Token-Level Entropy Modeling for Efficient Vision Transformers

Background

Most ViT compression efforts focus on weight pruning or quantisation, neglecting **token entropy** during inference. Recent works (e.g., TokenMerging [Bolya et al., 2023]) show that redundant tokens can be dynamically merged without significant accuracy loss.

Problem Description

Develop an **adaptive token-pruning/merging framework** for transformer-based vision networks aimed at compression efficiency. The student will implement entropy-based token importance estimation, enabling dynamic computation and storage reduction.

Milestones and Extensions

- Analyse entropy of token representations in pre-trained ViTs.
- Implement token importance scoring using Shannon entropy or attention maps.
- Integrate dynamic token removal into ViT inference.
- Measure computational savings and accuracy retention on ImageNet subset.
- **Extension:** integrate hardware-aware latency modeling.

Tools, Qualifications, and Outcomes

- **Skills:** Python, PyTorch, transformer internals, statistics.
- **Tools:** timm ViT models, OpenMMLab, HPC GPU.
- **Outcomes:** Efficient ViT inference with ≥30 % token reduction at minimal performance drop.

References

- Feng, Z. and Zhang, S., 2023. Efficient vision transformer via token merger. *IEEE Transactions on Image Processing*, *32*, pp.4156-4169.
- Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J. and Hsieh, C.J., 2021. Dynamicvit: Efficient vision transformers with dynamic token

sparsification. *Advances in neural information processing systems*, *34*, pp.13937-13949.

- Kim, S., Shen, S., Thorsley, D., Gholami, A., Kwon, W., Hassoun, J. and Keutzer, K., 2022, August. Learned token pruning for transformers. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 784-794).