Integrating Vision Foundation Models: Enhancing CLIP's Spatial Reasoning

Hyeon-Jun Yang, Hae-Gon Jeon*

Introduction

- CLIP (Contrastive Language-Image Pre-training) has achieved remarkable performance in zero-shot image-text retrieval tasks
- However, CLIP struggles to capture fine-grained details within images, such as spatial relationships between objects. This limitation persists even during fine-tuning

The sofa is farther than the bed (45.7%)

The bed is farther than the sofa (54.3%) ✗

The fireplace is right of the toilet (32.1%)

The toilet is right of the fireplace (67.9%) ✗

- This work attributes the issue to CLIP's reliance on captions during pre-training and proposes addressing it by leveraging foundation model from other computer vision task

Dataset

- EmbSpatial is VQA (Visual Question Answering) dataset focusing on spatial relations between objects in images
- It defines spatial positions from an ego-centric perspective, categorized into types such as Close, Far, Left, Right, Above, and Under
- This study generate negated caption by relacing the objects in the original caption to evaluate as binary classification

[Original caption]
The sofa is farther than the bed
↓
[Negated caption]
The bed is farther than the sofa

Main Approach

In this work, the power of the off-the-shelf vision encoder (DINOv2) from Depth anything V2 is leveraged to enhance the CLIP vision encoder for Spatial VQA

- Architecture

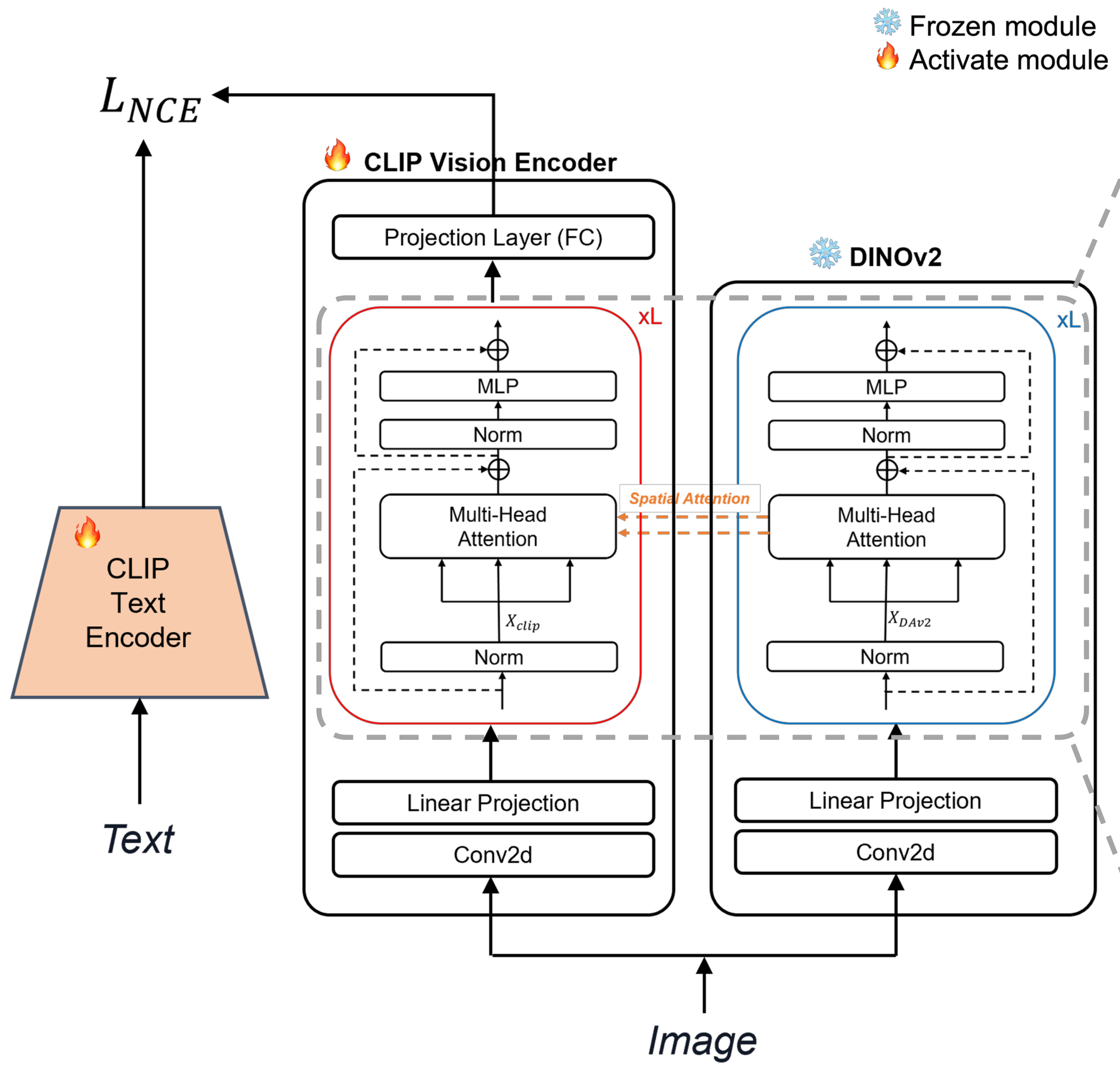


Fig. Overall Architecture

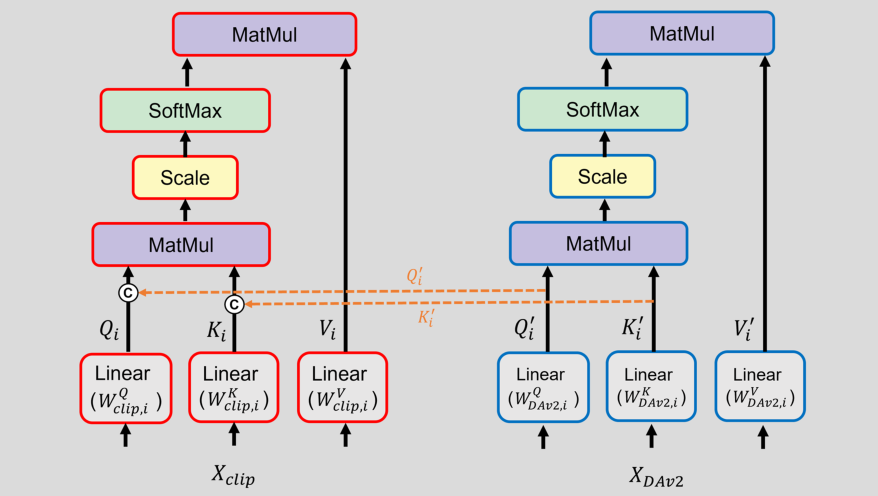


Fig. Revised Scaled-dot product attention for a single head

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{2d_k}}\right)V$$

$$MultiHead(X_{clip}, X_{DAv2}) = Concat(head_1, \dots, head_h)W^O_{clip}$$

$$\text{where } head_i = Attention(Concat(Q_i, Q_i'), Concat(K_i, K_i'), V_i)$$

where

$$Q_i = X_{clip}W^Q_{clip,i} \qquad Q_i' = X_{DAv2}W^K_{DAv2,i}$$

$$K_i = X_{clip}W^K_{clip,i} \qquad K_i' = X_{DAv2}W^K_{DAv2,i}$$

$$V_i = X_{clip}W^V_{clip,i}$$

Eq. Revised Multi-head Attention in CLIP

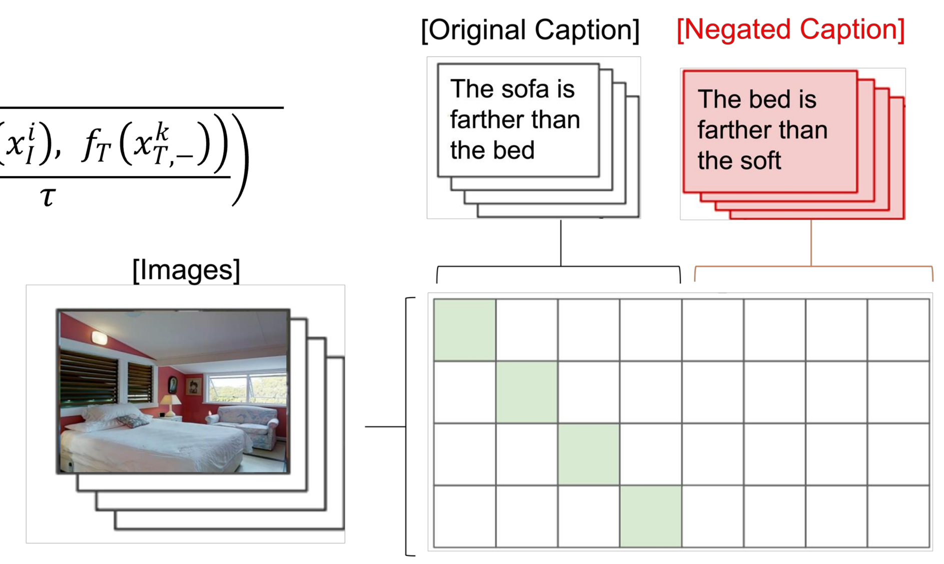
To preserve the features of the CLIP vision encoder while integrating the features of the Vision Foundation Model, This study proposes an architecture that includes uni-directional propagation of Query and Key for every transformer blocks

- Contrastive Learning with negative caption sampling

$$L_I = -\sum_{i=1}^{N} log \frac{\exp(sim(f_M(x_I^i),\ f_T(x_{T,+}^i))/\tau)}{\sum_{k=1}^{N} exp\left(\frac{sim\left(f_M(x_I^i),\ f_T(x_{T,+}^k)\right)}{\tau}\right) + \sum_{k=1}^{N} exp\left(\frac{sim\left(f_M(x_I^i),\ f_T(x_{T,-}^k)\right)}{\tau}\right)}$$

$$L_T = -\sum_{i=1}^{N} log \frac{\exp(sim(f_T(x_{T,+}^i), f_M(x_I^i))/\tau)}{\sum_{k=1}^{N} \exp(sim(\left(f_T(x_{T,+}^i), f_M(x_I^k)\right)/\tau)}$$

$$L = \frac{L_I + L_T}{2}$$



We also use negative sampling to introduce text variations, including negated captions in the batch but excluding them from loss calculation

Experiment

In fine-tuning result, we compare CLIP with Spatial Attention ***(Ours)*** to the original CLIP and CLIP with Residual Linear Connection

- Main Result

	Relation						All
	Close	Far	Left	Right	Above	Under	
Zero-shot evaluation							
CLIP	51.2	53.4	47.5	49.4	54.2	51.5	51.2
Fine-tuning evaluation							
CLIP	77.9	75.4	47.7	52.1	57.1	54.4	65.1
CLIP w/ Residual Linear Connection	77.9	78	55.6	54.7	63.9	58	68.2
CLIP w/ Spatial Attention (Ours)	**79.3**	**80.2**	**66.7**	**67.1**	56.9	**58.5**	**73.3**

- Ablation Study

	Relation						All
	Close	Far	Left	Right	Above	Under	
Fine-tuning evaluation							
CLIP w/ Spatial Attention (ours)	79.3	80.2	66.7	67.1	56.9	58.5	73.3
(-) Negative Sampling	80.8	80.7	67.0	68.2	56.3	53.9	72.4
(+) Residual Linear Connection on Value							

Conclusion & Future Work

- This work proposes a method to propagate query and key from a Vision Foundation Model to the CLIP Vision Encoder
- The approach method demonstrates improved fine-tuning performance on VQA tasks involving spatial relationships between objects in images
- As Vision Encoders gain importance with advancements in multimodal LLMs (MLLMs), this study is expected to contribute to extracting fine-grained information effectively
- For future work, we plan to utilize the proposed architecture as the vision encoder for a Multimodal Large Language Model (MLLM) and extend it through Instruction tuning