Effective post-training embedding compression via temperature control in contrastive training

Georgiana Dinu[1] Corey Barrett[2] Yi Xiang[1] Miguel Romero Calvo[1] Anna Currey[1] Xing Niu[1]
[1]Amazon [2]Oracle

amazon | science

Motivation

- Embeddings power many applications, particularly retrieval and retrieval-augmented generation; storing and searching large embedding collections is computationally costly.
- As downstream requirements are often unknown during training, fixed-size embeddings may be suboptimal. Post-training compression enables embeddings to flexibly adapt to variable system constraints at inference time.

This paper

- Role of temperature in contrastive training for text embeddings
- Impact on post-training compression

Temperature in contrastive learning

- Temperature parameter (tau) modulates impact of difficult negative examples
- Low temperature: Enhances the influence of most difficult negatives

Temperature

$$L_{\text{InfoNCE}}(f,x,\tau) = \mathbb{E}_{(x,x^+)\sim p_{pos}}\left[-log\frac{e^{s(f(x),f(x^+))/\tau}}{e^{s(f(x),f(x^+))/\tau} + \Sigma_{x^-} e^{s(f(x),f(x^-))/\tau}}\right]$$

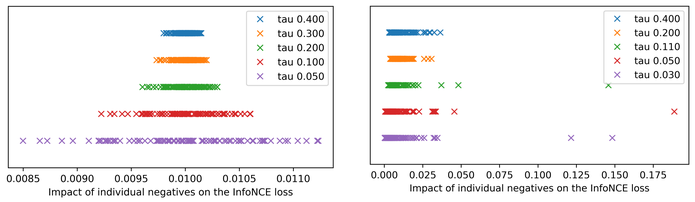


Figure 1: Each negative's contribution to InfoNCE normalization factor. Left: Random vectors (early training) and Right: mid-training vectors

Impact of temperature on tasks

Text embedding tasks:
- Retrieval: Measures relevance matching via ranking
- Clustering: Measures semantic similarity

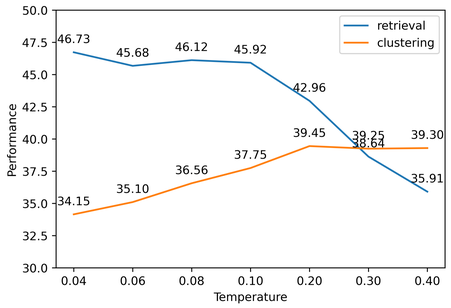


Figure 2: Performance trade-off across tasks: Retrieval performance decreases while clustering increases with larger temperatures.

- Hypothesis: high temperature produces a "more clustered" space, aiding clustering tasks
- Suggests interplay between capturing local and global structure under different temperature conditions

Intrinsic Dimensionality Analysis

Intrinsic dimensionality: The actual number of dimensions needed to represent the underlying structure of data. Calculated as the number of principal components needed to explain 95% of the variance (using PCA).

Key finding: Higher temperature → Lower intrinsic dimensionality (temp=0.04 requires 654 dimensions for 95% variance, 0.4 only needs 393)

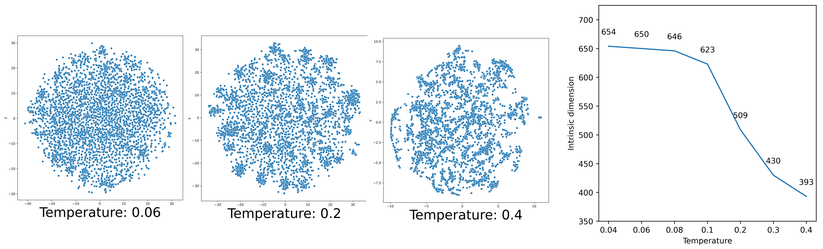


Figure 3: t-SNE projections of embeddings / temperature

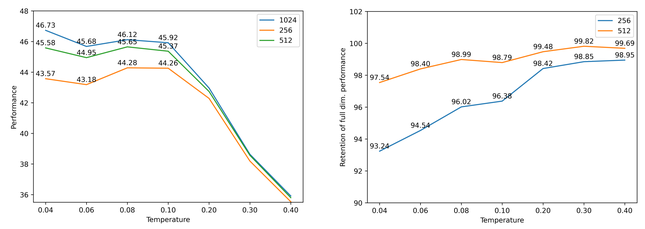
- Spaces with lower intrinsic dimensionality exhibit more correlation and redundancy between features.

More compact representations possible with lower intrinsic dimensionality, suggesting natural pathway for compression.

Post-training compression

Metrics:
- Absolute retrieval quality
- Quality retention: ratio of retrieval performance before and after compression

Method 1: Random Feature Selection (via vector truncation). Reduce dimensions from 1024 → 256/512. Up to 4x reduction in storage.



Method 2: Binarization (32-bit values quantized to 1-bit using sign function. Hamming distance instead of cosine similarity for faster and better retrieval. 32x reduction in storage.

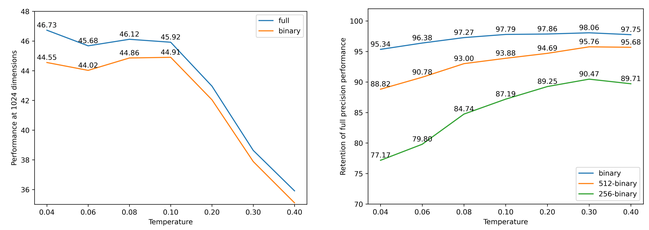


Figure 4. Vector truncation and binarization: Absolute quality drops / quality retention increases with larger temperatures

Temperature aggregation solutions

Goal: balance performance and compression objectives

Method 1: Plain Temperature Aggregation
- Sum of losses using different temperatures
- Use T=3 temperature values: [0.03, 0.06, 0.1] with uniform weights
- Covers best temperature range observed in experiments

$$L_{\text{TempAgg}}(f,x,\tau) = \Sigma_{t=1}^{T} w_t L_{\text{InfoNCE}}(f,x,\tau_t)$$

Method 2: Aggregation + MRL (Matryoshka Representation Learning)
- MRL Promotes embeddings that retain quality when truncated to smaller dimensions

$$L_{\text{MRL}}(f,x,\tau) = \Sigma_{i=1}^{k} \lambda_i L_{\text{InfoNCE}}(f_{[1:d_i]},x,\tau)$$

- Apply temperature aggregation at each truncated dimensionality

$$L_{\text{TempAggMRL}}(f,x,\tau) = \Sigma_{i=1}^{k} \lambda_i L_{\text{TempAgg}}(f_{[1:d_i]},x,\tau)$$

Method 3: Temperature Specialization + MRL
- Uses observation that lower temperatures are better for retrieval, especially for reduced dimensions
- Employs MRL loss with temperatures: {256:0.03, 512:0.06, 1024:0.1}
- Low temperature at 256 dimensions optimizes for retrieval. High temperature at 1024 provides aggregation effect and benefits clustering

	$\tau = 0.04$	$\tau = 0.1$	TempAgg	TempAggMRL	TempSpecMRL
Ret. (full)	**46.7**	45.9	46.4	46.6	46.0
Clust. (full)	34.1	**37.7**	36.4	36.9	37.4
Ret. 256	43.6	44.2	44.6	**45.1**	44.6
Ret. bin	44.6	44.9	45.2	**45.2**	44.8
Ret. bin re-rnk	45.6	45.6	45.9	**46.1**	45.5
Ret. 256/1024	93.2	96.4	96.2	96.9	**97.1**
Ret. bin/full prec	95.3	97.8	**97.4**	97.1	**97.4**
Ret. bin re-rnk /full prec	97.6	**99.0**	98.9	98.9	**99.0**
Intrinsic dim.	654	623	629	623	**617**

Table 1: Temperature aggregation matches performance of best single-temperature models while achieving compression quality retention of 97-99%

- Intrinsic dimensionality reduction correlates with compressibility

Key takeaways

- The temperature parameter in contrastive learning impacts the intrinsic dimensionality and the task-specific performance of resulting text embedding spaces
- Embedding spaces with lower intrinsic dimensionality enable more efficient post-training compression via truncation and binarization, but show lower absolute performance on retrieval tasks
- Temperature aggregation methods effectively balance retrieval performance and compressibility, achieving 99% retention with 32x compression

References

- Khosla et al. Supervised contrastive learning. NeurIPS 2020.
- Wang & Liu. Understanding the behaviour of contrastive loss. CVPR 2020. Kukleva et al. Temperature schedules for self-supervised contrastive methods on long-tail data. ICLR 2023.
- Kusupati et al. Matryoshka representation learning. NeurIPS 2022.

Conference ICLR

*Corresponding author: gddinu@amazon.com