Depth-Wise Algorithm Implementation On Attention CNN - Brief Report

IIT JODHPUR

Problem Statement

The aim of the project is to implement CNN models with some changes to existing standard CNN models so that they can be re-configured according to the need of edge devices without affecting too much accuracy.

Readme

The code is written in colab and thus can simple be run again by uploading the python files and the VisDrone dataset in colab and run normally everything will be automatically installed in colab and user only need to run all the cells to regenerate the result otherwise only to check the work if colab is uploaded results will be shown of last epochs they were trained on.

Approaches for Reconfigurable DNN

4 approaches have been implemented which are based on two main concepts.
1) Depth-wise separable Convolution
2) Linear Bottlenecks

Techniques used for Reconfigurable DNN

The main focus was to reduce the size of the model and this can be achieved by reducing the number of parameters this would cause accuracy to reduce but size of model will also be reduced.

Approach 1

Keeping the baseline model we look upon the convolution layer with maximum number of parameters and switch it with depthwise convolution layer and here we are replacing only 1 convolution layer from the block of the model.

Approach 2

We will now change multiple convolution layers from each block with the maximum number of parameters in that block of model.

Approach 3
We will now integrate reverse linear bottlenecks along with depth wise Convolution.

Approach 4
This will be the hybrid approach of all convolutional layers as depthwise convolutional layers with bottleneck blocks.

Results

Models:
CNN - Baseline CNN Model
CNN_v1 - Approach 1 implemented on Baseline CNN model
CNN_v2 - Approach 2 implemented on Baseline CNN model
CNN_v3 - Approach 3 implemented on Baseline CNN model
CNN_v4 - Approach 4 implemented on Baseline CNN model

Comparison Of Parameters And Model Size For Different Approaches

Model	Total Parameters	Size (MB)
CNN	23,653,314	90.23
CNN (With Approach 1)	21,565,378	82.27
CNN (With Approach 2)	20,199,776	77.06
CNN (With Approach 3)	21,733,058	82.91
CNN (With Approach 4)	20,367,456	77.70
Attention CNN	29,444,704	112.32

The reason why an attention CNN (Convolutional Neural Network) typically has more parameters than a regular CNN is because of the additional attention mechanism.

In a regular CNN, the convolutional layers apply filters to extract spatial features from the input data. These filters are shared across the entire input, leading to a relatively smaller number of parameters compared to the input size. The output of the convolutional layers is then typically fed into fully connected layers, which can have a large number of parameters depending on the size of the layer.

On the other hand, an attention CNN incorporates an attention mechanism, which allows the model to focus on specific regions or features of the input data that are considered more important. The attention mechanism introduces additional parameters to learn the attention weights or scores for different parts of the input. These attention weights are used to weight the importance of each feature map or spatial location in the convolutional layers.

The attention mechanism effectively introduces additional learnable parameters to the model, leading to a higher parameter count. This allows the attention CNN to dynamically attend to different parts of the input based on their relevance, potentially improving the model's performance by giving more emphasis to important features.

Overall, the higher parameter count in an attention CNN compared to a regular CNN is due to the added parameters associated with the attention mechanism, which enable the model to selectively focus on relevant information during the learning process.

Also, as clearly seen from the table, it is conclusive that the implementation of Depthwise Algorithm on the CNN Baseline model has been successful and has reduced the number of parameters without affecting the accuracy much.