

Stitch

AML Challenge Report 2025/26

Davide Perniconi
Matricola: 1889270

Daniele Marretta
Matricola: 1985747

Olga Corovencova
Matricola: 2249558

Leonardo Lavezzari
Matricola: 1984079

1 Proposed Method

We used a *contrastive approach* in order to learn a mapping function $T : R^{1024} \Rightarrow R^{1536}$ able to correctly align the captions of the text embeddings (`roberta-large-nli-stsb-mean-tokens` [1]) with the space of image embeddings (`DINOv2-giant` [2]).

- **Architecture:** Our model is a **MLP** network with the following configuration:

1. *Input dimension:* 1024
2. *Hidden layers:* [1256, 2048]
3. *output dimension:* 1536
4. *Activation:* Gelu
5. *Dropout:* 0.5
6. *Learnable Temperature parameter*
7. *Inizialization:* Xavier uniform
8. *Normalization:* The final embedding is L2-normalized

- **Loss Function:** We augmented the InfoNCE loss with a large pool of negative samples, conceptually inspired by the **Momentum Contrast (MoCo)** framework [3]. Our architecture adapts the MoCo Query/Key mechanism for multimodal alignment:

1. **Query Encoder:** This role is fulfilled by our **MLP projector**, which operates on the fixed text encoder embeddings. The MLP is the only component updated via backpropagation, learning the translation into the target space.
2. **Key Encoder (Target):** This role is conceptually fulfilled by the image encoder. Instead of maintaining a momentum-updated key encoder, its embeddings are treated as fixed, non-updating anchors (the key targets).
3. **Memory Bank:** A FIFO queue that stores key vectors from prior mini-batches. This bank provides a large, diverse set of negative samples without requiring a prohibitively large batch size, which is the core benefit of the MoCo approach.

- **Training Details:** We trained the model with AdamW optimizer (learning rate 0.01, weight decay 1×10^{-5}) using batches of 256. A ReduceLROnPlateau scheduler, adjusted the learning rate based on validation loss. The size of the queue is 10,000 and we used early stopping when validation loss stagnated. Validation performance was tracked using MRR and Validation Loss. All hyperparameters, including the queue size, learning rate, hidden layer sizes, and dropout rate, were selected via hyperparameter optimization using Optuna [4].

2 Results and Discussion

Our final model achieved strong performance on the public leaderboard, recording a MRR of 0.874. On the validation set, the model yielded: MRR = 0.942, Recall@1 = 0.906, Recall@3 = 0.973, and NDCG = 0.956. These metrics substantially outperform the challenge baseline, which reported an MRR of approximately 0.462. This method effectively performs cross-modal translation by utilizing the projection network to map text embeddings into the target image space, ensuring proximity for paired samples. We employ L2-normalization for training stability, preventing scale differences from dominating the optimization. The InfoNCE objective promotes strong alignment of positive pairs and separation from a large set of negatives. This is achieved efficiently by the memory queue without the need for large mini-batches. However, InfoNCE embeddings inherently prioritize maximizing the margin between positive and negative logits, which can result in less precise absolute distance minimization compared to losses like MSE or direct cosine similarity.

3 Conclusion

Our method, using an MLP trained with InfoNCE loss achieved an MRR of 0.874 on the public leaderboard. This contrastive approach learned the mapping between the two spaces, balancing retrieval performance against the inherent precision trade-offs of the InfoNCE objective. [LINK TO THE REPO](#)

4 What We Tried

During the competition, we found and trained models that did not outperform our final MoCo-based method. Instead, they provided valuable insights and contributed to our overall understanding of the problem.

Method 1: Latent Space Translation

Our initial strategy, which was based on the zero-shot translation principle [5], involved an affine mapping. We explored three variations of this simple approach: a basic **linear** mapping, a refinement using the **l-ortho** constraint, and the **ortho** solution derived from Procrustes analysis. All three methods underperformed, showing poor results. This performance shortfall led us to conclude that the relationship between the text and image embedding spaces is highly non-linear and fundamentally requires a more complex translation function.

Method 2: Direct Mapping with Regression

We attempted a regression approach to predict image embeddings. We tested both a standard shallow architecture and a deep architecture with residual connections. Models were trained using MSE loss, Cosine Distance or a combination of both. Ultimately, this approach proved ineffective. The inability of even the residual network to converge on a good solution highlights the structural incompatibility of the two embedding spaces for direct mapping.

Method 3: Inverse Pseudo-Transform

Our second strategy leveraged the Inverse Pseudo-Transform (IPT) method [6], which constructs two distinct relative spaces (\mathbf{X}_{rel} and \mathbf{Y}_{rel}) using separate but parallel anchor points. We utilized the concepts of anchor pruning and completion, along with the suggested values for δ and ω . The core IPT idea for cross-modal translation requires $\mathbf{X}_{\text{rel}} = \mathbf{Y}_{\text{rel}}$ in order to apply the formula:

$$\mathbf{Y}_{\text{abs}} = \mathbf{X}_{\text{rel}} \cdot (\mathbf{A}_y^T)^{-1}$$

We successfully reconstructed the absolute spaces for both text and image modalities with near-perfect accuracy. However, the crucial assumption for translation, $\mathbf{X}_{\text{rel}} = \mathbf{Y}_{\text{rel}}$, failed to hold sufficiently for accurate reconstruction of \mathbf{Y}_{abs} using the text relative space \mathbf{X}_{rel} . To enforce the necessary alignment $\mathbf{X}_{\text{rel}} \approx \mathbf{Y}_{\text{rel}}$, we used different approaches: Least-Squares regression, Procrustes analysis, and a MLP but all three alignment methods underperformed, resulting in overall poor retrieval performance. This approach is similar to direct regression between the absolute spaces, but there’s a key difference: the alignment is done on the much lower-dimensional relative spaces, whose dimensionality is determined by the number of anchors chosen. This potentially makes the direct translation easier and stops the problem from being considered a retrieval one like in the contrastive approach.

Method 4: Two-Head MLP with Memory Bank (Second best model)

We also experimented with a two-head architecture, in which the output embedding is derived by combining two components that are both trained within the same MLP:

- **Direction Head** d , which predicts a unit-norm vector representing the angular component of the embedding.
- **Scale Head** s , which predicts a scalar/vector (enforced to be positive via *Softplus*) representing the magnitude of the embedding.

The final predicted image embedding was constructed by combining them as:

$$T(x) = \text{normalize}(d(x)) \cdot s(x).$$

Both heads were trained using the same MoCo-style InfoNCE loss, then thanks to the support of a large negative queue of 32,000 samples, the model achieved competitive retrieval performance (MRR ≈ 0.871) on the public leaderboard. However, it did not surpass the simpler main approach, likely because scale modulation introduces additional instability in contrastive optimization.

Method 5: Flow Matching

We attempted to use a Flow Matching approach [7], modeling the transformation (Vector Field) from text to image embeddings. However, this sophisticated model architecture only achieved a maximum MRR of 0.76, indicating that the predicted image embeddings were still not accurate enough for effective cross-modal retrieval.

References

- [1] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. URL: <https://arxiv.org/abs/1908.10084>.
- [2] Maxime Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. 2023. arXiv: [2304.07193](https://arxiv.org/abs/2304.07193) [cs.CV].
- [3] Kaiming He et al. *Momentum Contrast for Unsupervised Visual Representation Learning*. 2020. arXiv: [1911.05722](https://arxiv.org/abs/1911.05722) [cs.CV]. URL: <https://arxiv.org/abs/1911.05722>.
- [4] Takuya Akiba et al. *Optuna: A Next-generation Hyperparameter Optimization Framework*. 2019. arXiv: [1907.10902](https://arxiv.org/abs/1907.10902) [cs.LG]. URL: <https://arxiv.org/abs/1907.10902>.
- [5] Valentino Maiorca et al. *Latent Space Translation via Semantic Alignment*. 2024. arXiv: [2311.00664](https://arxiv.org/abs/2311.00664) [cs.LG]. URL: <https://arxiv.org/abs/2311.00664>.

- [6] Valentino Maiorca et al. *Latent Space Translation via Inverse Relative Projection*. 2024. arXiv: [2406.15057](https://arxiv.org/abs/2406.15057) [cs.LG]. URL: <https://arxiv.org/abs/2406.15057>.
- [7] Yaron Lipman et al. *Flow Matching for Generative Modeling*. 2023. arXiv: [2210.02747](https://arxiv.org/abs/2210.02747) [cs.LG]. URL: <https://arxiv.org/abs/2210.02747>.