AIDI_1002_Final_Project_Template

December 15, 2023

1 AIDI 1002 Final Term Project Report

Members:

Thomas Shank - 200346862 Thomas.Shank@MyGeorgian.ca

Leon Quadros - 200532151 Leon.Quadros@MyGeorgian.ca

2 Introduction

2.1 Problem Description

Image classification involves assigning labels to images based on different categories, such as animals, vehicles or objects. Various patterns of pixels are decoded to numerical values and classified. These classifications of numerical vectors help models generalize input images and predict their content. However, this process is computationally exhaustive as it requires many steps including image processing and targetting (setting a bounding box).

The challenges of image classification are complex. These include lighting conditions, backgrounds, resolutions, and perspectives within the image. These properties **obfuscate the target within the image** where image complexity scales computational efforts exponentionally. For example, the difference between predicting black and white images (single channel) versus multi-channel (RGB) images of different categories, is considerable.

A major computational challenge in image classification is determining the location of the target within the background. The depth of this challenge increases exponentionally with image resolution and other image related attributes. Many modern models attempt to use Bounding-Box Regression, Region-Based CNN's, or RetinaNet. Which require extra steps resulting in extra mathematical operations.

2.2 Problem Context

Image classification is verbose, and requires acute attention to specific details. The verbosity of the problem expands with the complexity of the image, the more detailed the image, the harder it is to determine what the model is trying to predict. One major problem in image classification that this paper focuses on is locating the object in the image to be classified. Most traditional image classification models use tedious calculations to locate the target within the image to be classified, and this is a major loss when it comes to training over large datasets

Image classification is important because it is a foundational step in other applications such as object detection, medical imaging or content filtering. It is the first step in machine learning that allows systems to interpret and understand visual data. It's applications span a wide range of industries, contributing to advancements in many technological fields and simplifying many image combing related processes. Being able to quickly classify images can mean the difference between waiting 5-10 seconds or experience real-time classification.

2.3 Limitations of Other Image Classification Approaches

The limitations of image classification algorithms often revolve around a critical factor: speed. High computational complexity in determining the location of a target within an image introduces various constraints. The efficiency of target localization significantly influences the practical application of a specific algorithm.

For instance, if speed is not a primary concern, and there's ample time available for processing, implementing a system based on R-CNN (Region-Based Convolutional Neural Networks) may offer high accuracy. This can be advantageous in applications such as identifying cancerous skin patches or spots, where precision (technically recall) counts.

However, slower algorithms like R-CNN, Bounding-Box Regression, and Attention Mechanisms may not be suitable for systems requiring real-time predictions, such as self-driving vehicles or live object detection. In scenarios where immediate responses are crucial, faster models like YOLO (You Only Look Once) might be preferred due to their real-time capabilities.

It's essential to strike a balance between speed and accuracy based on the specific requirements of the application, considering factors such as processing time, real-time constraints and the nature of the target being detected.

The model proposed by Cordonnier et al. (2020) shares similarities with our selected paper, particularly in extracting 2 × 2 patches from the input image and applying full self-attention. This approach resembles the Vision Transformer (ViT). However, here the research extends beyond this by demonstrating that extensive pre-training enables vanilla transformers to compete with or outperform convolutional neural networks (CNNs). Additionally, Cordonnier et al. (2020) limits its applicability to small-resolution images due to a 2 × 2 pixel patch size.

2.4 Solution

The solution proposed by `An image is worth 16x16 words` for classifying images is called `patching`. Instead of scanning over a rasterization of pixels to determine the target, this paper suggests dividing the image into equal sized patches.

This optimizes the process of determining the label of the image; where instead of using globalized retention over the entire rasterization of individual pixels, it attempts to contextualize the information between patches and build an understanding of the image through clusters.

**This approach is very similar to YOLO (You only look once) algorithm where it divides the image into equal size chunks and attempts to predict the bounding box based from each chunk.*

These image patches are treated similar to an NLP token in a supervised environment. Where instead of a vectorized matrix of features related to a target label, we have a vector of bytes that are formatted/permuted depending on the images input channels (Grayscale=1, RGB=3).

This approach allows us to quickly iterate over images during the training process compared to other models that spend a lot of processing power in locating the target in the image through bounding box placement techniques.

# Background

Reference	Explanation	Dataset/Input	Weakness
https://www.youtube.com/watch?v=Vonyoz6Yt9c	Constructing a convolutional neural network from scratch that implements a patching process and an embedding transformer.	Numerical Grayscale Dataset	Inconsistent accuracy with colorized datasets that include more intensive boundary analyzation.
https://www.cs.toronto.edu/~kriz/cifar.html	The CIFAR-10 dataset (Canadian Institute For Advanced Research) is a collection of images that are commonly used to train machine learning and computer vision algorithms.	CIFAR-10 Dataset (60,000 images)	The dataset has fairly limited complexity containing 10 classes, which might not be sufficient for more complex multi-class classification problems
https://arxiv.org/abs/1810.03505	CINIC-10 dataset is a plug-in extended alternative for CIFAR-10. It was compiled by combining CIFAR-10 with images selected and downsampled from the ImageNet database.	CINIC-10 Dataset (270,000 images)	The quality and resolution of images in the dataset can impact the performance of models trained.

3 Methodology

- Image Patching: The authors split an image into fixed-size patches (16x16 pixels in this case). These patches are treated as the equivalent of words in NLP.
- Patch Embedding: Each image patch is linearly embedded into a vector. The sequence of these vectors forms the input to the Transformer.
- Transformer Model: A standard Transformer is applied directly to the sequence of embedded patches. The Transformer consists of multiple layers of self-attention and feed-forward neural networks.
- Pre-training and Fine-tuning: The ViT model is pre-trained on a large dataset (like ImageNet) and then fine-tuned on the specific task.

The authors show that this approach, when pre-trained on large amounts of data and transferred to multiple recognition benchmarks (like ImageNet, CIFAR-10, etc.), attains excellent accuracy, matching or outperforming the best convolutional networks while requiring substantially fewer computational resources to train.

Basically demonstrating that reliance on convolutional networks is not necessary for image classification tasks.

4 Implementation

1. Need to add code for replication of the paper

(Working) Vit pretrained model pepared on CIFAR-10, and tested against CINIC-10: ViT Base Pretrained

2. Code for our contribution and comparison on results. If half done, then include what can/could've been done in the 'Future Direction' section.

(Failed) Working code with extremely poor accuracy can be found in: ViT Torch Untrained

5 Conclusion and Future Direction

Image classification in detailed pictures is a difficult task to overcome. Our efforts to deploy a custom nn model built with torch from the assistance of youtube did not produce any fruitful results. One thing that was beneficial to beginning with this approach however, was the understanding of the research paper that came along with a manual implementation.

By hand writing a PatchEmbedding model and utilizing a Convolutional 2D Neural Network, we were able to gain a clear insight into the process of a ViT transformer. We gained an understanding of how images are preprocessed, considering the images color channels and resolution.

The ViT transformers concept is quiet simple when you understand it, but implementing it in a way thats compatible with different image formats requires tuning/coding relevant to those image formats. When done effectively on a 3+ channel image at at least 28x28 resolution, classification generalization starts to take place where the model begins to better understand images not native to its initial training format.

After a few failed attempts at modifying images resolutions, downscaling their channels, or rotating them, we investigated employing pre-trained vision transformers directly for image classification. Vit proves as a straightward yet scalable apporach thats remarkebly effective, especially when combined with pre-training on large datasets.

The vision transformer not only matches but surpasses the current state of the art performance on numerous image classification datasets and all while being relatively cost-effective in the pre-training phase. The process of image patching may not be resource effective, but its increase in speed is not to be underestimated as it can quickly train and validate over extensive sets of data.

6 References

- An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby

- Learning Multiple Layers of Features from Tiny Images, Alex Krizhevsky, 2009
- YouTube. (2023, September 29). Implement and train VIT from scratch for image recognition - pytorch. YouTube. https://www.youtube.com/watch?v=Vonyoz6Yt9c
- YouTube. (2020, October 4). An image is worth 16x16 words: Transformers for image recognition at scale (paper explained). YouTube. https://www.youtube.com/watch?v=TrdevFK_am4