Exploiting the Asymmetric Uncertainty Structure of Pre-trained VLMs on the Unit Hypersphere

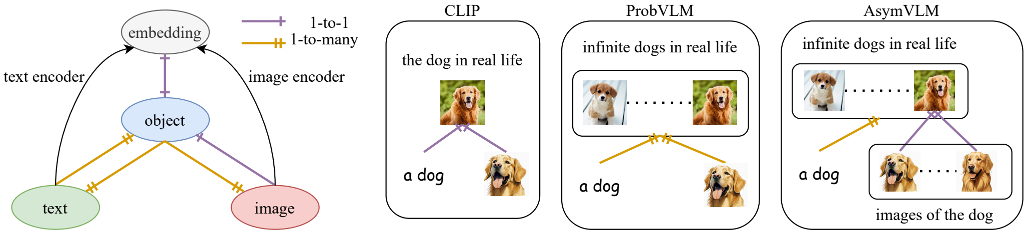
Li Ju

Division of Scientific Computing
Department of Information Technology
Uppsala University

November 6, 2025

Rethinking Building VLMs

- CLIP: "Image–text is an one-to-one mapping".
- ProbVLM[1]: "Image–text is a (symmetric) many-to-many mapping".
- AsymVLM: "Image–text is a many-to-many mapping with an asymmetric structure".



[1] Upadhyay et al., "Probvlm: Probabilistic adapter for frozen vison-language models".

Building the method

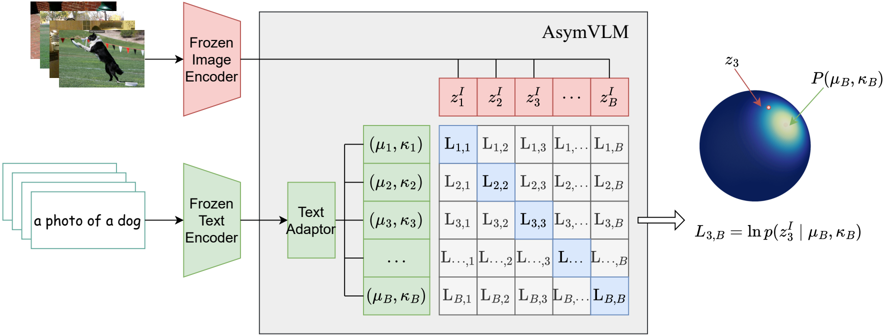
- Text encoder (text $\to$ embedding): one-to-many, modelled by probabilistic embeddings.
- Image encoder (image $\to$ embedding): one-to-one, modelled by deterministic embedding.

Additionally, we need to utlize the pre-trained models (CLIP, BLIP, SigLIP, etc), which have deterministic embeddings on $\mathbb{S}^{d-1}$:

- The method should be post-hoc.
- Probabilistic embeddings should be modelled by directional distributions.

Deriving the Loss

We want to maximize $p(z^I(i) \mid \theta(t))$ if t and i match, and minimize it if they do not:



To maximize the diagonals and minimize the off-diagonals, InfoNCE loss is applied.

Discussion

Unified objectives:

$$\arg\min_{\theta\in\Theta} -\frac{1}{2B}\sum_{n=1}^{B}\left[\ln\frac{\exp\left(\tau\delta(n,n)\right)}{\sum_{m=1}^{B}\exp\left(\tau\ln\delta(n,m)\right)}+\ln\frac{\exp\left(\tau\delta(n,n)\right)}{\sum_{m=1}^{B}\exp\left(\tau\delta(m,n)\right)}\right].$$

Denoting $\mathrm{CosSim}(r,s)=\mu(t_r)^\top z_s^I$, for any $r,s\in[B]$ we have,

for CLIP: $\delta_{\mathrm{CLIP}}(r,s)=\mathrm{CosSim}(r,s)$,

for AsymVLM$_{\mathrm{vMF}}$: $\delta_{\mathrm{vMF}}(r,s)=\kappa(t_r)\cdot\mathrm{CosSim}(r,s)+F_d(\kappa(t_r))$,

for AsymVLM$_{\mathrm{PS}}$: $\delta_{\mathrm{PS}}(r,s)=\kappa(t_r)\ln(1+\mathrm{CosSim}(r,s)+\ln C_d(\kappa(t_r))$.

Empirical results: Uncertainty evaluation