Gradient Episodic Memory (GEM) for Continual Learning

CS677 Deep learning – fall 2020

Fernando Rios | Hassan Ouanir | Maha Faruque

Introduction

The capability of continuously learning, improving, and sharing information is known as lifelong learning. Throughout a human's lifespan, they are learning from day one and building upon knowledge previously learned and lastly, sharing that knowledge with others. To always learn and grow is built into Humans and that's why Humans are lifelong learners. Lifelong learning is a set of neurocognitive mechanisms that build upon our sensorimotor skills, memory consolidation, and recover. As time goes on, technology is advancing and so is the amount of information we produce daily. Over 2.5 quintillion bytes of data is created every day and with all that data, it is pivotal for computational systems to be able to have lifelong learning capabilities. Unfortunately, there are many barriers for neural network models and machine learning for the lifelong learning capabilities. A very crucial challenge that arises with this is, catastrophic forgetting/interference.

Catastrophic forgetting is when the neural network tends to completely or suddenly forget previously learned information when learning new information which as a result, creates a problem for deep neural network models. Imagine if you learned something new in class today but then you suddenly forgot what you learned last week. It creates a big problem and therefore, many people have spent a lot of time on finding a solution for catastrophic forgetting. The three neural network approaches for solving catastrophic forgetting in lifelong learning are:

- **Regularization:** avoids overfitting and decreases the error by fitting a function accurately on the training data. Regularization has a penalty when changing the weights of the Neural Networks. The penalty increases proportionally with the weight change. Using regularization helps with keeping the previous knowledge.

- **Dynamic Architecture:** when learning a new task, the new resources are assigned to learn the new information, and this ultimately helps with learning new information without forgetting the previously learned information.
- **Complimentary Learning System (CLS):** CLS is a combination of both regularization and dynamic architecture. Not only are new resources created but few of the old weights are also changed. Again, this helps with learning new information while keeping the previously learned information as well.

Our project goes over the Gradient Episodic Memory (GEM) which uses the CLS Neural Network approach.

Gradient Episodic Memory (GEM)

The CLS approach tries to solve the catastrophic forgetting issue while also allowing sharing of knowledge to previous information. The experiment is on the MNIST dataset and the GEM model being used for working on overcoming some challenges which include catastrophic forgetting. GEM is a continual learning model in which each gradient updates for the current task and uses quadratic programming and helps solve catastrophic forgetting for previously learned tasks. The main attribute of GEM is to reduce forgetting, an episodic memory is needed to store parts of the learned information from a task. This approach demands much more memory than other approaches during training time but in return, runs greater in a single pass setting.

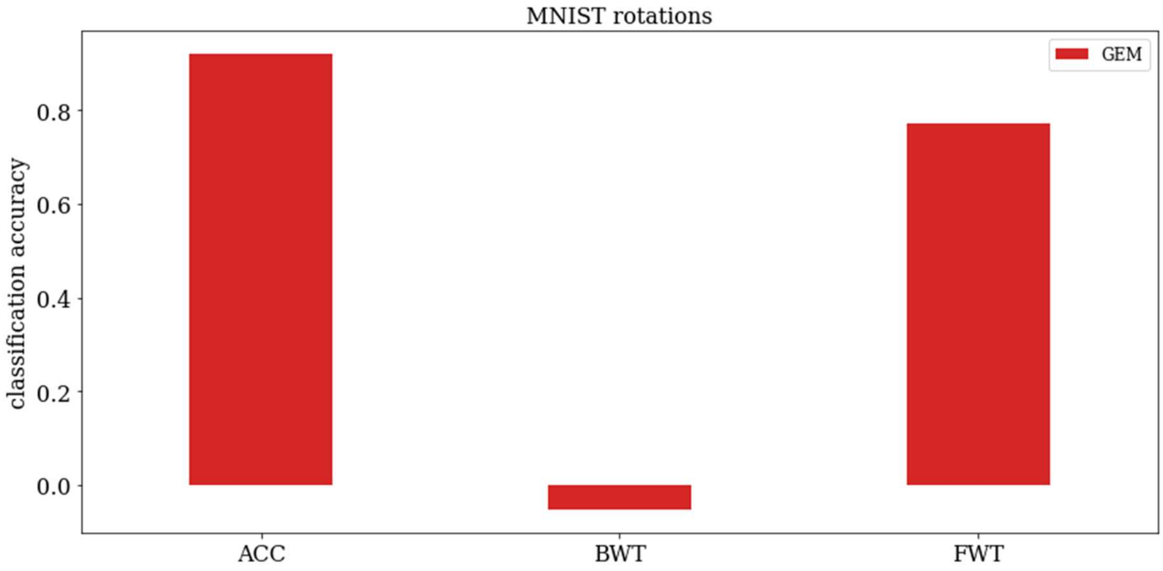
Dataset and Architecture

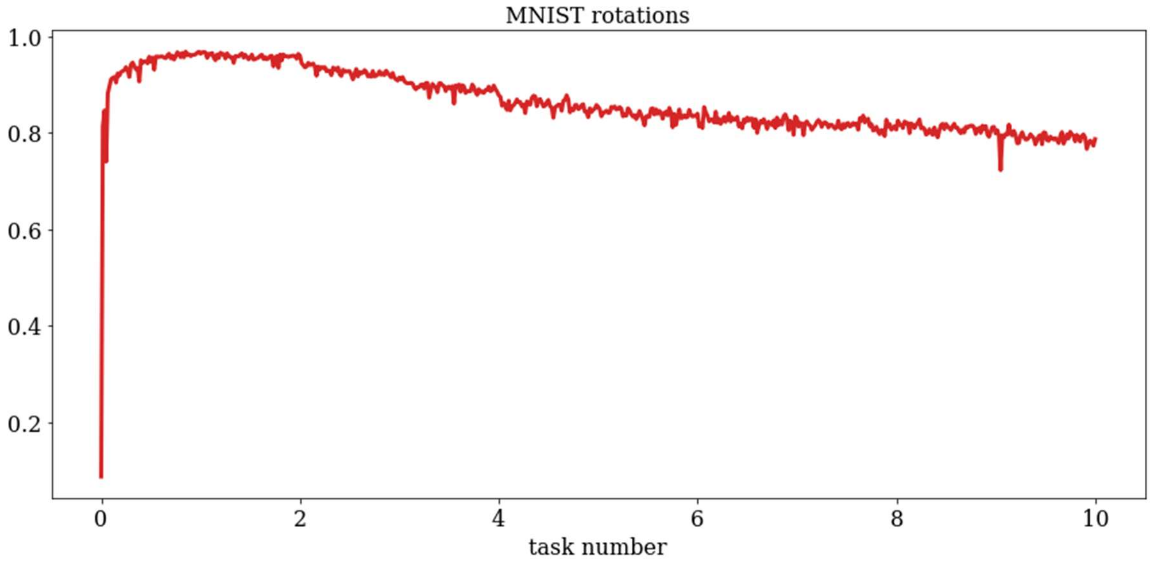
We used the MNIST dataset with T=10 tasks. The MNIST dataset consists of handwritten digits that is often used for training and testing for machine learning. Each image is 28*28 pixels and each image is rotated a certain angle and then tries to predict the number at the rotated angle. This project uses 10 tasks with 10 different rotation angles which creates 60,000 images for training and 10,000 images for testing. The model studies the tasks in order and each example once. Then the evaluation for each task is performed in the test partition of each dataset. On the MNIST tasks, we utilize completely connected neural networks with two (2) hidden layers of 100 ReLU units. All the networks are trained by plain SGD on mini batches of 10 samples.

Goal

The goal for this project is to use the MNIST dataset along with the GEM model to solve the challenge of catastrophic forgetting which occurs in Neural Networks. In the notebook with the code, we use a set of metrics to assess the model learning over a continuum of data. The metrics used identify the model by their accuracies and the capabilities of transferring information between tasks. The model used for continuous learning in this project is GEM which aims to diminish catastrophic forgetting while aso transferring information to previously learned tasks. More details on the metrics and framework can be found in the notebook along with the code.

Results





The figure above shows us the average accuracy (ACC), backward transfer (BWT) and forward transfer (FWT) for the dataset MNIST.

Final Accuracy = 0.928497

Backward Transfer = -0.046793

Forward Transfer = 0.757682

We can visualize that GEM minimizes backward transfer, while exhibiting a positive forward transfer.

References

- David Lopez-Paz and Marc' Aurelio Ranzato. Gradient Episodic Memory for Continual Learning.
- Continual Lifelong Learning with Neural Networks: A Review . German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, Stefan Wermter.