NV-EMBED: IMPROVED TECHNIQUES FOR TRAINING LLMS AS GENERALIST EMBEDDING MODELS

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, Wei Ping

ICLR 2025

Presented by:

Usneek Singh & Parth Nanda

Continuing from embeddings discussion.....

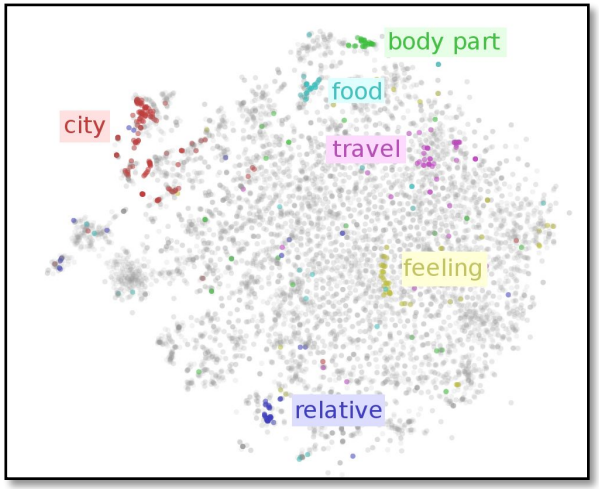
1. Why we need embeddings?

Retrieval, classification, clustering, etc..

2. Benchmarks: MTEB, MMTEB, AIR

3. Techniques: Traditional, Bidirectional, Decoder-only

Examples: word2vec, bag-of-words



Bidirectional Embedding Models

- Attend to tokens on both sides

Input: I love <mask> Tech ; *Output:* Georgia

Input: I love Georgia Tech. <sep> It is the best school ; *Output:* Entailed

- Mean Pooling across all layers to get single embedding.

Examples: BERT, T5

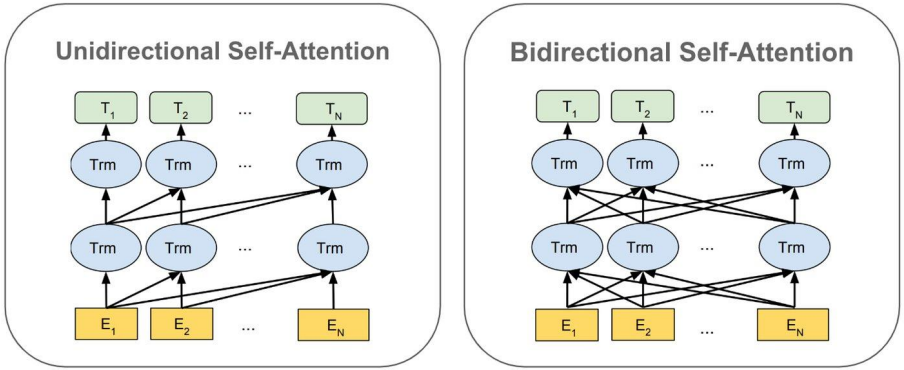
Decoder Embedding Models

1. Problems:
   A)Limited learning capability.
   B)Curse of dimensionality

Ideas:

1. The embeddings from last token <eos> from last layer.
2. Contrastive Training, e.g E5-Mistral
3. Instruction Tuning, e.g. SFR-Embedding-Mistral
4. Knowledge Distillation, eg. Gecko

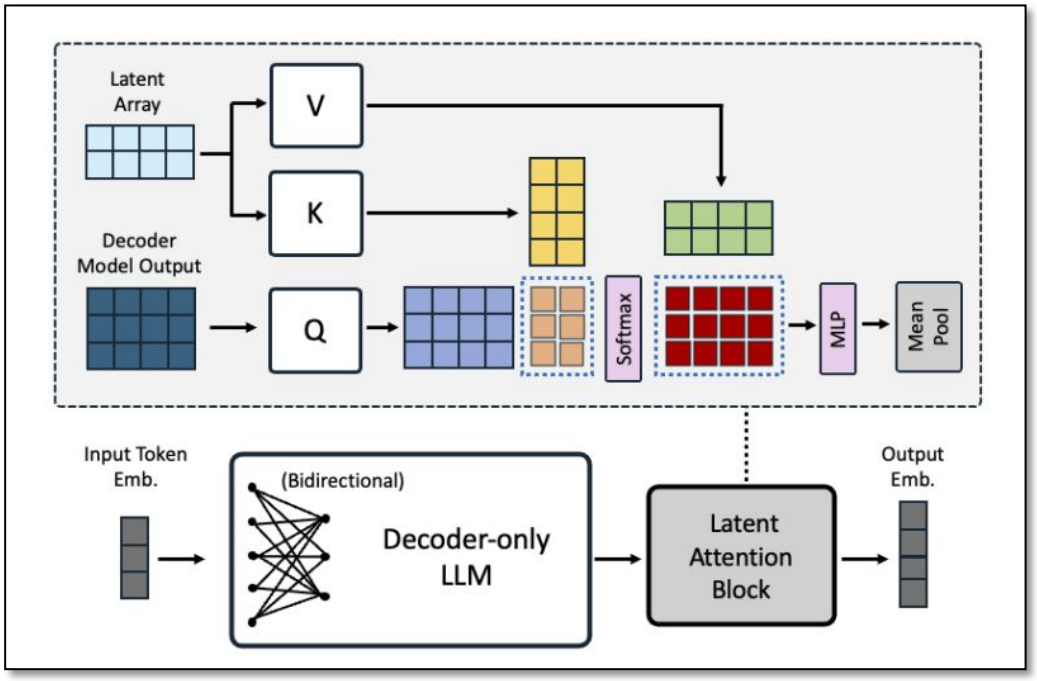
Concept 1: Bidirectional Attention



Modify the attention block to look in both directions.

attn_out = attention(q, k, v, attn_mask=padding_mask, causal=False)

Concept 2: Latent Attention Layer



Decoder Last layer

Query: Q = ℓ×d
Latent Key: K = r×d
Latent Value: V = r×d

QK^T : (ℓ×d) · (d×r) = ℓ×r

A = softmax(QK^T / √d) : ℓ×r

O = AV : (ℓ×r) · (r×d) = ℓ×d

Concept 3: Two-Stage Instruction Training

Instruction Tuning for Retrieval tasks:
In a batch of size **B**, each query's positive is paired with **B−1** other passages in the batch as **negatives**.

What will happen if we use the same trick for classification?
X1: A,
x2: B,
x3: A,
x4: B
What will be the hard negatives for x1 ?
[B,A]

Split into two-phases:

1. Retrieval only Tasks (with in-Batch Training)
2. Retrieval and Non-Retrieval tasks (without in-Batch Training)

Concept 3: Two-Stage Instruction Tuning

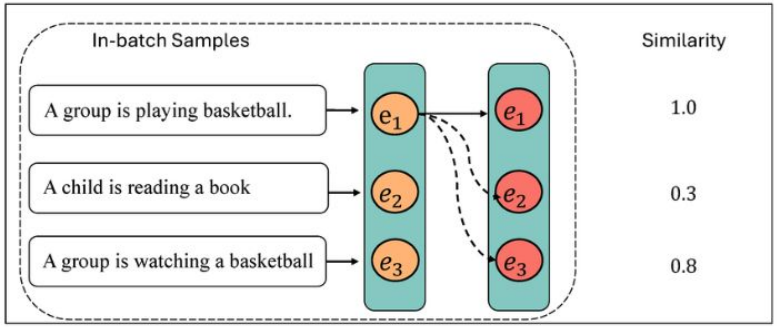
- How the training works in Stage 2 ? (without in-batch)

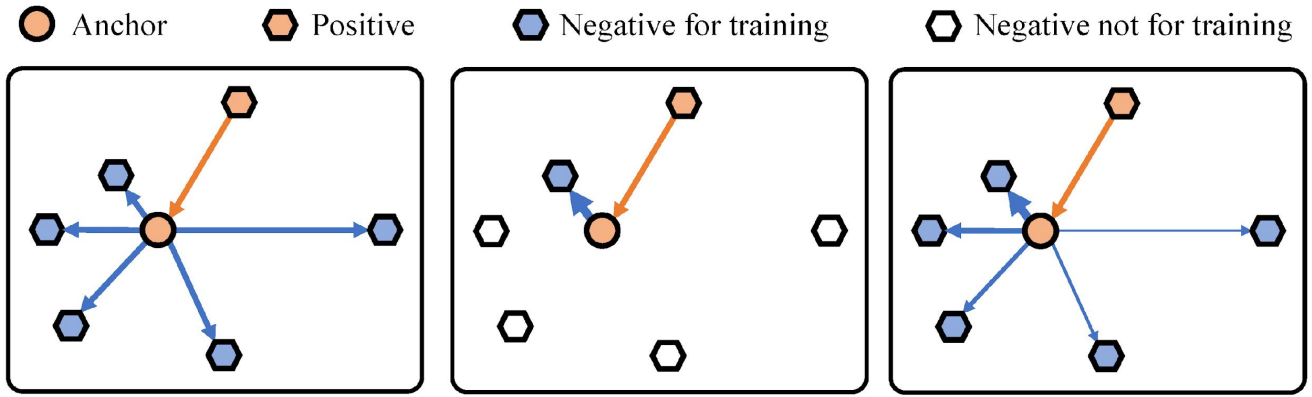
Task	**Query**	**Positive Class**	**Negative Class**
Retrieval	q	Ground truth d+	Hard negative mined documents
Classification	q	Examples from the same class d+	Examples from other classes d-
Similarity	q	High similarity score d+	Hard negative mined d-

Concept 4: Hard negative Mining Technique

Use a teacher model to get relevance scores of other documents (E5-mistral-7b-instruct)

Determine a threshold based on the set percent margin
max_negative_score_threshold = pos_score * percentage_mar





Training Dataset

Retrieval: MSMARCO, HotpotQA, Natural Question, PAQ, Stack Exchange, SQuAD, ArguAna, BioASQ, FiQA, FEVER, HoVer, SciFact, NFCorpus, MIRACL, and Mr.TyDi.

Classification: Amazon (Reviews, Polarity, Counterfactual), IMDB, Banking77, Emotion, MTOP, Massive, ToxicConversations, and TweetSentimentExtraction.

Clustering: Arxiv, Biorxiv, Medrxiv, TwentyNewsgroups, Reddit, and StackExchange.

Semantic Textual Similarity (STS): STS12, STS22, and STS-Benchmark

---

Synthetic Dataset: Mixtral-Instruct

Task Name	Instruction Template	Number of Samples
ArguAna	Given a claim, retrieve documents that support or refute the claim	16k
Natural Language Inference	Retrieve semantically similar text Given a premise, retrieve a hypothesis that is entailed by the premise	270k
PAQ, MSMARCO	Given a web search query, retrieve relevant passages that answer the query Given a question, retrieve passages that answer the question Given a question, retrieve documents that can help answer the question	500k, 500k
SQUAD	Given a question, retrieve passages that answer the question	87k
StackExchange	Given a web search query, retrieve relevant passages that answer the query	80k
Natural Question	Given a question, retrieve passages that answer the question	100k
HotpotQA	Given a multi-hop question, retrieve documents that can help answer the question	170k
FEVER	Given a claim, retrieve documents that support or refute the claim	140k
FiQA2018	Given a financial question, retrieve relevant passages that answer the query	5k
BioASQ	Given a query, retrieve documents that can help answer the question	2.4k
HoVer	Given a claim, retrieve documents that support or refute the claim	17k
Nfcorpus	Given a question, retrieve relevant documents that answer the question	3.6k
MIRACL	Given a question, retrieve passages that answer the question	2k
Mr.TyDi	Given a question, retrieve passages that answer the question	2k
SciFact	Given a scientific claim, retrieve documents that support or refute the claim	0.9k
STS12, STS22, STSBenchmark	Retrieve semantically similar text.	1.8k, 0.3k, 2.7k
AmazonCounterfactual-Classification	Classify a given Amazon customer review text as either counterfactual or not-counterfactual	6k
AmazonPolarity-Classification	Classify Amazon reviews into positive or negative sentiment	20k
AmazonReviews-Classification	Classify the given Amazon review into its appropriate rating category	40k
Banking77-Classification	Given a online banking query, find the corresponding intents	10k
Emotion-Classification	Classify the emotion expressed in the given Twitter message into one of the six emotions:anger, fear, joy, love, sadness, and surprise	16k
Imdb-Classification	Classify the sentiment expressed in the given movie review text from the IMDB dataset	24k
MTOPIntent-Classification	Classify the intent of the given utterance in task-oriented conversation	15k
MTOPDomain-Classification	Classify the intent domain of the given utterance in task-oriented conversation	15k
MassiveIntent-Classification	Given a user utterance as query, find the user intents	11k
MassiveScenario-Classification	Given a user utterance as query, find the user scenarios	11k
ToxicConversationsClassification	Classify the given comments as either toxic or not toxic	50k
TweetSentimentExtractionClassification	Classify the sentiment of a given tweet as either positive, negative, or neutral	27k
Arxiv-Clustering-P2P	Identify the main and secondary category of Arxiv papers based on the titles and abstracts	50k
Arxiv-Clustering-S2S	Identify the main and secondary category of Arxiv papers based on the titles	50k
Biorxiv-Clustering-P2P	Identify the main category of Biorxiv papers based on the titles and abstracts	15k
Biorxiv-Clustering-S2S	Identify the main category of Biorxiv papers based on the titles	15k
Medrxiv-Clustering-P2P	Identify the main category of Medrxiv papers based on the titles and abstracts	2.3k
Medrxiv-Clustering-S2S	Identify the main category of Medrxiv papers based on the titles	2.3k
Reddit-Clustering	Identify the main category of Medrxiv papers based on the titles and abstracts	50k
Reddit-Clustering-S2S	Identify the main category of Medrxiv papers based on the titles and abstracts	40k
Stackexchange-Clustering	Identify the main category of Medrxiv papers based on the titles and abstracts	50k
Stackexchange-Clustering-S2S	Identify the main category of Medrxiv papers based on the titles and abstracts	40k
TwentyNewsgroups-Clustering	Identify the topic or theme of the given news articles	1.7k

***Note**: MTEB test examples were removed from train*

Summary of Design

- **Changes in Model architecture**

  A) Latent Attention Layer

  B) Bidirectional Attention

- **Changes in training phase**

  A) Two stage Training

----------NV-Embed-v1----------

- B) Positive aware Hard negative using an expert model

- **Curation of training dataset**

Adding Retrieval datasets, Example based multi class labels

- **New Synthetic Training Datasets**

60000 synthetic tasks using Mistral-instruct model

----------NV-Embed-v2----------

-

MMTB Leader Board

Embedding Task Mertric	Retrieval (15) nDCG@10	Rerank (4) MAP	Cluster. (11) V-Meas.	PairClass. (3) AP	Class. (12) Acc.	STS (10) Spear.	Summ.(1) Spear.	Avg. (56)
NV-Embed-v2	**62.65**	60.65	**58.46**	88.67	**90.37**	84.31	30.7	**72.31**
Bge-en-icl (zero shot)	61.67	59.66	57.51	86.93	88.62	83.74	30.75	71.24
Stella-1.5B-v5	61.01	61.21	57.69	88.07	87.63	84.51	31.49	71.19
SFR-Embedding-2R	60.18	60.14	56.17	88.07	89.05	81.26	30.71	70.31
Gte-Qwen2-7B-instruct	60.25	61.42	56.92	85.79	86.58	83.04	31.35	70.24
NV-Embed-v1	**59.36**	60.59	52.80	86.91	87.35	82.84	31.2	**69.32**
Bge-multilingual-gemma2	59.24	59.72	54.65	85.84	88.08	83.88	31.2	69.88
Voyage-large-2-instruct	58.28	60.09	53.35	89.24	81.49	84.58	30.84	68.28
SFR-Embedding	59.00	60.64	51.67	88.54	78.33	85.05	31.16	67.56
GritLM-7B	57.41	60.49	50.61	87.16	79.46	83.35	30.37	66.76
E5-mistral-7b-instruct	56.9	60.21	50.26	88.34	78.47	84.66	31.4	66.63
Text-embed-3-large (OpenAI)	55.44	59.16	49.01	85.72	75.45	81.73	29.92	64.59

- NV Embed achieves 72.31 average score
- Achieved highest accuracy for Retrieval, Clustering and Classification tasks

A1: 2-stage Training Process

Embedding Task	Retrieval	Rerank	Cluster.	PairClass.	Class.	STS	Summ.	Avg.
Single Stage (Inbatch Enabled)	61.25	60.64	57.67	87.82	86.6	83.7	30.75	70.83
Single Stage (Inbatch Disabled)	61.37	60.81	58.31	88.3	90.2	84.5	30.96	71.94
Two Stage Training	**62.65**	60.65	58.46	88.67	90.37	84.31	30.70	**72.31**
Reversed Two Stage	**61.91**	60.98	58.22	88.59	90.26	83.07	31.28	**71.85**

- Worse performance with single stage (in batch enabled)
- Best performance with two stage training.
- Reversed Two stage: Just change the order

A2: Causal vs Bidirectional Attention

Pool Type	EOS		Mean		Latent-attention		Self-attention	
Mask Type	bidirect	causal	bidirect	causal	bidirect	causal	bidirect	causal
Retrieval (15)	62.13	60.30	61.81	61.01	**62.65**	61.15	61.17	60.53
Rerank (4)	60.02	59.13	60.65	59.10	60.65	59.36	60.67	59.67
Clustering (11)	58.24	57.11	57.44	57.34	**58.46**	57.80	58.24	57.11
PairClass. (3)	87.69	85.05	87.35	87.35	88.67	87.22	87.69	85.05
Classification (12)	90.10	90.01	89.49	89.85	**90.37**	90.49	90.10	90.01
STS (10)	82.27	81.65	84.35	84.35	84.31	84.13	84.22	83.81
Summar. (1)	30.25	32.75	30.75	30.88	30.70	30.90	30.93	31.36
Average (56)	71.63	70.85	71.71	71.38	**72.31**	71.61	71.61	70.6

With almost all pooling methods and across all tasks, bidirectional attention mechanism gives higher performance than causal attention.

A3: Pooling Methods

Pool Type	EOS		Mean		Latent-attention		Self-attention	
Mask Type	bidirect	causal	bidirect	causal	bidirect	causal	bidirect	causal
Retrieval (15)	62.13	60.30	61.81	61.01	**62.65**	61.15	61.17	60.53
Rerank (4)	60.02	59.13	60.65	59.10	60.65	59.36	60.67	59.67
Clustering (11)	58.24	57.11	57.44	57.34	**58.46**	57.80	58.24	57.11
PairClass. (3)	87.69	85.05	87.35	87.35	88.67	87.22	87.69	85.05
Classification (12)	90.10	90.01	89.49	89.85	**90.37**	90.49	90.10	90.01
STS (10)	82.27	81.65	84.35	84.35	84.31	84.13	84.22	83.81
Summar. (1)	30.25	32.75	30.75	30.88	30.70	30.90	30.93	31.36
Average (56)	71.63	70.85	71.71	71.38	**72.31**	71.61	71.61	70.6

Latent Attention performs better than:

1. <EOS> token embedding : *Recency bias*
2. Mean of all layers
3. Self-attention (one more layer after the latent-attention)

A4: Classification Representation

Embedding Task	Retrieval	Rerank	Cluster.	PairClass.	Class.	STS	Summ.	Avg.
Label-based approach	62.40	59.7	53.04	88.04	89.17	84.25	30.77	70.82
Example-based approach	62.65	60.65	**58.46**	88.67	**90.37**	84.31	30.70	**72.31**

- Label-based Approach: Hard negatives are class labels
- Example-based approach: Hard negatives are examples from other classes
- Example based approach performs better

A5: Hard negative Mining/Synthetic Datasets

Embedding Task	Retrieval	Rerank	Cluster.	PairClass.	Class.	STS	Summ.	Avg.
[S0] Without HN, Without AD, Without SD	59.22	59.85	57.95	85.79	90.71	81.98	29.87	70.73
[S1] With HN, Without AD, Without SD	61.52	59.80	58.01	88.56	90.31	84.26	30.36	71.83
[S2] With HN, With AD, Without SD	62.28	60.45	58.16	88.38	90.34	84.11	29.95	72.07
[S3] **With HN, With AD, With SD**	62.65	60.65	58.46	88.67	90.37	84.31	30.70	**72.31**

- Hard negative mining helps.
- AD = Additional Public datasets, SD= synthetic datasets
- Using both kind of datasets improve performance

Conclusion

- NV-Embed: A decoder-only model better than bidirectional models for embedding tasks.
- Latent Attention Layer helps in more expressive pooling.
- Two-stage training process helps to handle variety of tasks.
- Good out-of-domain performance on other benchmarks.