Modern Hopfield Networks on the CIFAR10 dataset

Haosheng Wang, Edrick Guerrero, Alfonso Gordon Cabello de los Cobos

Introduction

Memory involves the efficient storage and retrieval of information, and it comes in various forms—short-term, long-term, sensory, procedural, among others. Hopfield networks, also known as associative memories, are a class of recurrent neural networks (RNNs) designed to function as content-addressable memory systems. A defining characteristic of these networks is their ability to reconstruct entire patterns from partial or noisy inputs.

The original Hopfield network, introduced by John J. Hopfield in 1982, was based on binary feature representations and binary activation functions. Since then, significant advancements have been made. Modern Hopfield networks generalize the original model to continuous states, dramatically increasing storage capacity and stability. These developments, particularly in networks with continuous dynamics and large memory capacity, have been explored in a series of works since 2016.

In this project, we focus on replicating and analyzing the 2020 paper *"Hopfield Networks is All You Need"*, which presents a modern formulation of Hopfield networks that bridges them with attention mechanisms commonly used in deep learning nowadays.

Related Work

There has been recent work on applying human memory mechanisms into LLMs. This one specifically identities three major drawbacks with current popular LLMs: excessive training data and power consumption, forgetfulness, and a lack of logical reasoning capabilities within black-box models. It mentions new and innovative architectures that mimic many of the memory functions of the human brain, such as spiking neural networks, which allow for asynchronous learning and are inspired by the brain's synaptic plasticity ("ability of synapses to strengthen or weaken over time, in response to increases or decreases in their activity"). The paper also identifies challenges with mimicking the human brain, such as the reconstruction theory, which states that human memory is a reconstruction rather than a copy of experiences and that human memories are spatially decomposed into different regions. This is all while consuming very little power (20-23W) which is about 5 times lower than modern LLMs. The paper concludes that further research into “machine memory intelligence” can address current models’ drawbacks and also has the potential to contribute back to brain science.

https://www.sciencedirect.com/science/article/pii/S2095809925000293

Data

The dataset used for this project will be the CIFAR10, which consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. Some of these images will be processed by adding noise to them, creating a parallel dataset that will be used for testing the accuracy of the different goals.

Methodology

Our model will be a Modern Hopfield Network, a network useful for remembering images whose nodes are connected to each other. This model weights are mainly trained based on the energy function explained above.

A part of the CIFAR10 dataset will be cloned and those images will be edited by adding noise to them or by cutting them. Later, the original images will be used for the model to memorize them. The testing will consist of giving the model the noisy images for it to retrieve the originals based on its memory.

The most difficult parts of this implementation are:

1. Computational power: as this dataset is really big, the use of even a portion, combined with the memory stored by the model, will require a great computational power.
2. Translation: The original model was implemented using the PyTorch library, in our case we will use TensorFlow. Although this libraries are really similar, the way their functions are declared and work are not exactly the same, leading to a careful check of the documentation while implementing our version.

Metrics

To evaluate our implementation, we will use the metric of “accuracy”. The accuracy can be measured in the following 2 ways:

1. Measuring the average number of cases where the similarity between the retrieved image and the stored image is above a certain threshold
2. Summing (1-similarity) for each case, this can be viewed as a loss function

The original paper wants to show the following three properties of their new energy function:

1. Global convergence to a local minimum
2. Exponential storage capacity
3. Convergence after one update step

In this project, we decided the following goals:

1. BASE: high accuracy with small part of the dataset
2. TARGET: analyze how accuracy is affected by the quality (i.e. correlation between images) and quantity (number of images to store) of the stored images/features
3. STRETCH: analyze how accuracy is affected by the quality (e.g. amount of noise, completeness of features) of the images/features used for retrieval

Ethics

Immediately, the question of privacy and surveillance comes up with this kind of technology. Being able to re-construct a large block of data from a smaller and noisier block of data has many applications, many in which case can be used “for good” such as medical diagnoses, but applying this to re-construct faces, biometric data, and more is much more glaring. With the United States specifically, much of the government technology that is developed under contract overseas and that is used in tandem with human rights violations is the same technology that is currently being used on American soil. Many of the arguments used in support of such technology involves criminal justice, but it’s hard to have a genuine conversation with this argument while ignoring that investment in community, education, social nets, and social justice has been shown to be much more effective than any kind of surveillance technology.

It’s obvious there’s been a push for more surveillance in the United States, which can be supported by the fact that many local governments have considered banning masks (including medical masks) in public for “safety” all while ignoring the fact that medical masks are a proven pillar of public safety and disability solidarity. With this technology, it would be possible to re-construct a face, even if the input face is wearing a mask.

This brings up the question of whom the stakeholders of this technology are: Governments (including foreign governments), civilians, companies (defense contractors), people registered as criminal offenders, people not registered as criminal offenders.

There are many consequences of mistakes made by our algorithm. For example, if our algorithm was developed to re-construct faces and used by the government to deport people, our algorithm could inaccurately construct a face that accurately resembles a real person, in which case the government could justify sending that person to a foreign prison without due process and deny any wrongdoing.

Division of Labor

Haosheng Wang: training
Edrick Guerrero: evaluating
Alfonso Gordon Cabello de los Cobos: preprocessing