Week 3 Lesson Plan

Contrastive loss

Asif Qamar

pepal leaves publisher

www.supportvectors.ai

First draft, February 14, 2026

Contents

Why Contrastive Learning?

Contents

Not every likelihood induces a geometry worth inhabiting.

— Musings from a morning hike

Orientation

Today we take a decisive turn.
We ask: why is next-token prediction (or more generally, negative log-likelihood) insufficient for shaping the geometry of semantic space? And why does contrastive learning—triplets, InfoNCE, SimCLR—feel like the right sculpting tool?
The day begins experimentally, not philosophically.

Opening Experiments (11:00–12:30)

Experiment 1: Vanilla BERT Embeddings

Task. Use pretrained `bert-base-uncased`. Extract the `[CLS]` embedding for short texts across dissimilar domains:

- Medical diagnosis snippets
- Legal clauses
- Poetry fragments
- Python docstrings

Compute cosine similarities across:

- same-domain pairs
- cross-domain pairs

Observation to record. Distribution overlap between intra-class and inter-class similarity.

You are measuring geometry, not accuracy.

Experiment 2: Contrastive-Trained Embeddings

Repeat with a contrastive model (e.g., Sentence-BERT or SimCSE).
Plot similarity histograms again.
Expected phenomenon. Tighter intra-class clusters. Greater margin between unrelated samples.

Discussion Prompt

Why does NLL not enforce separation? Why does contrastive loss actively carve margins?

Limitations of Negative Log-Likelihood

Negative log-likelihood (NLL) optimizes:

$$\mathcal{L}_{\text{NLL}} = -\sum_{t} \log p(x_t \mid x_{<t})$$

It ensures predictive competence. It does not ensure metric structure. Two sequences can be equally predictable yet occupy nearby embeddings despite semantic opposition.
NLL encourages correct continuation. It does not enforce:

Likelihood shapes probability. Contrastive shapes geometry.

- isotropy
- margin maximization
- cluster separation
- angular uniformity

The result: anisotropic embedding space.

The Case for Contrastive Loss

Contrastive learning introduces explicit relational constraints.

Triplet Loss

Given anchor a, positive p, negative n:

$$\mathcal{L}_{\text{triplet}} = \max\left(0, d(a,p) - d(a,n) + \alpha\right)$$

where α is a margin.
Triplet loss encodes:

You are enforcing geodesic inequality.

$$d(a,p) + \alpha < d(a,n)$$

It is a local geometric law.

From Triplets to InfoNCE

Triplets are sparse supervision.
InfoNCE generalizes:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \exp(\text{sim}(z_i, z_k)/\tau)}$$

τ is temperature. Lower τ sharpens curvature.

Instead of one negative, all other samples act as negatives.
This induces:

- Uniformity on the hypersphere
- Alignment of positives
- Repulsion of unrelated samples

Why SimCLR?

SimCLR [1] removes architectural complications. No memory banks. No momentum encoders.
Large batch = many negatives.
Core principle:

- Data augmentation produces positive pairs.
- All others are negatives.

It demonstrated scale alone could yield high-quality representations.

[1] Ting Chen et al. (2020). "A Simple Framework for Contrastive Learning of Visual Representations". In: *ICML*

Vision Transformers

The Vision Transformer (ViT) [2] proved attention is not modality-bound. Patchify image. Treat patches as tokens. Apply transformer encoder. Contrastive training + ViT = strong visual embeddings.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: **International Conference on Learning Representations (ICLR)**

CLIP: Language Meets Vision

CLIP [3] learns joint embedding space.
Loss:

- Image-to-text contrast
- Text-to-image contrast

Symmetric InfoNCE.
Result:

- Zero-shot classification
- Open-vocabulary recognition

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. (2021). "Learning Transferable Visual Models From Natural Language Supervision". In: **Proceedings of the 38th International Conference on Machine Learning (ICML)**

Joint geometry across modalities.

SigLIP and SigLip-*

SigLIP [4] replaces softmax normalization with independent sigmoid losses.

[4] Xiaohua Zhai et al. (2023). "Sigmoid Loss for Language-Image Pre-Training". In: *arXiv preprint arXiv:2303.15343*

Key difference: No need for large global batch normalization. More stable scaling behavior.
SigLip-* variants extend:

- multilingual alignment
- scaling efficiency
- higher resolution patch encodings

Afternoon Lab (2:00–4:00)

Lab 1: Visualizing Embedding Geometry

- PCA projection
- t-SNE / UMAP
- Compare BERT vs Contrastive model

Lab 2: Margin Sensitivity

Modify temperature τ. Observe cluster collapse vs dispersion.

Lab 3: Small Contrastive Fine-Tune

Fine-tune sentence encoder on domain-specific dataset. Measure improvement on retrieval.

Key Takeaways

- NLL optimizes probability, not geometry.
- Contrastive learning explicitly enforces relational structure.
- Triplet loss introduces margin.
- InfoNCE scales supervision.
- Temperature governs curvature.
- CLIP unifies modalities via shared embedding space.
- SigLIP simplifies scaling and improves stability.

What Must You Carry Forward

- Always ask: what geometry is your loss imposing?
- Contrastive loss creates separability, not just predictability.
- Retrieval systems depend on embedding margins.
- Temperature is not cosmetic — it shapes the manifold.
- Multi-modal models are geometric unifiers.

Essential Reading

1. **SimCLR: A Simple Framework for Contrastive Learning of Visual Representations** [5]
   Why read this: establishes clean baseline for contrastive scaling.
2. **An Image is Worth 16x16 Words: Vision Transformer** [6]
   Why read this: attention beyond text.
3. **CLIP: Learning Transferable Visual Models from Natural Language Supervision** [7]
   Why read this: multimodal alignment breakthrough.
4. **Sigmoid Loss for Language-Image Pre-Training** [8]
   Why read this: scaling refinement over softmax contrastive loss.

Oliver Twist List (For the Voracious Mind)

1. **On the Uniformity and Alignment of Representations** [9]
2. **SimCSE: Simple Contrastive Learning of Sentence Embeddings** [10]
3. **Sentence-BERT: Sentence Embeddings using Siamese BERT Networks** [11]

Looking Ahead

Contrastive learning is not merely a training trick. It is geometric engineering.
Next week:

- Contrastive loss in retrieval systems
- Hard-negative mining

[5] Ting Chen et al. (2020). "A Simple Framework for Contrastive Learning of Visual Representations". In: *ICML*

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: **International Conference on Learning Representations (ICLR)**

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. (2021). "Learning Transferable Visual Models From Natural Language Supervision". In: **Proceedings of the 38th International Conference on Machine Learning (ICML)**

[8] Xiaohua Zhai et al. (2023). "Sigmoid Loss for Language-Image Pre-Training". In: *arXiv preprint arXiv:2303.15343*

[9] Tongzhou Wang and Phillip Isola (2020). "Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere". In: *ICML*

[10] Tianyu Gao, Xingcheng Yao, and Danqi Chen (2021). "SimCSE: Simple Contrastive Learning of Sentence Embeddings". In: *EMNLP*

[11] Nils Reimers and Iryna Gurevych (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *arXiv preprint arXiv:1908.10084*

- Matryoshka embeddings
- Reranking architectures

Experiments with the Architecture of Representation

Contents

Introduction

In the transition from Software Engineering to AI Engineering, the most critical shift in mental models is moving from **Discrete Logic** (if-else, key-value) to **Vector Topology** (manifolds, distance, and density). In a standard software system, a "Physics" article and a "Politics" article are separated by a database tag. In a Neural Information Retrieval system, they are separated by the **angular margin** of their embeddings.

However, not all "intelligent" models produce usable geometry. This lab explores the fundamental divide between **Negative Log-Likelihood (NLL)**—the objective that powers models like BERT—and **Contrastive Loss**, which powers modern Bi-Encoders and SBERT. We will prove that while a model can "know" a lot of facts (NLL), it may still be "disorganized" (Anisotropic) in its internal representation.

Technical Background: The Objective Function Gap

To understand why our experiment works, we must look at what the models were "paid" to do during training.

1. The NLL Objective (Predictive)

Standard BERT is trained on **Masked Language Modeling (MLM)**. The loss function is a variation of Cross-Entropy:

$$L = -\sum_i y_i \log(\hat{y}_i)$$

The model wins if it guesses the missing word in a sentence. This forces the model to learn syntax and grammar. However, it never explicitly compares two sentences. As a result, BERT embeddings often suffer from **Anisotropy**: they occupy a narrow, high-density cone in the vector space. Because all vectors point in roughly the same direction, the "average" cosine similarity between any two random sentences is often 0.8 or higher.

2. The Contrastive Objective (Discriminative)

Contrastive models (like those using **InfoNCE** or **Triplet Loss**) are trained specifically to distinguish between pairs. They use a formula designed to pull "Positive" pairs together and push "Negative" pairs apart:

$$L = -\log \frac{\exp(\text{sim}(q, p^+)/\tau)}{\sum_{i=0}^{N} \exp(\text{sim}(q, p_i)/\tau)}$$

The model treats the hypersphere as a map where orthogonal concepts must be placed at opposite poles.

Experiment 1: Visualizing the "Collapsed" Space

The Task

Load two models: `bert-base-uncased` (NLL) and `all-MiniLM-L6-v2` (Contrastive). Generate a 300×300 Cosine Similarity matrix. Extract:

1. **Intra-class similarities:** row and column belong to the same topic.
2. **Inter-class similarities:** row and column belong to different topics.

What did we find? BERT shows a "unimodal" distribution where Inter and Intra-class curves overlap at ~ 0.9. The Contrastive model shows a "bimodal" distribution with a clear gap at 0.5.

Experiment 2: The Linear Probe Stress Test

The Task

Train a `LogisticRegression` on 20% of the embeddings. This is a **Linear Probe**. After training, add Gaussian Noise: $V_{noisy} = V + \mathcal{N}(0, 0.1)$.

Lessons Learned BERT "knows" the difference, but the information is **tangled**. Contrastive loss "unfolds" the manifold, making the boundary robust to noise and interlopers.

Summary

As an AI Engineer, remember: **The vector is the UI.** If your embeddings are anisotropic, your retrieval will be noisy. Always check your histograms before you trust your search results.

There were 14 pages in this document.

FEEDBACK

Please send any feedback on this document, or report errors to the author at `asif@supportvectors.com`