ProLIP: Probabilistic Language-Image Pretraining

DeepSeek

1 Architecture Overview

ProLIP uses two parallel neural networks:

- **Visual Encoder**: Processes images through layers: Image $\rightarrow$ CNN $\rightarrow$ ResNet $\rightarrow$ Embedding Vector $\mathbf{v}$
- **Textual Encoder**: Processes text through layers: Text $\rightarrow$ BERT $\rightarrow$ Transformer $\rightarrow$ Embedding Vector $\mathbf{t}$

Why This Matters: Separate pathways handle fundamentally different data types (pixels vs. words), mimicking human sensory processing. The encoders create a shared semantic space where images and text become directly comparable.

Connection to ProLIP: The embeddings $\mathbf{v}$ and $\mathbf{t}$ act as a "universal language" for cross-modal reasoning.

Math Connection: Encoders implement nonlinear transformations akin to basis decompositions in functional analysis.

2 Core Components

2.1 Probabilistic Tokens

Special tokens govern uncertainty parameters:

$$[\text{CLS}_v] \rightarrow \mu_v, \log \sigma_v^2 \quad \text{(Visual)}$$
$$[\text{UNC}_t] \rightarrow \mu_t, \log \sigma_t^2 \quad \text{(Textual)}$$

Embeddings are sampled as:

$$\mathbf{v} \sim \mathcal{N}(\mu_v, \sigma_v^2), \quad \mathbf{t} \sim \mathcal{N}(\mu_t, \sigma_t^2)$$

Why This Matters: Explicit uncertainty modeling prevents overconfidence, crucial for ambiguous inputs (e.g., blurry images).

Connection to ProLIP: The $\log \sigma^2$ terms act as trainable confidence indicators.

Math Connection: Rooted in Bayesian inference, where distributions represent beliefs updated via evidence.

2.2 Contrastive Loss

$$\mathcal{L}_{\text{contrast}} = -\log \frac{e^{s(\mathbf{v}_i, \mathbf{t}_i)/\tau}}{\sum_j e^{s(\mathbf{v}_i, \mathbf{t}_j)/\tau}} \tag{1}$$

- Positive pairs: Align matching image-text
- Negative pairs: Separate mismatched pairs
- τ: Temperature parameter controls "strictness"

Why This Matters: Teaches relative similarity like humans learning by comparison.

Connection to ProLIP: Forms the backbone of cross-modal alignment.

Math Connection: Analogous to Boltzmann distributions in statistical mechanics.

2.3 Inclusion Loss

$$\mathcal{L}_{\text{inclusion}} = \mathbb{E}_{t \sim T} \left[\|\mathbf{t} - \text{Proj}_V(\mathbf{t})\|_2 \right] \tag{2}$$

Why This Matters: Prevents text embeddings from hallucinating unrealistic concepts.

Connection to ProLIP: Ensures text features stay grounded in visual reality.

Math Connection: Subspace projection from linear algebra, minimizing reconstruction error.

2.4 Variance Regularization

Why This Matters: Quantifies model uncertainty—critical for safety-critical applications.

Connection to ProLIP: High variance triggers "I don't know" states for ambiguous inputs.

Math Connection: Mirrors the evidence lower bound (ELBO) in variational inference.

2.5 L2-Norm Constraint

$$\|\mathbf{v}\|_2 \leq \gamma \quad \forall \mathbf{v} \in V \tag{3}$$

Why This Matters: Stabilizes training by preventing embedding magnitudes from diverging.

Connection to ProLIP: Creates a compact geometric space for contrastive learning.

Math Connection: Constrained optimization via Lagrange multipliers.

3 Training Dynamics

1. Encoders output probabilistic embeddings
2. Contrastive loss clusters related pairs
3. Inclusion loss projects text to visual space
4. Variance terms regulate confidence
5. L2 constraint bounds embedding magnitudes

Why This Matters: Components interact like instruments in an orchestra—each plays a distinct role but must harmonize.

4 Masked Language Handling

For input: `"A grey [M] cat [M] a [M] hat"`:

- Reconstructs masked tokens using visual context
- Inclusion loss enforces $V \subset V^{\text{masked}}$

Why This Matters: Trains robustness to incomplete data—a universal challenge in real-world AI.

Math Connection: Relates to matrix completion and low-rank approximations.

5 Equation Summary

Component	LaTeX
Contrastive Loss	$\mathcal{L}_{\text{contrast}} = -\log \frac{e^{s(\mathbf{v}_i,\mathbf{t}_i)/\tau}}{\sum_j e^{s(\mathbf{v}_i,\mathbf{t}_j)/\tau}}$
Inclusion Loss	$\mathcal{L}_{\text{inclusion}} = \mathbb{E}_{t\sim T}\left[\|\mathbf{t} - \text{Proj}_V(\mathbf{t})\|_2\right]$
L2 Constraint	$\|\mathbf{v}\|_2 \leq \gamma$
Variance	$\log\sigma^2$

Final Notes: ProLIP bridges AI with fundamental mathematics—from Bayesian uncertainty to geometric manifolds. Its principles extend beyond vision-language tasks, offering a blueprint for reasoning under uncertainty in any domain.