
  - GISTEmbed[1]: Teacher Model을 활용하여 In-batch negatives 내 false negatives를 동적으로 filtering
  - ANCE[2]: In-batch negatives가 local하게 샘플링되는 기존 방법론
    → In-batch negatives를 global corpus에서 비동기적으로 업데이트하는 approximation 방법론 제안

- 이러한 Contrastive Learning의 특성을 극대화하기 위한 방법론을 연구적으로 접근하였고 ICLR에서 높은 점수를 받아서 공유하고 싶었음





1. Solatorio, A. V. (2024). Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:2402.16829*.
2. Xiong, L., Xiong, C., Li, Y., Tang, K. F., Liu, J., Bennett, P., ... & Overwijk, A. (2020). Approximate nearest neighbor negative contrastive learning for dense text retrieval. arXiv preprint arXiv:2007.00808.

◆ Matryoshka Representation Learning

Matryoshka Representation Learning

Aditya Kusupati[*†◇], **Gantavya Bhatt**[*†], **Aniket Rege**[*†],
Matthew Wallingford[†], **Aditya Sinha**[◇], **Vivek Ramanujan**[†], **William Howard-Snyder**[†],
Kaifeng Chen[◇], **Sham Kakade**[‡], **Prateek Jain**[◇] **and Ali Farhadi**[†]
[†]University of Washington, [◇]Google Research, [‡]Harvard University
{kusupati,ali}@cs.washington.edu, prajain@google.com

NeurIPS 2022

◆ Matryoshka Representation Learning

Abstract

- 표현 학습(Representation Learning)은 다양한 자질의 문맥을 **고정된 크기의 표현 벡터(Representation vector)**로 표현하려는 학습으로, 다양한 다운스트림 작업에 활용될 수 있음

- 그러나, 고정된 크기의 표현 벡터는 작업별로 과도하거나 부족한 자원을 할당하게 되어 비효율성을 초래할 수 있음
  - E.g.) "고양이"와 같이 시각적으로 뚜렷한 특징을 가진 클래스는 비교적 간단한 모델 또는 저차원 벡터로 충분히 분류 가능
    "샤무아염소"처럼 비슷한 외형을 가진 동물과 구분해야 하는 클래스는 더 높은 차원의 세부적인 표현이 필요

  - 수백만 개의 후보군 중 관련성이 있는 일부 후보군만 찾기 위한 "검색(Retrieval)" 작업과
    검색된 후보군 사이에서 보다 정밀한 "재순위화(Reranking)"는 작업의 정밀도/속도의 관점에서 서로 다른 요구사항을 가짐

  - LLMs와 같은 고차원/고자원 모델을 표현 학습에 활용할 수록 이러한 작업 복잡도와 효율성 사이의 고려가 더욱 필요함

- 이러한 **다양한 컴퓨팅 자원과 다운스트림 작업의 요구**에 flexibly adaptive할 수 있는 Representation Learning 방법론인 Matryoshka Representation Learning (MRL)을 제안

◆ Matryoshka Representation Learning

Overview

- **Matryoshka?**
  - 러시아를 대표하는 전통 공예품으로, 서로 다른 크기의 나무 인형들이 하나의 큰 인형 안에 겹겹이 들어 있는 구조
  ➔ ***고차원 벡터가 저차원 벡터를 포함하는 Coarse-to-Fine Representation***

  - 모든 인형은 동일하거나 유사한 디자인을 가짐
  ➔ ***서로 다른 크기의 Coarse-to-Fine Representations를 동일한 목표로 학습하고, 다운스트림 작업의 복잡도에 따라 적합한 차원의 벡터를 선택적으로 활용***

참고 자료. 러시아 전통 공예품인 마트료시카 인형

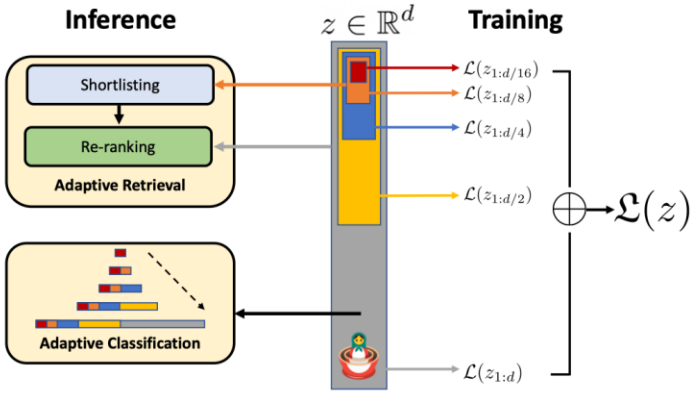


Figure 1: Matryoshka Representation Learning is adaptable to any representation learning setup and begets a Matryoshka Representation z by optimizing the original loss $\mathcal{L}(.)$ at $O(\log(d))$ chosen representation sizes. Matryoshka Representation can be utilized effectively for adaptive deployment across environments and downstream tasks.

◆ Matryoshka Representation Learning

Matryoshka Representation Learning

$$\min_{\{\mathbf{W}^{(m)}\}_{m\in\mathcal{M}},\,\theta_F} \frac{1}{N}\sum_{i\in[N]}\sum_{m\in\mathcal{M}} c_m \cdot \mathcal{L}\left(\mathbf{W}^{(m)} \cdot F(x_i;\theta_F)_{1:m}\,;\, y_i\right)$$

For $d \in \mathbb{N}$, consider a set $\mathcal{M} \subset [d]$ of representation sizes.
For a datapoint x in the input document $\mathcal{X}$, ***the goal of MRL is to learn a d-dimensional representation vector $z \in \mathbb{R}^d$***.

For every $m \in \mathcal{M}$, MRL enables each of first m dimensions of the embedding vectors, $z_{1:m} \in \mathbb{R}^m$ ***to be independently capable of being a transferable and general-purpose representation of the datapoint x***.

$$c_m = 1\ for\ all\ m \in \mathcal{M}$$
$$F(\cdot;\theta_F)\colon \mathcal{X} \to \mathbb{R}^d$$

$\mathcal{L}\colon \mathbb{R}^L \times [L] \to \mathbb{R}_+$ is the multi-class softmax cross-entropy loss

$$\mathcal{M} = \{8,16,\dots,1024,2048,\dots\}$$

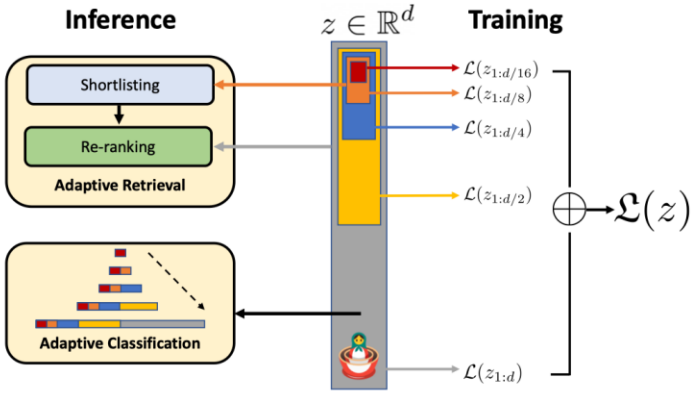


Figure 1: Matryoshka Representation Learning is adaptable to any representation learning setup and begets a Matryoshka Representation z by optimizing the original loss $\mathcal{L}(.)$ at $O(\log(d))$ chosen representation sizes. Matryoshka Representation can be utilized effectively for adaptive deployment across environments and downstream tasks.

◆ Matryoshka Representation Learning

Experimental Setting

(a) Supervised learning for vision:

- ResNet50 on ImageNet-1K
- ViT-B/16 on JFT-300M

(b) Contrastive Learning for vision + language:

- ALIGN model with ViT-B/16 vision encoder and BERT language encoder on ALIGN data

(c) Masked language modeling:

- BERT on English Wikipedia and BooksCorpus

Matryoshka dimensions

- ResNet50 ($d = 2048$):
  $\mathcal{M} = \{8,16,32,64,128,256,512,1024,2048\}$

- ViT-B/16 and BERT ($d = 768$):
  $\mathcal{M} = \{12,24,48,96,192,384,768\}$

◆ Matryoshka Representation Learning

Image Classification

(a) Linear classification (Figure 2)

- MRL vs. FF:
  - MRL 모델은 모든 representation sizes에서 일관적으로 FF 모델과 유사한 성능 관측
  - MRL-E는 16차원부터 FF 모델과 거의 동일한 정확도를 보임

➔ 다양한 sizes에서 성능 유지 + 효율성

(b) 1-NN Accuracy (Figure 3, 4)

- MRL vs. FF:
  - 낮은 차원에서는 Matryoshka 표현이 최대 2% 더 높은 정확도를 보임
  - 다른 차원에서는 FF 모델과 유사한 성능

➔ 특히 낮은 차원에서 효과적(일반화 관점)

SVD/Slim. Net과 같은 차원 축소/서브 네트워크 방법론 적용의 경우, 다양한 크기로의 유연성을 제공하나, 정보 손실 불가피한 결과. 특히 저차원에서 크게 낮은 성능을 보임

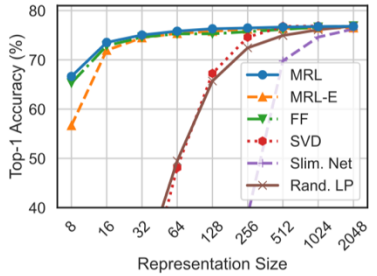


Figure 2: ImageNet-1K linear classification accuracy of ResNet50 models. MRL is as accurate as the independently trained FF models for every representation size.

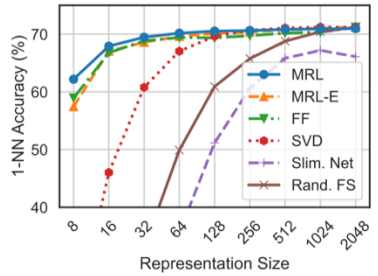


Figure 3: ImageNet-1K 1-NN accuracy of ResNet50 models measuring the representation quality for downstream task. MRL outperforms all the baselines across all representation sizes.

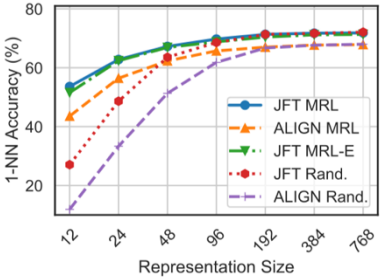


Figure 4: ImageNet-1K 1-NN accuracy for ViT-B/16 models trained on JFT-300M & as part of ALIGN. MRL scales seamlessly to web-scale with minimal training overhead.

◆ Matryoshka Representation Learning

Adaptive Classification (AC)

(a) Policy

- 표현 크기 m_i에서 m_{i+1}로 증가시키는 정책
  - 예측 신뢰도 p_i가 학습된 임계값 t_i^*를 초과하는지 여부에 따라 결정
  - $p_i \geq t_i^*$이면 현재 표현 크기 m_i를 사용하고, 그렇지 않으면 m_{i+1}

- 임계값 결정:
  - 10K ImageNet-1K validation set 활용하여 grid search

➔ 간단한 입력에 대해서는 적은 자원만 사용하고, 복잡한 입력에 대해서는 더 많은 자원을 사용하기 위함

(b) Adaptive Classification on MRL ResNet50 (Figure 6)

- MRC-AC는 512d-FF보다 동일한 accuracy (76.3)을 유지하면서 약 37차원의 표현만을 사용 → 14배 작은 벡터만을 사용해서 유사한 성능

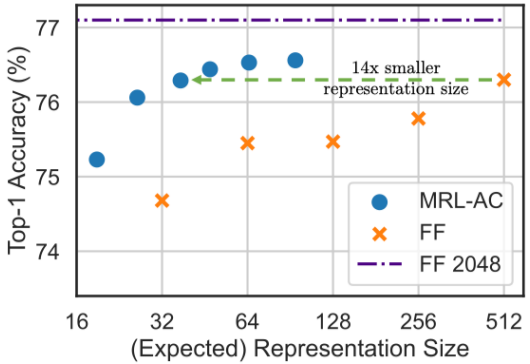


Figure 6: Adaptive classification on MRL ResNet50 using cascades results in 14× smaller representation size for the same level of accuracy on ImageNet-1K (∼ 37 vs 512 dims for 76.3%).

◆ Matryoshka Representation Learning

Image Retrieval

- Query와 same class인 images를 retrieval하는 task

(a) mAP@10 for Image Retrieval (Figure 7)
- 저차원($d \leq 256$)에서 SVD 및 Slim. Net은 성능이 크게 떨어지나, MRL은 일관되게 높은 검색 성능을 보임

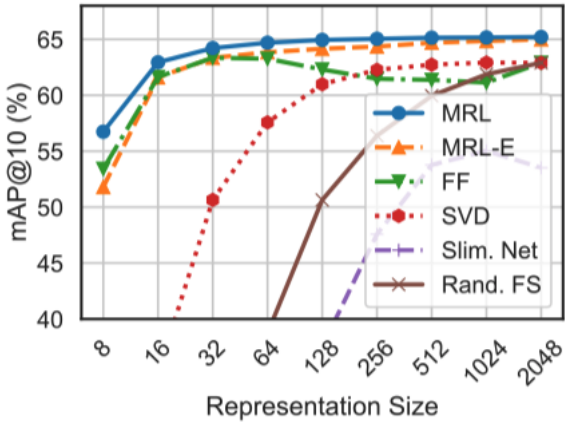


Figure 7: mAP@10 for Image Retrieval on ImageNet-1K with ResNet50. MRL consistently produces better retrieval performance over the baselines across all the representation sizes.

◆ Matryoshka Representation Learning

Adaptive Retrieval (AR)

- Given query image에 대해 저차원 표현(e.g., $d = 16$)을 사용하여 shortlisting
- 이후 고차원 표현($e.g., d = 2048$)로 reranking하는 방식
- Top rank에 대한 성능이 중요한 real-world 시나리오에서 효율성을 개선하기 위한 방법론

(a) ImageNet-1K (Figure 8 좌측)

- $D_s = 16$ & $D_r = 2048$을 사용한 MRL-AR 모델은 $d = 2048$인 단일 검색 모델과 동일한 정확도를 보이면서 이론적으로 128배, 실제로 14배 더 효율적인 결과를 보였음

(b) ImageNet-4K (Figure 8 우측)

- 분류 class 범주가 커져서 난이도가 증가함에 따라, $D_s = 64$가 필요하나, 여전히 이론적으로 약 32배, 실제로 약 6배 효율적인 결과를 보였음

(c) ImageNet-4K (Figure 8 Funnel)

- D_s와 D_r 선택의 어려움 → 계단식 검색 (16->32->64->128->256->2048)을 각각 사용하면서 reranking하고, 각 pool을 (200->100->50->25->10)으로 줄이는 방법론
- Funnel scenario에서도 여전히 이론적으로 128배 효율적임

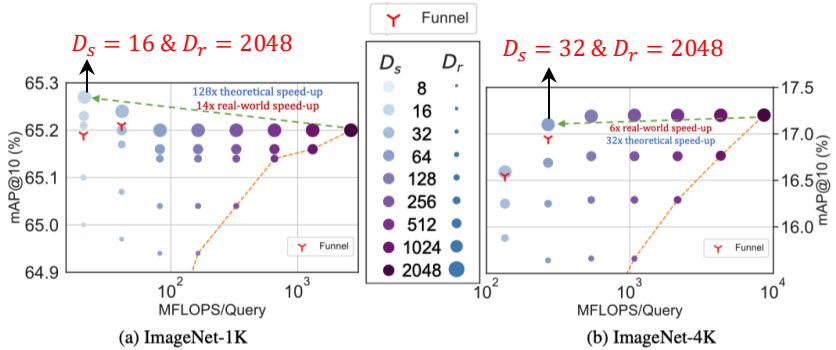


Figure 8: The trade-off between mAP@10 vs MFLOPs/Query for Adaptive Retrieval (AR) on ImageNet-1K (left) and ImageNet-4K (right). Every combination of D_s & D_r falls above the Pareto line (orange dots) of single-shot retrieval with a fixed representation size while having configurations that are as accurate while being up to 14× faster in real-world deployment. Funnel retrieval is almost as accurate as the baseline while alleviating some of the parameter choices of Adaptive Retrieval.

◆ Matryoshka Representation Learning

Disagreement across Dimensions

(a) Plastic Bag
- 8차원에서 샤워 캡으로 예측 되나, 차원이 증가하면서 잘 예측함

(b) Rock Python
- 8차원에서 같은 상위 클래스인 Boa Constrictor로 잘못 예측되었으나, 고차원에서 잘 예측함

(c) Sweatshirt
- 8, 16차원에서 인형의 눈(선글라스)에 초점을 맞추어 잘못 예측하였으나, 고차원에서 잘 예측함

➔ 고차원으로 갈수록 더 어려운 예측이 가능하며, 저차원에서는 상대적으로 단순한 작업에 활용 가능. 작업의 난이도에 따라서 선택적으로 활용할 수 있음

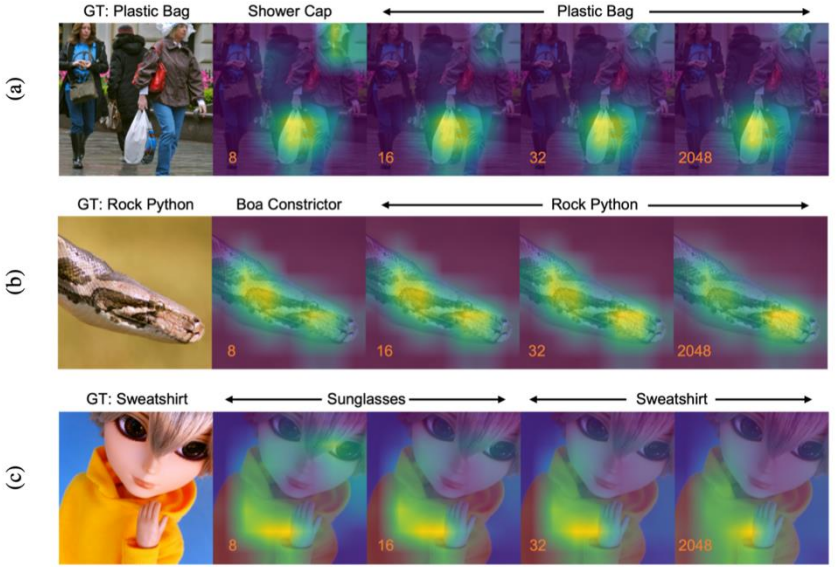


Figure 9: Grad-CAM [80] progression of predictions in MRL model across 8, 16, 32 and 2048 dimensions. (a) 8-dimensional representation confuses due to presence of other relevant objects (with a larger field of view) in the scene and predicts "shower cap" ; (b) 8-dim model confuses within the same super-class of "boa" ; (c) 8 and 16-dim models incorrectly focus on the eyes of the doll ("sunglasses") and not the "sweatshirt" which is correctly in focus at higher dimensions; MRL fails gracefully in these scenarios and shows potential use cases of disagreement across dimensions.

◆ Contextual Document Embeddings

CONTEXTUAL DOCUMENT EMBEDDINGS

John X. Morris
Cornell University
jxm3@cornell.edu

Alexander M. Rush
Cornell University
arush@cornell.edu

ICLR 2025 Submit (8/6/6)

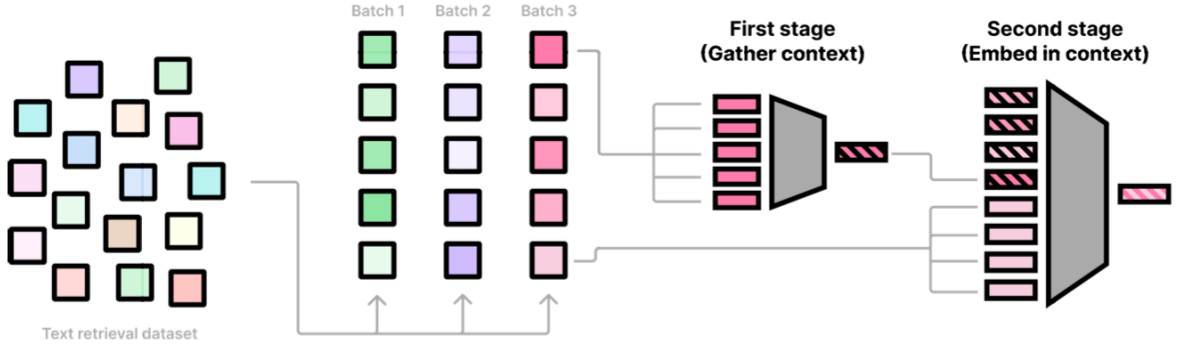
◆ Contextual Document Embeddings

Abstract

- 기존 Document에 대한 Dense Embeddings는 문서 자체만을 독립적으로 인코딩하므로, 특정 검색 맥락에서 발생하는 문맥 의존성 부족 문제가 있음
  - "The National Football League Draft is an annual event in which the National Football League (NFL) teams select eligible college football players..."
  - **검색 의도**에 따라서 **같은 문서**도 **다른 정보**가 요구될 수 있음
    - 스포츠 도메인: "NFL", "Draft", "football"
    - 방송/뉴스: "Annual Event"
  - 단일 문서가 아닌 **문서 간의 관계와 맥락**이 중요한 요소임
    - *예) 의료 문서 검색: 특정 증상에 대한 정보를 검색할 때, 해당 증상을 설명하는 문서뿐 아니라, **관련 질병, 치료법 등이 포함된 이웃 문서 정보**도 중요함*
- 본 논문에서는 문서 임베딩에 문서 의존성을 추가하여 특정 도메인과 맥락에서의 성능을 개선하는 방법론을 제안함

◆ Contextual Document Embeddings

Proposal Overview



Contextual batching partitions a dataset of documents and queries into batches that share similar context.

Contextual embedding produces an embedding for text that incorporates corpus-level information.

- **Contextual Training**: 문서 임베딩에서 인접 문서의 개념을 Contrastive Learning 학습 과정에 통합하는 방법론
  - 질의-문서 클러스터링을 활용하여 각 학습 배치에 대한 보다 어려운 이웃 그룹(In-batch negatives)을 샘플링

- **Contextual Embedding Architecture**: 기존 임베딩 아키텍쳐에 문서의 맥락 정보를 주입하는 방법론
  - Two-stage Encoding을 위한 두 개의 독립적인 인코더를 활용, 첫 번째 단계에서 맥락 정보를 인코딩한 뒤, 두 번째 인코더에 질의의 프롬프트로 활용
  - 추가 저장 공간이나 검색 프로세스의 변경 없이 기존의 bi-encoder 구조에 통합 가능

◆ Contextual Document Embeddings

Contextual Training with Adversarial Contrastive Learning

- **Contextual Training의 필요성**: 'NFL'이라는 단어가 일반적인 훈련 데이터셋 $\mathcal{D}_T$에 드물게 나타나며, 테스트 시점에서 $\mathcal{D}$가 스포츠 기사로 구성된 데이터셋이면, 'NFL'은 테스트셋에서는 매우 흔하게 등장할 것임. → 평가가 통계적으로 $\mathcal{D}_T$에 적대적

  - 이를 해결하기 위해 쿼리-문서 쌍을 독립적으로 샘플링하는 대신, 먼저 도메인을 샘플링한 후 예제를 샘플링하는 방법론 등이 효과적임이 입증되었음

  - General Text Embeddings (GTE)[1] 에서는 언어/도메인 subsets으로부터 multinomial distribution 기반의 샘플링을 적용하여 이러한 차이를 줄이기 위한 시도를 한 바 있음

  - BGE-M3[2], NV-Embed[3], SFR-Embeddings[4]은 in-batch negatives를 수행할 때, different tasks/granularities 가 same batch에 샘플링되는 것을 방지하여 homogeneous batching을 보장하고, 이를 통해 효율성/효과성을 확보하였음

1. Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., & Zhang, M. (2023). Towards general text embeddings with multi-stage contrastive learning. arXiv preprint arXiv:2308.03281.
2. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv preprint arXiv:2402.03216.
3. Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B., & Ping, W. (2024). NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. arXiv preprint arXiv:2405.17428.
4. Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. Sfrembedding-mistral: enhance text retrieval with transfer learning. Salesforce AI Research Blog, 3, 2024b.

◆ Contextual Document Embeddings

Contextual Training with Adversarial Contrastive Learning

- Training data $\mathcal{D}_T$를 a self-similar pseudo-domain groups $(\mathcal{B}^1, \ldots, \mathcal{B}^B)$로 분할

$$\max_{\phi,\psi} \sum_b \sum_{(d,q)\in\mathcal{B}^b} \log p(d \mid q) = \max_{\phi,\psi} \sum_b \sum_{(d,q)\in\mathcal{B}^b} \log \frac{\exp f(d,q)}{\sum_{(d',\cdot)\in\mathcal{B}^b} \exp f(d',q)}$$

- Groups가 가능한 challenging하도록 구성하기 위해 다음과 같은 가장 어려운 조합을 찾는 최적화 문제로 formalize:

$$\max_{(\mathcal{B}^1,\ldots\mathcal{B}^B)} \sum_b \sum_{\substack{(d,q)\in\mathcal{B}^b \\ (d',q')\in\mathcal{B}^b}} f(d,q') + f(d',q) = \max_{(\mathcal{B}^1,\ldots\mathcal{B}^B)} \sum_b \sum_{\substack{(d,q)\in\mathcal{B}^b \\ (d',q')\in\mathcal{B}^b}} \phi(d)\cdot\psi(q') + \phi(d')\cdot\psi(q) \quad (2)$$

- 비현실적인 계산.. 근사하기 위한 다른 방법 고안:
  - 기하학적으로 Dot product와 L2 norm은 normalized space에서 반비례 관계이므로
  - 수식 (2)을 최대화하는 것 대신 L2 distance를 최소화하는 방식으로 최적화
  - $m\big((d,q),(d',q')\big) = \|\phi(d) - \Phi(q')\| + \|\phi(d') - \Phi(q)\|$, m(a,b)는 L2 distance를 활용하여 a와 b사이에 유사도가 높은 샘플을 찾아내고자 함

Contextual Document Embeddings

Contextual Training with Adversarial Contrastive Learning

- Triangle inequality에 따라, $m\big((d,q),m\big)+m\big(m,(d',q')\big)\geq m((d,q),(d',q'))$ 가 성립하므로, 수식 (2)의 최대화 문제를 다음과 같은 최소화 문제로 간소화하여 정의할 수 있음

$$\min_{\substack{(\mathcal{B}^1,\dots\mathcal{B}^B)\\(m^1,\dots,m^B)}} \sum_b \sum_{(d,q)\in\mathcal{B}^b} m((d,q),m^b) \qquad (3)$$

  - m^b는 $\mathcal{B}^b$ domain (cluster)에 대한 centroid
  → 샘플 간 모든 쌍의 거리를 직접 계산하지 않고, centroid와의 거리 계산만으로 근사

  - 기존 임베딩 모델인 GTR을 활용하여 Cluster Embeddings ($\mathcal{B}^b$)를 계산함

◆ Contextual Document Embeddings

Filtering False Negatives

- GIST와 유사하게, Teacher model을 사용하여 샘플간 유사도를 미리 계산하고 이를 in-batch negatives시 고려함

$$\log p(d \mid q) = \frac{\exp f(d,q)}{\exp f(d,q) + \sum_{d' \notin S(q,d)} \exp f(d',q)} \tag{4}$$

- $S(q,d) = \{d' \in \mathcal{D} | f(q,d') \geq f(q,d)\}, f$ to be a simple pre-trained teacher embedding model
- 이 방법은 True negatives에 대한 Over-pruning을 초래할 수 있지만, 모델 성능에 중요한 역할을 한다고 함

◆ Contextual Document Embeddings

Packing

- 이러한 방법으로 생성된 Clusters의 크기가 일정하지 않아서 다음의 문제가 발생할 수 있음
  - 큰 클러스터는 batch size를 초과할 수 있음
  - 작은 클러스터는 batch size보다 작아서 샘플 수가 부족할 수 있음

➔ ***균일한 크기의 batch를 packing하는 방법 필요***

- 큰 클러스터의 경우 랜덤하게 나누거나 적절히 샘플을 분할하여 여러 작은 배치로 구성
- 작은 클러스터의 경우 유사한 클러스터와 병합하여 배치 크기를 조정
- 여러 에포크를 학습할 때 클러스터 구성을 매번 무작위로 조정하여 다양성 보장
- 클러스터 결합에는 Greedy method (Greedy Cluster-Level Traveling Salesman)[1]을 적용

1. Shi, W., Min, S., Lomeli, M., Zhou, C., Li, M., Szilvasy, G., ... & Lewis, M. (2023). In-context pretraining: Language modeling beyond document boundaries. arXiv preprint arXiv:2310.10638.

◆ Contextual Document Embeddings

Contextual Document Embedding (CDE)

- Two-stage process를 통해 contextualized embeddings 계산:
  - **Gather and embed context (first stage):**
    context 문서 $d^1, \ldots, d^J \in \mathcal{D}$에 대하여 firsts stage encoder M_1을 활용하여 $M_1(d^1), \ldots, M_1(d^J)$ 시퀀스로 인코딩
  - **Embed document with additional context tokens (second stage):**
    $\phi(d'; \mathcal{D}) = M_2\big(M_1(d^1), \ldots, M_1(d^J), E(d'_1), \ldots, E(d'_T)\big)$,
    M_2는 second-stage encoder, E는 token embedding matrix of M_2
- 질의 임베딩:
  - 질의 임베딩 계산은 문서 임베딩과 유사하지만, test time에서는 질의 문맥 정보를 사용할 수 없으므로 문서 문맥만을 활용
    $\phi(q; \mathcal{D}) = M_2\big(M_1(d^1), \ldots, M_1(d^J), E(q_1), \ldots, E(q_T)\big)$

◆ Contextual Document Embeddings

Contextual Document Embedding (CDE) 중간 점검

- Q1) 그러면 context document $d^1, \ldots, d^J$는 어떤 문서로 정의되는가?
  - *직접적으로 언급되어있진 않으나, Contextual Training with Adversarial Contrastive Learning 단계에서 가장 관련성이 높은 문서들이 context document로 사용될 것 같다는 추론이 가능함*

- Q2) context document는 문서마다/클러스터마다 다른가?
  - *Q1 답변에 적은 직관에 따르면 문맥 문서는 문서마다 다를 것임*

◆ Contextual Document Embeddings

Contextual Document Embedding (CDE) 학습 전략

- **효율성 최적화**:
  - 배치 내에서 문맥 문서를 공유하여 모든 샘플마다 개별적으로 문맥 정보를 재계산하지 않음
  - Context embedding $M_1(d_1), \ldots, M_1(d_J)$는 한 번 계산된 후 배치 내 모든 문서에서 재사용함

- **Embedding without context**:
  - 문맥 정보가 없거나 사용 불가능한 경우의 generalizability 개선을 위한 전략
  - Sequence dropout: randomly replace context embeddings $M_1(d^*)$ with some null token $v_\emptyset$ according to some a uniform probability p.

- **Position-agnostic embedding**:
  - Context document는 순서가 없으므로, 위치 정보를 제거함

◆ Contextual Document Embeddings

Experimental Setup

- **Small setting**:
  - 6-layer transformers with maximum sequence length of 64 + 64(context tokens)
  - Evaluated on a truncated version of the BEIR benchmark
  - A variety of batch sizes in $\{256, 512, 1024, 2048, 4096\}$ and cluster sizes $\{64, 256, 1024, 4096, \ldots, 2097152, 4194304\}$
  - Typical embedding models의 학습 시나리오에 따라, two phases의 학습 적용
    1) a large weakly-supervised pre-training phase and 2) a short-supervised phase

- **Large setting**:
  - BERT-based model (**NomicBERT**) on sequences of length 512 + 512 (contextual tokens)
  - Evaluated on the full MTEB benchmark

- **Partitioning dataset**:
  - 데이터셋을 batch로 나눌 때 GTR 모델을 사용하여 인코딩하며, FAISS를 활용하여 클러스터링을 적용하였음
  - 도메인별로 100steps clustering을 수행, 총 3번의 시도 중 최상의 clustering을 선택

◆ Contextual Document Embeddings

Experimental Setup

- **Training**:
  - M_1 and M_2를 모두 NomicBERT로 초기화하여 학습
  - M_1의 경우 512 length로 학습, M_2는 M_1의 512개 embeddings를 prompt로 사용하므로, 1024 length로 학습

◆ Contextual Document Embeddings

Experimental Result

	Clssfctn	Cluster	PairCls	Rerank	Retrvl	STS	Summ.	Mean
nomic-embed-v1	74.1	43.9	85.2	55.7	52.8	82.1	30.1	62.39
stella-base-en-v2	75.3	44.9	86.5	58.8	50.1	83.0	32.5	62.61
bge-base-en-v1.5	75.5	45.8	86.6	58.9	53.3	82.4	31.1	63.56
GIST-Embedding-v0	76.0	46.2	86.3	59.4	52.3	83.5	30.9	63.71
gte-base-en-v1.5	77.2	46.8	85.3	57.7	54.1	82.0	31.2	64.11
anon-model-v1								
[Random]	81.3	46.6	84.1	55.3	51.1	81.4	31.6	63.81
[Contextual]	81.7	48.3	84.7	56.7	53.3	81.6	31.2	**65.00**

1. [Random] 대비 [Contextual]의 높은 성능

Table 2: Performance of models with 250M or fewer parameters on the MTEB benchmark for text embedding models. "Random" indicates the performance of our model with random training documents included instead of per-domain contextual documents.

2. Batching 전략/Architecture 유무에 따른 성능 변화

Contextual Batch	Contextual Arch	Batch Size	Cluster Size	Train loss	Train acc.	NDCG@10
		16384	-	0.39	90.3	59.9
✓		512	512	0.81	77.7	61.7
	✓	16384	-	0.37	90.7	62.4
✓	✓	512	512	0.68	80.9	**63.1**

Table 1: Performance of our small models with and without the two improvements proposed in this paper, measured on a shortened version of the BEIR benchmark. Numbers are NDCG@10.

3. Loss가 높을 수록 더 높은 성능 관측 (상관관계O)

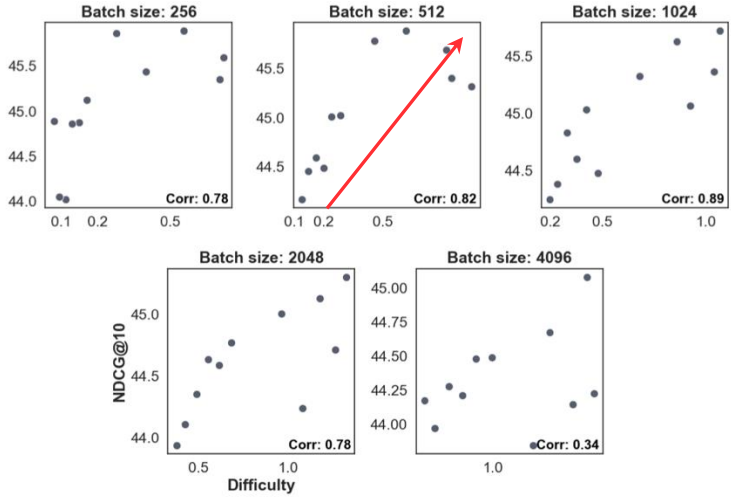


Figure 2: Performance vs. average batch difficulty (as measured by loss at the end of pre-training and supervised training) across batch sizes, after supervised contrastive training. Within a given batch size, we observe a clear increase in performance by making individual batches harder. Correlations are Pearson.

◆ Contextual Document Embeddings

How Hard are the clusters?

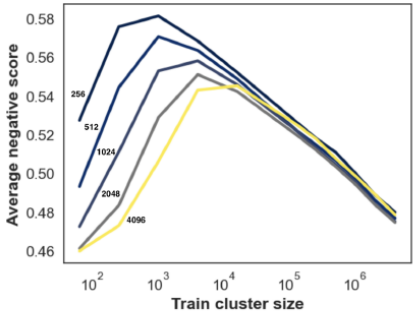


Figure 6: Average difficulty of in-batch negatives as measured by a surrogate model as cluster size and batch size change.

- Larger batch sizes
→ bring easier non-negative examples

- Decreasing cluster size
→ increases the average hardness of negative examples in a given cluster.

Which contextual documents help?

X-axis: test domain
Y-axis: input domain

각 도메인에 대한 contextual document를 활용하는 것이 가장 좋음

일부 domain 에서는 cross-over interactions 관측됨

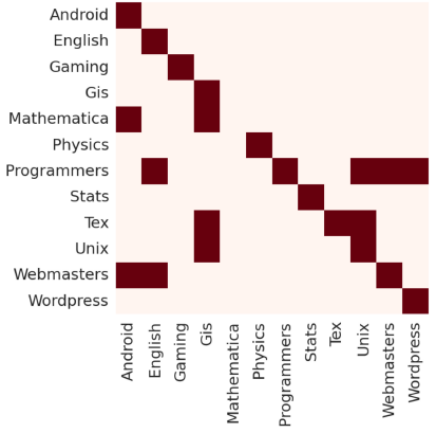


Figure 7: Impact of context by testing our model with different Stackexchange forum input types. Y-axis indicates the input domain, X-axis indicates the test domain. Dark squares come within one point NDCG@10.

◆ Contextual Document Embeddings

결론 및 생각 정리

- Large batch size를 고집하는 최근 Embedding training paradigms 에 충격적인 실험 결과 (batch_size=512)
- In-batch negatives 내 샘플링되는 examples가 매우 중요함. 각 문서마다의 인접 문서 정보? 와 같은 부가적 정보를 최대한 활용하여 이들의 품질을 개선하는 것이 단순히 hard negatives의 수를 늘리고 batch size를 키우고 하는 작업보다 중요할 수도 있겠다는 생각이 듦

감사합니다.

Copyright© 2022 Natural Language Processing & Artificial Intelligence Lab