

VERSE: Visual Embedding Reduction and Space Exploration - Latent-space clustering for improving document understanding

I. de Rodrigo Tobías; A.J. López López; J. Boal Martín-Larrauri

Abstract-

Conventionally, synthetic training data quality is evaluated through human perception, prioritizing visual realism. From the model's perspective, what truly matters is whether a sample lies within the right region of its embedding space. This work introduces VERSE, a methodology for analyzing and improving the performance of Vision–Language Models by exploring their visual embedding space. VERSE enables the visualization of latent representations to assess model feasibility, identifies problematic regions, and guides synthetic data generation to enhance performance in those clusters. We validate the proposed methodology for Visually-rich Document Understanding by training on the synthetic MERIT Dataset and evaluating on its real-world counterpart, MERIT Secret, focusing on key information extraction as a sequence-generation task scoped to transcripts of records in Spanish. Results show that VERSE uncovers the visual features associated with error-prone clusters, and that retraining with samples containing these features substantially boosts F1 performance without degrading generalization. On-premise models optimized with VERSE—Donut (F1 = 0.76) and Idefics2 (F1 = 0.81)—match or surpass SaaS solutions such as GPT-4o (F1 = 0.78) and Pixtral (F1 = 0.73), preserving data privacy and avoiding external APIs.

Index Terms- Visually-rich Document Understanding; Vision-Language Models; Visual embeddings; Interpretability; Explainability

Due to copyright restriction we cannot distribute this content on the web. However, clicking on the next link, authors will be able to distribute to you the full version of the paper:

[Request full paper to the authors](#)

If your institution has an electronic subscription to Pattern Recognition, you can download the paper from the journal website:

[Access to the Journal website](#)

Citation:

Rodrigo, I. de; López López, A.J.; Boal, J. "VERSE: Visual Embedding Reduction and Space Exploration - Latent-space clustering for improving document understanding", *Pattern Recognition*, , .