Lightweight Vision Transformers via Knowledge Distillation and Mixed-Precision Quantization

Background

Despite excellent accuracy, Vision Transformers are computationally heavy. Knowledge distillation (KD) and mixed-precision quantization offer promising ways to maintain accuracy while lowering complexity [Hinton et al., 2015; Zhang et al., 2023].

Problem Description

Develop a **student-teacher ViT compression framework** using KD and mixed-precision quantization. The student model should preserve visual quality while operating with reduced bit precision per layer.

Milestones and Extensions

- Select teacher ViT (e.g., DeiT-B) and define lightweight student variant.
- Implement KD loss combining logits + feature alignment.
- Apply per-layer quantization (4–8 bit) using QAT.
- Evaluate accuracy, FLOPs, and compression ratio.
- **Extension:** hardware deployment using TensorRT or mobile GPU.

Tools, Qualifications, and Outcomes

- **Skills:** PyTorch, ViT, quantization, KD theory.
- **Tools:** timm models, TensorRT, HPC GPU.
- **Outcomes:** Lightweight ViT achieving near-teacher accuracy with large efficiency gains.

References

- Hinton, G., Vinyals, O. and Dean, J., 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and Jégou, H., 2021, July. Training data-efficient image transformers &

distillation through attention. In *International conference on machine learning* (pp. 10347-10357). PMLR.

- Liu, Y., Yang, H., Dong, Z., Keutzer, K., Du, L. and Zhang, S., 2023. Noisyquant: Noisy bias-enhanced post-training activation quantization for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 20321-20330).