Honolulu

AML Challenge Report 2025/26

Sandro Khizanishvili
Matricola: 2175979

Emre Yesil
Matricola: 2186113

Juan Iñaki Larrea
Matricola: 2241722

Mohammadreza heibati
Matricola: 2166497

1 Proposed Method

We propose an approach for text-to-image embedding translation using contrastive learning with multiple positives. Our method learns to map text embeddings to corresponding image embeddings in a shared semantic space, enabling effective cross-modal retrieval.

- **Architecture:** We employ a 3-layer MLP adapter network with Batch Normalization and GELU activations. The network takes standardized (standard scaling) 1024-dimensional text embeddings as input and produces 1536-dimensional image embeddings as output. The architecture consists of: input dimension 1024 → hidden layer 1536 → hidden layer 2048 → output dimension 1536, with dropout probability of 0.5 for regularization. This design provides sufficient capacity to learn complex text-to-image mappings while maintaining training stability.

- **Loss Function:** We optimize a multi-positive contrastive loss that extends the standard InfoNCE loss [1, 2] to handle multiple positive pairs. For a batch of N samples, we compute both text-to-image and image-to-text directions:

$$\mathcal{L}_{\text{t2i}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\sum_{j\in\mathcal{P}_i}\exp(s_{ij}/\tau)}{\sum_{k=1}^{N}\exp(s_{ik}/\tau)}\right)$$

$$\mathcal{L}_{\text{i2t}} = -\frac{1}{N}\sum_{j=1}^{N}\log\left(\frac{\sum_{i\in\mathcal{P}_j}\exp(s_{ij}/\tau)}{\sum_{k=1}^{N}\exp(s_{kj}/\tau)}\right)$$

  where s_{ij} is the cosine similarity between the predicted embedding for text i and target image j, $\mathcal{P}_i$ is the set of positive images for text i, $\mathcal{P}_j$ is the set of positive texts for image j, and τ is a learnable temperature parameter. The final loss is the average of both directions: $\mathcal{L} = (\mathcal{L}_{\text{t2i}} + \mathcal{L}_{\text{i2t}})/2$. **Crucially, we remove self-comparisons from the positive mask to prevent trivial solutions and force cross-sample relationship learning.**

- **Training Details:** We train our model for 150 epochs using the Adam optimizer with learning rate 5×10^{-4} and weight decay 10^{-5}. We use a large batch size of 2048 to provide sufficient negative samples for contrastive learning. The temperature parameter τ is learned jointly with the model parameters using a higher learning rate of 10^{-2} and clamped to prevent numerical instability. We employ Cosine Annealing Warm Restarts scheduling with initial period $T_0 = 10$ epochs and minimum learning rate 10^{-6}. Gradient clipping with maximum norm 1.0 ensures training stability. Model selection is based on Mean Reciprocal Rank (MRR) on the validation set, with the best model checkpoint saved for evaluation.

2 Results and Discussion

Our final model achieved an MRR of **0.85637** on the public leaderboard, significantly outperforming the baseline score.

The key to this improvement was reframing the problem from regression to retrieval. While the baseline used MSE loss with a simple linear model, we employed a **Multi-Positive Contrastive Loss** with large batches (2048) and **removed self-comparisons** to prevent trivial solutions. Additional enhancements included a **learnable temperature** parameter and **Cosine Annealing Warm Restarts** for better optimization.

This contrastive approach proved substantially more effective at capturing the complex relationship between text embeddings and VAE latents, demonstrating that retrieval-oriented learning is better suited for this cross-modal alignment task..

3 Conclusion

Our multi-positive contrastive learning framework successfully addresses the challenge of mapping multiple text captions to shared image embeddings by preventing trivial solutions through strategic self-comparison removal.

LINK TO THE REPO

What We Tried

This section details the various approaches we explored during the competition, including those that were not part of our final submission.

Method 1: Orthogonal Procrustes Analysis

We implemented a parameter-free, closed-form solution based on Orthogonal Procrustes Analysis [3]. This method finds the optimal rotation/reflection matrix W that aligns the text embedding space to the VAE latent space by solving $\text{argmin}_W \|Y - XW\|^2$ subject to $W^T W = I$, where X is the matrix of text embeddings and Y is the target VAE embeddings. The solution is calculated via Singular Value Decomposition (SVD) as $W = UV^T$, given $X^T Y = U\Sigma V^T$.

This method achieved strong performance (MRR: 0.87 train, 0.84 validation), demonstrating the efficacy of orthogonal alignment. However, the strict orthogonality constraint proved too limiting for our modality translation task. While it outperformed linear baselines, it was surpassed by more flexible non-linear projections.

Method 2: Simple MLP with Contrastive Loss

After trying different hyper parameters in a simple MLP with MSE as the loss function, we decide to try changing it with the contrastive loss. This modification succed on a better score in the public leaderboard of **0.82741**.

Model Architecture:

The neural network has three fully connected layers, mapping 1024-dimensional text embeddings to the 1536-dimensional image space. The hidden layer has the same size of the previous model with the MSE loss, **2048 units**. In the activation function we tried different options: ReLU, LeakyRELU, SiLU, GELU. We decide to keep the one which has the best performance: **LeakyRELU**.

Training Procedure:

The model was trained using a **contrastive learning** approach via the **symmetric InfoNCE loss** (similar to the method proposed in CLIP), optimizing the alignment of text embeddings (**P**) and image embeddings (**Y**) within the same minibatch. The objective is to maximize the similarity between the positive pair (text and its corresponding image) and minimize it with all other negative pairs within the batch. We try different sizes of the batches, and the best one was 256. As in the Wide Contrastive Loss, we added a **learnable temperature** parameter, initialized at ln(1/0.07). For optimization, the **Adam** optimizer was used with a weight decay of 10^{-5}.

Method 3 : Wide Contrastive MLP

As a strong alternative to our main submission, we explored a **Wide Multi-Layer Perceptron (MLP)** trained with contrastive learning. This approach achieved an impressive MRR of **0.86279** on the public leaderboard.

Model Architecture:

The network consists of three fully connected layers, mapping 1024-dimensional text embeddings to the 1536-dimensional image space. A key design choice was to expand the hidden layer to **4096 units**. This wider layer helped the model capture detailed semantic patterns better than deeper, narrower networks. We also replaced the usual ReLU activation with **GELU**, which improved gradient flow and reduced the risk of dead neurons during training.

Training Procedure:

We trained the model using the **InfoNCE (contrastive) loss**, adopting the approach proposed in CLIP [4]. Importantly, we added a **learnable temperature** parameter, initialized at ln(1/0.07). This allowed the model to dynamically scale the logits and sharpen the difference between positive and negative pairs as training progressed. To ensure numerical stability, we imposed an upper bound on this scaling factor by clamping it at ln(100), preventing the logits from growing indefinitely and avoiding potential gradient explosions. We further boosted performance by using a **Cosine Annealing Warm Restarts** scheduler [5] ($T_0 = 10$), which periodically reset the learning rate and helped escape local minima.

Remarks:

Although this method performed very well, we ultimately chose **another model** -that we explain above- for the final submission due to its architecture seeming more reliable and likely to perform better on the private leaderboard. Nevertheless, this experiment demonstrates that a carefully tuned MLP with modern contrastive techniques can achieve competitive results even against more complex architectures.

References

[1] Maksym Bekuzarov. *Losses Explained: Contrastive Loss*. https://medium.com/@maksym.bekuzarov/losses-explained-contrastive-loss-f8f57fe32246. 2020.

[2] Brian Williams. *Contrastive Loss Explained*. https://medium.com/data-science/contrastive-loss-explaned-159f2d4a87ec. 2020.

[3] Valentino Maiorca et al. *Latent Space Translation via Semantic Alignment*. https://arxiv.org/pdf/2311.00664. 2024.

[4] Rishabh Misra. *Understanding CLIP for Vision Language Models*. Blog Post. Accessed: 2025-11-18. 2021. URL: https://medium.com/self-supervised-learning/understanding-clip-for-vision-language-models-43b700a4aa2b.

[5] Ilya Loshchilov and Frank Hutter. *SGDR: Stochastic Gradient Descent with Warm Restarts*. 2016. arXiv: 1608.03983 [cs.LG]. URL: https://arxiv.org/abs/1608.03983.