viPolyQwen: A Curriculum-Based Approach to Multi-Task Multimodal Embedding Learning with Dynamic Loss Balancing

Nguyen Anh Nguyen* (EraX) & Gtel Mobile JSC (GMobile) – Vietnam.

*Corresponding Author: nguyen@hatto.com

Abstract

We present viPolyQwen, a unified multimodal embedding framework that leverages curriculum learning and dynamic loss balancing to effectively train on heterogeneous data types. Built upon the Qwen2-VL-2B-Instruct foundation model, our approach introduces several key innovations: (1) a two-phase curriculum strategy that establishes strong text representations before introducing multimodal complexity, (2) multi-head attention pooling with learnable query vectors for effective sequence aggregation, (3) an enhanced projection architecture with residual connections for stable training dynamics, and (4) prefix-guided task conditioning with adaptive loss weighting to balance the embedding space across diverse data types. Training on 7.5M samples spanning text similarity, OCR, and visual question answering tasks, our framework addresses the fundamental challenge of learning unified representations that excel across multiple modalities and tasks. Through careful architectural design and training strategies, viPolyQwen demonstrates the effectiveness of curriculum learning in multimodal embedding spaces.

1. Introduction

The development of unified multimodal embedding models presents significant challenges in balancing representation quality across diverse data types and tasks. While vision-language models (VLMs) have shown remarkable capabilities in cross-modal understanding, translating these capabilities into effective, general-purpose embeddings requires careful consideration of training dynamics, loss function design, and data presentation strategies.

Traditional approaches to multimodal embedding learning often struggle with several key challenges:

1. **Task Imbalance**: Different data types naturally exhibit different loss scales and convergence rates, leading to imbalanced optimization where dominant tasks overshadow others.
2. **Representation Collapse**: Without proper regularization and architectural design, embeddings may collapse to trivial solutions that fail to capture meaningful semantic relationships.
3. **Multimodal Complexity**: Jointly training on text-only and image-text pairs requires careful coordination to prevent interference between modalities.

To address these challenges, we propose viPolyQwen, a comprehensive framework that combines curriculum learning with dynamic loss balancing. Our key contributions include:

- **A two-phase curriculum strategy** that first establishes strong text-based representations on 1M text pairs before introducing 6M multimodal samples, enabling stable learning progression.
- **Multi-head attention pooling** that adaptively aggregates sequence representations through learned query vectors, providing more nuanced feature selection than traditional pooling methods.
- **Enhanced projection architecture** with residual connections and careful initialization strategies that maintain training stability while learning complex multimodal mappings.
- **Prefix-guided task conditioning** that enables explicit task-aware optimization while sharing parameters across all data types, combined with adaptive loss weighting to balance the embedding space.

2. Related Work

2.1 Multimodal Embedding Models

Recent advances in vision-language pretraining have produced powerful models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) that learn joint representations through contrastive learning. However, these models typically optimize for a single objective (e.g., image-text matching) and may not generalize well to diverse downstream tasks requiring different similarity notions.

ColPali (Faysse et al., 2024) addresses document understanding through multi-vector representations, capturing fine-grained information at the cost of increased retrieval complexity. Our work explores whether careful training strategies can enable single-vector embeddings to capture similar richness while maintaining computational efficiency.

2.2 Multi-Task Learning and Loss Balancing

Multi-task learning has long recognized the challenge of balancing different objectives (Caruana, 1997). Recent approaches like GradNorm (Chen et al., 2018) and uncertainty weighting (Kendall et al., 2018) propose dynamic strategies for loss balancing. Our work differs by introducing curriculum learning as a complementary strategy, establishing foundational representations before tackling the full complexity of multimodal data.

2.3 Curriculum Learning

Curriculum learning (Bengio et al., 2009) proposes that models benefit from seeing examples in a meaningful order, typically from simple to complex. While widely studied in supervised learning, its application to multimodal embedding learning remains underexplored. Our two-phase curriculum specifically addresses the challenge of learning unified representations across modalities.

3. Methodology

3.1 Model Architecture

viPolyQwen builds upon the Qwen2-VL-2B-Instruct foundation model, extending it with specialized components for embedding generation:

3.1.1 Multi-Head Attention Pooling

Traditional pooling methods (mean, max, or last-token) may not optimally aggregate information from variable-length sequences. We introduce a multi-head attention pooling mechanism that learns to weight sequence elements based on their relevance:

Given encoder outputs $\mathrm{H} \in \mathbb{R}^{L \times d}$ where L is sequence length and d is hidden dimension, we compute:

$$\mathrm{u}_i = \mathrm{h}_i^T \mathrm{v}_a$$

where $\mathrm{v}_a \in \mathbb{R}^d$ is a learned query vector. The attention weights are computed as:

$$a_i = \frac{\exp(u_i) \cdot M_i}{\sum_{j=1}^{L} \exp(u_j) \cdot M_j}$$

where M is the attention mask. This process is repeated for $k = 4$ attention heads, producing aggregated representations that are concatenated and projected.

3.1.2 Enhanced Projection with Residual Connections

The projection from hidden states to the final embedding space employs a sophisticated architecture:

$$\mathrm{p} = \mathrm{LayerNorm}(\mathrm{W}_2 \cdot \mathrm{GELU}(\mathrm{W}_1 \mathrm{c} + \mathrm{b}_1) + \mathrm{W}_2 \cdot \mathrm{b}_2)$$

Crucially, we incorporate a learnable residual connection with adaptive scaling:

$$\mathrm{e} = \alpha \cdot \mathrm{p} + (1 - \alpha) \cdot \mathrm{W}_r \mathrm{c}$$

where α is initialized to 0.5 and learned during training, and W_r projects the input to match dimensions when necessary.

3.2 Loss Functions and Task-Specific Objectives

3.2.1 Prefix-Guided Task Conditioning

Each input is prepended with a task-specific prefix $p_i \in \{$<text_pair>, <ocr>, <vqa_single>, <vqa_multi>$\}$, enabling the model to apply task-appropriate processing while sharing parameters.

3.2.2 Text Similarity Loss

For text pairs with similarity scores $s \in [0, 1]$, we employ a three-component loss:

$$\mathcal{L}_{\text{text}} = \mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{score}} \cdot \mathcal{L}_{\text{MSE}} + \lambda_{\text{rank}} \cdot \mathcal{L}_{\text{rank}}$$

where: - $\mathcal{L}_{\text{InfoNCE}}$ is the standard contrastive loss with temperature $\tau = 0.07$ - $\mathcal{L}_{\text{MSE}} = (s - \hat{s})^2$ where $\hat{s} = \frac{1}{2}(\mathrm{e}_a^T \mathrm{e}_b + 1)$ - $\mathcal{L}_{\text{rank}}$ ensures correct ordering of similarity scores

The ranking loss is particularly important for maintaining relative relationships:

$$\mathcal{L}_{\text{rank}} = \frac{1}{|P|} \sum_{(i,j) \in P} \max(0, m_r - (\hat{s}_i - \hat{s}_j))$$

where $P = \{(i, j) | s_i > s_j\}$ and $m_r = 0.05$ is the margin.

3.2.3 OCR and VQA Losses

For OCR and VQA tasks, we combine contrastive learning with triplet loss:

$$\mathcal{L}_{\text{OCR/VQA}} = \mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{triplet}} \cdot \mathcal{L}_{\text{triplet}}$$

The triplet loss with margin $m = 0.2$ ensures separation between positive and negative pairs:

$$\mathcal{L}_{\text{triplet}} = \max(0, d(\mathrm{e}_a, \mathrm{e}_p) - d(\mathrm{e}_a, \mathrm{e}_n) + m)$$

3.3 Curriculum Learning Strategy

Our two-phase curriculum addresses the challenge of learning from heterogeneous data:

Phase 1 (Steps 0-2000): Train exclusively on 1M text similarity pairs to establish strong language understanding and similarity metrics.

Phase 2 (Steps 2000+): Introduce the full 6.5M dataset including multimodal samples.

This strategy is motivated by the observation that text-based similarity provides a strong foundation for multimodal understanding. During the transition, we apply a temporary boost factor (2.0x) to text-pair losses to maintain their influence as multimodal samples dilute their proportion in batches.

3.4 Dynamic Loss Balancing

To address the inherent scale differences between tasks, we employ adaptive loss weighting:

$$\lambda_{\text{task}} = \begin{cases} 1.0 & \text{for text pairs} \\ 1.0 & \text{for OCR} \\ 1.0 & \text{for single-turn VQA} \\ 1.2 & \text{for multi-turn VQA} \end{cases}$$

Additionally, we implement warmup schedules for specific loss components: - Temperature: $0.1 \rightarrow 0.07$ - Score loss weight: $0.5 \rightarrow 3.0$ - Rank loss weight: $0.1 \rightarrow 1.0$

4. Implementation Details

4.1 Training Configuration

- **Hardware**: 4 NVIDIA GPUs with mixed precision (bfloat16)
- **Batch size**: 16 per device $\times$ 4 GPUs = 64 global batch size
- **Learning rates**:
  - Language model backbone: 1×10^{-4}
  - Vision encoder: Frozen
  - Projection layers: 3×10^{-4}
- **Gradient clipping**: Adaptive schedule from 50.0 to 5.0
- **Weight decay**: 0.05 for projection layers, 0.001 for backbone

4.2 Dataset Composition

Our training dataset comprises 7.5M samples: - **Text similarity pairs**: 3.5M samples with scores in [0, 1] - **OCR**: 1.5M image-text pairs - **VQA single-turn**: 1.5M question-answer pairs - **VQA multi-turn**: 1M conversational examples

The dataset includes substantial Vietnamese content alongside English and Chinese, addressing the need for multilingual embedding models.

4.3 Training Stability Measures

Several design choices ensure stable training:

1. **Careful initialization**: Xavier initialization with gain=1.0 for projection layers
2. **Gradient accumulation**: Effective batch size of 480 through 10 accumulation steps
3. **DDP synchronization**: Always execute backward pass, using zero loss for problematic batches
4. **Edge case handling**: Explicit handling of batch_size=1 scenarios in loss functions

5. Experimental Observations

While comprehensive benchmarking is ongoing, early training dynamics show promising signs:

5.1 Embedding Space Evolution

Monitoring the gap between positive and negative similarities reveals healthy learning dynamics: - Initial gap: ~0.01 (near collapse) - After 120 steps: ~0.07 (clear separation emerging) - Target by step 2000: 0.10-0.15

5.2 Loss Component Contributions

The multi-component loss successfully balances different objectives: - InfoNCE provides the primary discriminative signal - MSE loss aligns absolute similarity values - Ranking loss preserves relative relationships

5.3 Curriculum Effectiveness

The two-phase curriculum demonstrates clear benefits: - Phase 1 establishes stable text representations - Phase 2 successfully incorporates multimodal complexity without destabilizing learned representations

6. Discussion

6.1 Architectural Innovations

The combination of multi-head attention pooling and residual projection connections addresses key challenges in embedding learning:

1. **Attention pooling** enables task-specific feature selection without explicit task-specific parameters
2. **Residual connections** in the projection layer prevent information loss during dimensionality reduction
3. **Learnable scaling factor** α allows the model to adaptively balance between transformed and residual paths

6.2 Curriculum Learning Insights

Our results suggest that curriculum learning provides substantial benefits for multimodal embedding training:

1. **Foundation establishment**: Text-only pretraining creates a semantic scaffold for multimodal learning
2. **Stable transitions**: Gradual introduction of complexity prevents catastrophic forgetting
3. **Task balance**: The curriculum naturally addresses data imbalance issues

6.3 Limitations and Future Work

Several areas warrant further investigation:

1. **Scalability**: Extending to larger backbone models and datasets
2. **Task diversity**: Incorporating additional modalities (audio, video)
3. **Evaluation**: Comprehensive benchmarking across diverse retrieval tasks
4. **Theoretical analysis**: Formal understanding of curriculum benefits in embedding spaces

7. Conclusion

viPolyQwen demonstrates that careful architectural design and training strategies can enable effective multi-task multimodal embedding learning. Our curriculum-based approach, combined with multi-head attention pooling and sophisticated loss balancing, addresses fundamental challenges in learning unified representations across diverse data types.

The framework's success in maintaining stable training dynamics while learning from heterogeneous data suggests that curriculum learning may be a crucial component for future multimodal models. As we continue training and evaluation, we anticipate that viPolyQwen will provide valuable insights into the design of general-purpose embedding models.

Acknowledgments

We thank the Qwen team for their foundational vision-language model and the open-source community for tools enabling this research.

References

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In Proceedings of the 26th International Conference on Machine Learning.

Caruana, R. (1997). Multitask learning. Machine learning, 28(1), 41-75.

Chen, Z., Badrinarayanan, V., Lee, C. Y., & Rabinovich, A. (2018). Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In International Conference on Machine Learning.

Faysse, M., Sibille, H., Wu, T., et al. (2024). ColPali: Efficient document retrieval with vision language models. arXiv preprint arXiv:2407.01449.

Jia, C., Yang, Y., Xia, Y., et al. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning.

Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning.