PHANTOM OF LATENT FOR LARGE LANGUAGE AND VISION MODELS

Byung-Kwan Lee
KAIST
leebk@kaist.ac.kr

Sangyun Chung
KAIST
jelarum@kaist.ac.kr

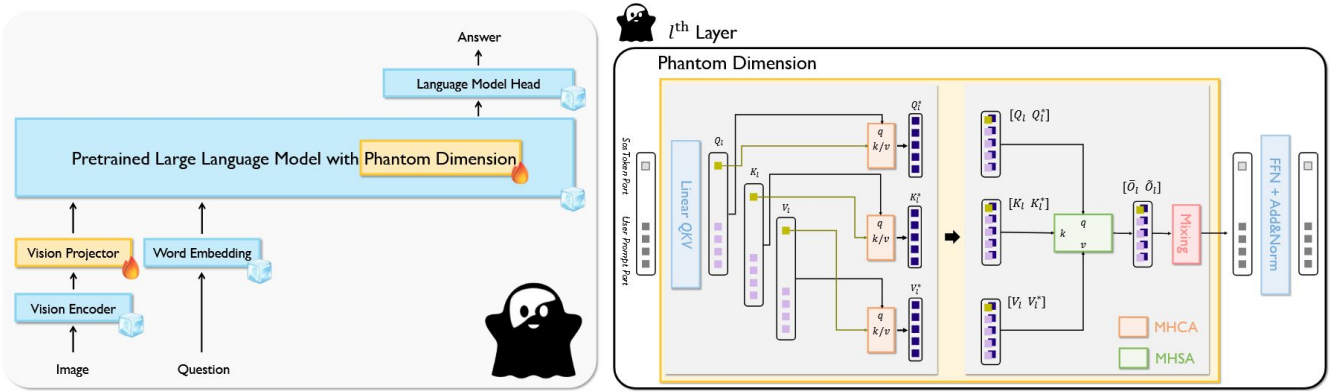
Chae Won Kim
KAIST
chaewonkim@kaist.ac.kr

Beomchan Park
KAIST
bpark0810@kaist.ac.kr

Yong Man Ro
KAIST
ymro@kaist.ac.kr

- Problem / objective
  - Efficient LLVM (모델 크기↓, 성능↑)
- Contribution / Key idea
  - Efficient LLVM family, **Phantom**, with model sizes of 0.5B, 1.8B, 3.8B, 7B.
    - Temporarily increase the latent hidden dimension during MHSA (Multi-Head Self-Attention)
    - Training strategy, **Phantom Optimization (PO)**, using the 2M number of **Phantom triples**.

- **Overview**



(a) Model Architecture (b) Phantom Dimension

Figure 3: (a) Overview of model architecture and the detail of first training step with Phantom Dimension and Phantom Optimization. In second training step, we train all of the parameters described in this figure. (b) Illuminating how Phantom Dimension temporarily enlarges the latent hidden dimension in forward propagation at l-th layer in Phantom, where 'Linear QKV', MHSA, and 'FFN+Add&Norm' is generally used module from pretrained LLM. Only MHCA module is added.

- **Overview of Model Architecture.**

❏ 구조: 1) vision encoder, 2) vision projector, 3) multimodal language model

1. vision encoder: InternViT-300M
2. vision projector: 2 fc layers with GELU
3. multimodal LLM: Qwen2-0.5B, InternLM2-1.8B, Phi3-mini-3.8B, and InternLM2.5-7B

- **Gathered Visual Instruction Tuning Sample Configuration.**
  - ❏ 2.8M visual instruction tuning samples across multiple datasets

- **Curation of Phantom Triples.**

1. 2M Phantom Triples
2. 구성: 1) question, 2) correct answer, and 3) confusing answer
3. Confusing answer 생성 과정:
    a. GPT-4o-mini 기반 오답 생성
    b. GPT-4o 기반 자동 검증
    c. 사람 검토

- **Realization of Phantom Dimension.**
  - ❏ dd