Deep Learning CS 577 Project Proposal

Project Title

Image classification using Vision Swin Transformers

Team Members

1. Raghunath Babu(A20511598)
   rbabu@hawk.iit.edu
2. Naveen Raju Sreerama Raju Govinda Raju(A20516868)
   nsreemarajugovinda@hawk.iit.edu

Description of the problem

Handling high dimensional images (i.e high resolution images) with poor image quality is a big adversary to efficiently leverage architectures like convolutional neural networks (and deep learning), RNNs and state of the art vision transformer. Even though Vision transformers thrive in such applications, it requires a higher computational power to capture the global context and dependencies. This demands an architecture that uses a hierarchical structure which has a better ability to work on high dimensional images with lower computational power thereby capturing global dependencies within an image. Moreover, for these applications it requires some additional features to handle various input size(scalability) and adapt to a variety of computer vision applications. This brings us to explore the potential of swim transformers.

A concise overview of prior efforts and an explanation of what sets the proposed work apart.

For the applications mentioned above, so far vison transformers, convolutional neural networks and hybrid networks of multi-layer perceptron and transformers have been used at different scale of complexity and optimization.

The proposed work on Swin Transformers would capture long range dependencies in images effectively at lower computational power. The traditional CNN'S (with a shallow network) struggles to capture long range dependencies and fails to handle images at different scales. Going deeper may capture the desired objective but with a higher computation. Whereas the Swin Transformers with shallow layer with much less computation capture long range context in the image

Preliminary plan

1. Data set collection
2. Data Analysis and cleaning

3. Implementing vision Transformers
4. Training and generating evaluation metrics.
5. Comparing metrics of different ViTs
6. Analyzing the FPs and FNs and fine-tuning hyper parameters for improvement

Reference Papers

1. D. Kang, P. Koniusz, M. Cho and N. Murray, "Distilling Self-Supervised Vision Transformers for Weakly-Supervised Few-Shot Classification & Segmentation," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023 pp. 19627-19638.
2. Agilandeeswari, L., Meena, S.D. SWIN transformer based contrastive self-supervised learning for animal detection and classification. *Multimed Tools Appl* **82**, 10445–10470 (2023).
3. Z. Liu, et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021 pp. 9992-10002.
   doi: 10.1109/ICCV48922.2021.00986
4. Alexander Kolesnikov and Alexey Dosovitskiy and Dirk Weissenborn and Georg Heigold and Jakob Uszkoreit and Lucas Beyer and Matthias Minderer and Mostafa Dehghani and Neil Houlsby and Sylvain Gelly and Thomas Unterthiner and Xiaohua Zhai.," An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale",