CNN Algorithm Optimization

Examining Inefficiencies within the Pooling Process

Nick Fogg
Computer Science
NC State University
Raleigh NC USA
nsfogg@ncsu.edu

David Sweasey
Computer Science
NC State University
Raleigh NC USA
dbswease@ncsu.edu

IDEA

The goal of our project is to optimize the performance of a Convolutional Neural Network (CNN). By examining and reproducing the steps in the current CNN workflow, we plan to expose any inefficiencies in performance time or memory management.

During a preliminary examination of the current CNN structure, it was identified that there are potential slowdowns in the algorithm, specifically regarding the pooling layers. Max pooling tends to outperform average pooling, but it does so with a loss of information. Too much loss in detail can result in a much lower model accuracy [1]. To address this challenge, we are considering altering the current approach to pooling in the CNN algorithm. In the current model average pooling has better loss yields, while max pooling has higher loss resulting in more back propagation. A possible solution can use a different approach to maintain efficiency and accuracy. This will lessen the effect of backpropagation and improve performance.

MOTIVATION

CNNs are primarily used for computer vision: the perception of the world as humans by computers. [1] In real-world applications, both efficiency and accuracy of computer vision is paramount. Pooling layers are a primary factor in computational speedup within a CNN. They perform a "dimensional shrink" to significantly decrease the number of computations that must be completed. However, much of the previous layer's information was lost, and in some cases, this information contained pertinent details. [1]

We want to explore ways of optimizing pooling layers to minimax computational overhead and output accuracy. In many models, the hyperparameters serve as the basis for how pooling layers behave. The parameters include filter size and stride. [1] Fine-tuning these parameters takes time, and the efficacy of these parameters determines the accuracy and efficiency of the CNN. We hope to speed up the determination of these parameters while maintaining accuracy.

Once the parameters are set, another problem arises: data preservation. The different forms of pooling yield differing results. Average pooling preserves a bit of all data by computing the average value of a region whereas max pooling returns the largest value. We hope to find a healthy balance between the two methods, potentially with an additional parameter corresponding to a weighted average that combines the benefits of average and max pooling.

By the end of the project, we hope to find optimizations in hyperparameter generation and pooling layer methods such that our CNN is accurate, efficient, and training requires fewer iterations of backpropagation.

TIMELINE

2/23/24: Proposal Deadline

Finalize project idea for project. Create proposal outlining project idea, motivation, goals, and timeline.

3/08/24: CNN Research

Research current CNN models to understand algorithm involved. Search inefficiencies within pooling algorithms.

3/29/24: Pooling Optimization

Implement researched strategies and compare performance to current CNN standards, document results.

4/12/24: Presentation Deadline

Prepare presentation. Discuss project overview, motivations for work, attempted optimization methods, and findings.

4/26/24: Final Report Deadline

Summarize project findings into a final report.

REFERENCES

[1] Saha, S. (2022) *A comprehensive guide to Convolutional Neural Networks - the eli5 way*, *Medium*. Available at: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 (Accessed: 21 February 2024).