Deep Learning CS 577

Intermediate Project Proposal

Project Title

Image classification using hierarchical based shifted window Vision Transformers.

Team Members

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

Problem Description

The motivation of this project is to handle a predominant adversary in the field of computer vision in handling high dimensional images with poor image quality and thereby preserving the computational efficiency. Prominent approaches with Convolutional Neural Networks(CNNs), Deep learning networks, RNNs, LSTMs and state of the art Vision Transformers were leveraged for this use case. The problem they faced with CNNs, DNNs, RNNS etc were, there was a need to add several layers to support the dimension of the input. This increased the parameters of the model, resulting in a burden to computation which were considered unnecessary. Also in applications with huge datasets CNNs failed to capture long range dependencies and global context. When Vision Transformers were used, it helped in capturing global dependencies within an image, but with huge computation efforts.

During the initial processing of the high dimensional images, where converting the images to image patches for tokenization, this vision transformer resulted in a lot of key, value and queries and performing self attention demanded a lot of computation with huge efforts in training. The model resulting would be too complex and would be practically difficult to host and perform api calls. This boils down to our approach where we want to get the underlying global dependencies in an image and minimize the computation.

In this project we are focused on implementing a hierarchy vision transformer where we define windows as a collection of image patches(similar to the regular vision transformer) and this window size will be hierarchically decreased. Another advantage in the proposed approach is that we can leverage several architectures like Feature pyramid network(FPN), U-Net etc. The self

attention would be performed within the local window and the window would be shifted within the images subsequently. A cyclic shifted window technique(resulting in non overlapping windows) is implemented to capture the global context.

Analyzing this, we could infer that the properties of the Vision transformers are preserved and quadratically decreasing computational complexity.

Data Description:

- Data set Source : ETH Zurich (https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/)
- Data Set used “Food 101”.
- Number of classes : 101 food categories
- Data set size : 101,000
- Each class has 1000 images
- Train and Test set split: In each class we consider 750 images for the train and 250 images for the test.
- Data set used in previous work:

  This data set was initially used in the paper titled “Food-101 – Mining Discriminative Components with Random Forests”. In this research, the authors use a machine learning technique called Random Forests to identify the key features that distinguish different food items. By analyzing a dataset of 101 food categories, they aim to find the most important characteristics that can help classify and distinguish various foods.

- Reason for selecting this dataset:

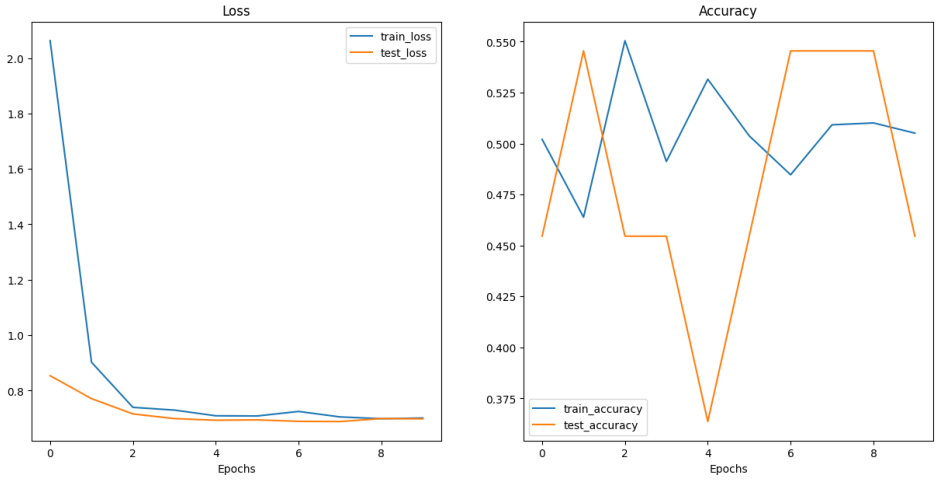
  The Food-101 dataset's unique characteristics, such as high inter-class variability, diverse backgrounds, and the need to capture long-range dependencies in food images, make it an ideal candidate for showcasing the benefits of Swin-Transformers/Vit Transformer. Traditional CNNs may struggle with the dataset's wide variety of food categories and complex visual distinctions between them. In contrast, Swin Transformers leverage their self-attention mechanisms to capture fine-grained details and complex patterns, enabling them to effectively differentiate between visually diverse food categories. Moreover, they excel at handling unwanted background noise by selectively attending to the relevant food item while ignoring clutter, enhancing robustness and accuracy in food recognition. Additionally, their ability to model long-range dependencies makes them well-suited for recognizing complex food structures, which is crucial in a dataset like Food-101.

  By training a Swin Transformer/Vit Transformer on the Food-101 dataset, you can demonstrate how this architecture excels in addressing real-world challenges in image classification, particularly in domains with high visual complexity and variability, such as food recognition. Swin Transformers' capacity to adapt to diverse food categories, mitigate background interference, and capture intricate details and long-range relationships within

images makes them a powerful tool for accurate and robust food classification, showcasing their effectiveness in handling complex and variable image data

What have we done so far:

- Read and understood “Vision Transformer” and “Swin Transformer” research papers, technical blogs. And also understood technical details of model architectures watching youtube videos.
- Finding the proper dataset category needed to showcase our project motive. The other best data set that we came through is caltech-ucsd_birds-200-2011 where it has 200 bird categories with a lot of inter class variations and other structural complexities.
- Tried to implement custom architecture of standard Vision Transformer and Swin Transformer
- Trained the code on a simple dataset i.e “Daisy and Dandelion” dataset just to check working of code. The below figure is the trial run metrics.



What remains to be done

- Implement an end to end Vision Transformer and Swim transformer specific to the Food Dataset.
- Dataset cleaning
- Fine tuning various paraments(hyper parameter)
    - Optimizers
    - Number of heads in encoder
    - Learning rate
    - Regularization
    - Batch normalization

- Compare the performance with ViT and Swin transformer metrics using pretrained weights.