MetaEmbed: Scaling Multimodal Retrieval at Test-Time with Flexible Late Interaction

Zilin Xiao[1,2], Qi Ma[1], Mengting Gu[1], Jason Chen[1], Xintao Chen[1], Vicente Ordonez[2], Vijai Mohan[1]

Meta Superintelligence Labs[1] Rice University[2]

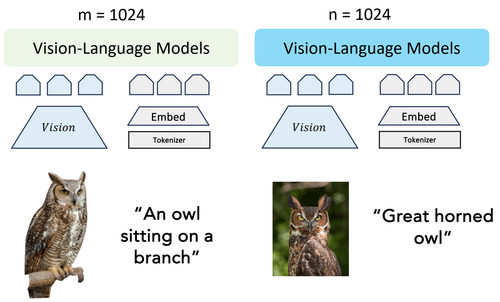
ICLR

Oral Session 4C (Friday 3:51pm)

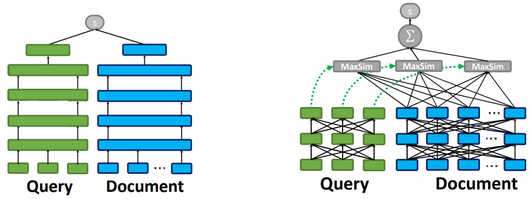
RICE

Meta

The Problem: Multimodal Retrieval



Baseline Methods



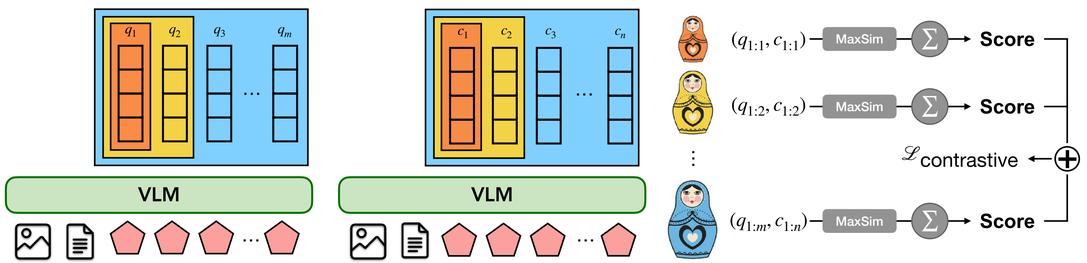
$$s(Q,D) = \mathbf{q}^\top \mathbf{d}$$

Single-vector Retrieval
e.g. CLIP

$$s(Q,D) = \sum_{i=1}^{m} \max_{1 \le j \le n} \mathbf{q}_i^\top \mathbf{d}_j$$

Multi-vector Retrieval
e.g. ColBERT, ColPali

Key Idea: Matryoshka Multi-Vector Retrieval (MMR)

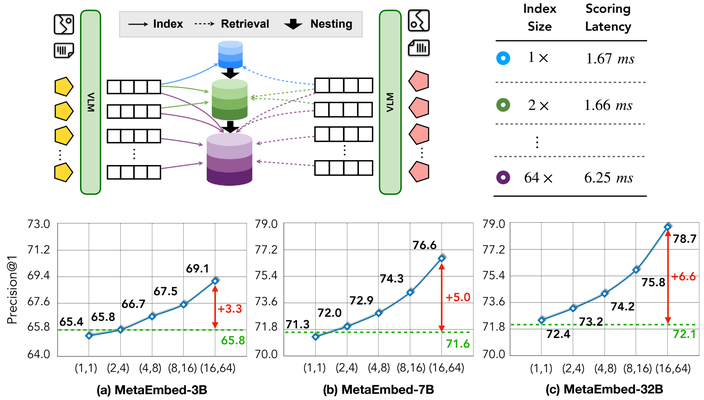


$$\text{Score}\left(q_{1:r_q^{(g)}}, c_{1:r_c^{(g)}}\right) = \sum_{i=1}^{r_q^{(g)}} \max_{j \in [1, r_c^{(g)}]} \langle q_i, c_j \rangle$$

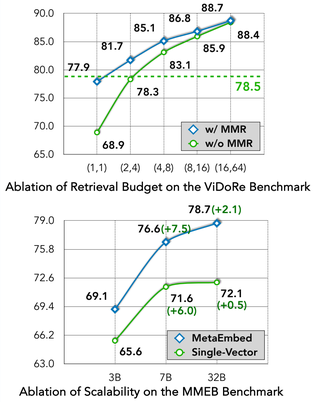
$$\mathcal{L}_{\text{contrastive}} = \mathcal{L}_{\text{NCE}}\big(\text{Score}(q_{1:1}, c_{1:1})\big) + \mathcal{L}_{\text{NCE}}\big(\text{Score}(q_{1:2}, c_{1:2})\big) + \cdots + \mathcal{L}_{\text{NCE}}\big(\text{Score}(q_{1:m}, c_{1:n})\big)$$

- Contrastive losses on nested prefixes so the first few vectors form a coarse summary and additional vectors progressively refine it.

Inference: Flexible Trade-offs: Latency vs Accuracy



Ablations



Multimodal Retrieval Evaluation

- Models trained on MMEB and ViDoRe-Train.
- Best accuracy obtained when using 16 query vectors and 64 retrieval vectors.
- Competitive or superior accuracy even with a single vector.

Models	Size	Per Meta-Task Score				Average Score		
		Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
Baseline Models								
CLIP	428M	55.2	19.7	53.2	62.2	47.6	42.8	45.4
MagicLens	613M	38.8	8.3	35.4	26.0	–	–	27.8
UniIR	428M	42.1	15.0	60.1	62.2	–	–	42.8
MM-EMBED	7B	48.1	32.2	63.8	57.8	–	–	50.0
GME	7B	56.9	41.2	67.8	53.4	–	–	55.8
mmE5	11B	67.6	62.7	71.0	89.7	72.4	66.6	69.8
MoCa-3B	3B	59.8	62.9	70.6	88.6	72.3	61.5	67.5
MoCa-7B	7B	65.8	64.7	75.0	**92.4**	74.7	67.6	71.5
B3-7B	7B	70.0	66.5	74.1	84.6	75.9	67.1	72.0
MetaEmbed – PaliGemma Initialized								
MetaEmbed-3B$^{\text{Gemma}}$	3B	64.9	53.5	70.9	79.5	68.6	61.3	65.4
MetaEmbed – Llama-3.2-Vision Initialized								
MetaEmbed-11B	11B	66.4	42.1	74.3	<u>91.6</u>	65.7	64.3	65.1
MetaEmbed – Qwen2.5-VL Initialized								
MetaEmbed-3B	3B	62.7	68.1	71.9	79.6	73.5	63.8	69.1
MetaEmbed-7B	7B	<u>71.3</u>	<u>74.2</u>	<u>78.7</u>	85.4	<u>81.8</u>	<u>70.0</u>	<u>76.6</u>
MetaEmbed-32B	32B	**73.7**	**78.6**	**78.9**	88.1	**82.8**	**73.7**	**78.7**