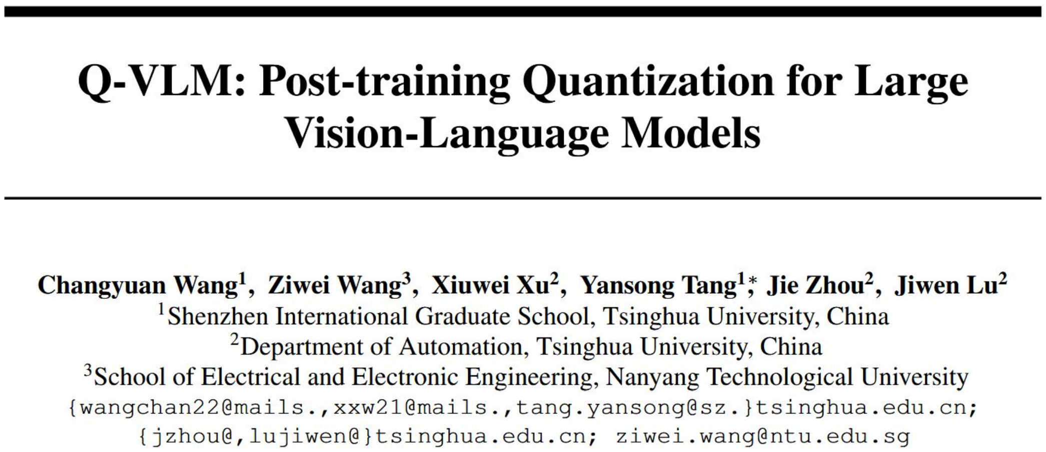
NeurIPS 2024



Q-VLM: Post-training Quantization for Large Vision-Language Models

Changyuan Wang[1], Ziwei Wang[3], Xiuwei Xu[2], Yansong Tang[1]*, Jie Zhou[2], Jiwen Lu[2]
[1]Shenzhen International Graduate School, Tsinghua University, China
[2]Department of Automation, Tsinghua University, China
[3]School of Electrical and Electronic Engineering, Nanyang Technological University
{wangchan22@mails.,xxw21@mails.,tang.yansong@sz.}tsinghua.edu.cn;
{jzhou@,lujiwen@}tsinghua.edu.cn; ziwei.wang@ntu.edu.sg

- Problem / objective
  - Efficient multi-modal inference

- Contribution / Key idea
  - Post-training quantization framework for LVLMs (Large Vision Language Model)

전유진

Wang, Changyuan, et al. "Q-VLM: Post-training Quantization for Large Vision-Language Models." *arXiv preprint arXiv:2410.08119* (2024).

NeurIPS 2024

Overview

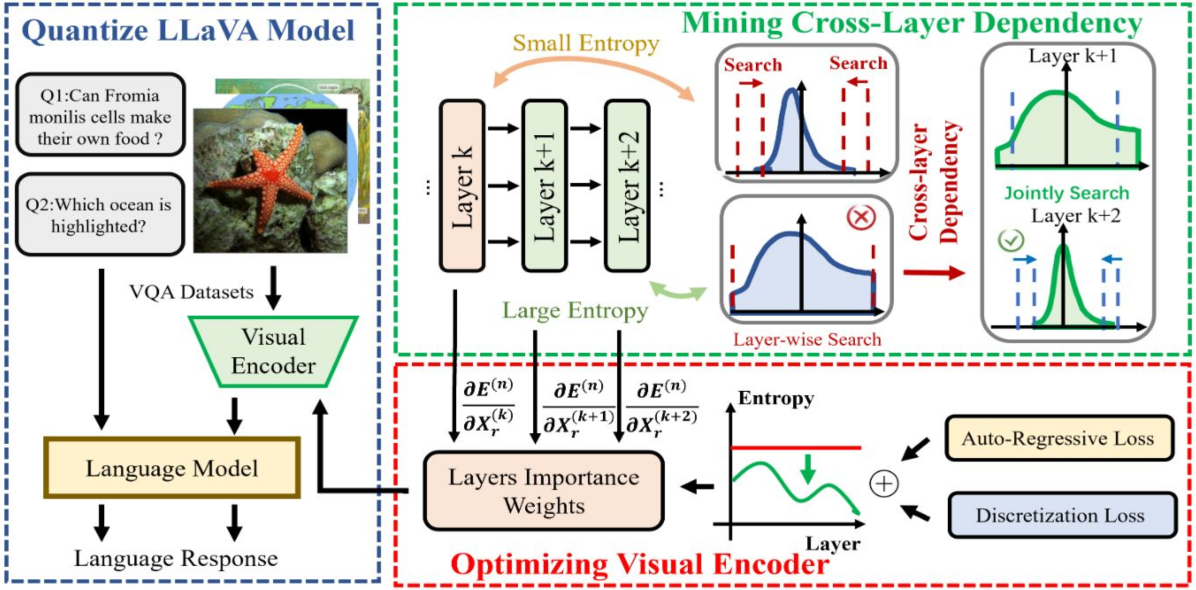

Quantize LLaVA Model
Q1:Can Fromia monilis cells make their own food ?
Q2:Which ocean is highlighted?
VQA Datasets
Visual Encoder
Language Model
Language Response
Mining Cross-Layer Dependency
Small Entropy
Layer k
Layer k+1
Layer k+2
Large Entropy
Search
Search
Layer-wise Search
Cross-layer Dependency
Layer k+1
Jointly Search
Layer k+2
$\frac{\partial E^{(n)}}{\partial X_r^{(k)}}$
$\frac{\partial E^{(n)}}{\partial X_r^{(k+1)}}$
$\frac{\partial E^{(n)}}{\partial X_r^{(k+2)}}$
Layers Importance Weights
Entropy
Layer
Auto-Regressive Loss
Discretization Loss
Optimizing Visual Encoder

Figure 1: The overall pipeline of our method. We employ entropy as the proxy to represent cross-layer dependency for efficient block assignment, which decomposes the large search space from the entire model to blocks containing multiple layers. Moreover, the visual encoder is further optimized for fine-grained search space decomposition.

전유진

Wang, Changyuan, et al. "Q-VLM: Post-training Quantization for Large Vision-Language Models." *arXiv preprint arXiv:2410.08119* (2024).

NeurIPS 2024

Preliminaries - Post-training Quantization for LVLMs

- Global Optimization : 정확하지만 탐색 비용 너무 크다.

$$\min_{\{Q_k\}} \quad J = \left\| W_q^{(n)} X_q^{(n)} - W_r^{(n)} X_r^{(n)} \right\|_2^2$$
$$s.t. \quad X_q^{(k+1)} = Q_k(W_q^{(k)} X_q^{(k)}) \qquad (1)$$

- Greedy Layer-wise Optimization : 계산은 빠르지만 cross-layer dependency 무시로 인한 오차 누적된다.

$$\min_{Q_k} \quad J = \left\| W_q^{(k)} X_q^{(k)} - W_r^{(k)} X_r^{(k)} \right\|_2^2 \qquad (2)$$

Ours - Mining Cross-layer Dependency for LVLM Quantization

- Block-wise Quantization 제안

Wang, Changyuan, et al. "Q-VLM: Post-training Quantization for Large Vision-Language Models." *arXiv preprint arXiv:2410.08119* (2024).

NeurIPS 2024

Optimizing Visual Encoders for LVLM Quantization

전유진