**Training Machine Learning Models**

Currently, the standard method to train machine learning models to classify images is to split a dataset into batches and repeatedly pass these batches to the model. These batches of data are then used to optimise the parameters of the model via backpropagation on a loss function which is computed using the batched data. This is known as Offline Training.

Offline Training is capable of producing high quality results. For example, Figure 1 shows the classification accuracy of a ResNet-18 model trained for 100 epochs on the CIFAR-10 dataset, it attains an average classification accuracy of over 77%.

However, Offline Training has some limitations. If the data is not available in its entirety, for example it is real-time data, or it is too expensive to store it all then Offline Training can be difficult as it requires the whole dataset to be available at training time.

This raises the question: what happens if the network is just trained as data becomes available?

**Catastrophic Forgetting**

Catastrophic Forgetting is a phenomenon that is caused by training an already trained model on new data. The model is unable to retain the knowledge that has learnt previously leading it to complete erase this learnt knowledge. Offline Training is unable to preserve the existing knowledge in the network.

This ability to learn continuously is often overlooked.

**Example of Catastrophic Forgetting**

In this example, a ResNet-18 model is trained on CIFAR-10 split into 5 separate tasks each consisting of 2 classes. These tasks are used to train the model sequentially, each task is trained for 20 epochs. Each time, the model forgets everything it has learnt about the previous tasks and overwrites the knowledge leading to a consistently low overall accuracy.

**Continual Learning**

Continual Learning is the study of techniques to reduce the effect of Catastrophic Forgetting. The majority of the literature focuses on alternatives to Offline Training but there has been a recent interest into how the architecture of a model may contribute to forgetting.

The key thing that links everything in this field is lack of need for the entire dataset to be present at training time. Elastic Weight Consolidation was one of the first techniques that gave prominence to the idea of Continual Learning, this is an approach that penalises changes to weights that are important to previous tasks.

**Applications of Continual Learning**

There are also real-world applications of this area of research. Retraining models on new data to improve their accuracy is costly in terms of energy usage, computational resources, and time. Continual Learning offers a potential way to reduce these costs and provide training benefits.

It is also important to note that Image Classification is the main domain used for research, but the techniques can also be applied to other domains such as Natural Language Processing.

**Setup for Continual Learning**

The setup of the Continual Learning is vitally important. This has been a key issue in the literature as highlighted by the authors of a technique known as GDumb. They are argue that many techniques in the literature have too simplistic setups to be applicable to the real-world.

There are a few key factors affecting the difficulty of the problem:

* Online vs Offline: Online assumes that the data is streamed continuously and not fixed unlike Offline Training
* Disjoint vs Overlapping: Disjoint task formulations partitions the problem into tasks with distinct classes without overlap. Non-disjoint is more real-world applicable as seen in Figures 2 and 3.
* Class Incremental vs Task Incremental: Task Incremental uses the disjoint task formulation but importantly the model is told which task a sample belongs to at classification time greatly simplifying the problem. Class Incremental can use either disjoint or overlapping and is more realistic
* Imposing limits on the resource consumption of the techniques is also important to ensure real-world applicability

**Part I: Literature Evaluation**

Initially, the focus of my project will be on implementing and comparing techniques from the literature. As mentioned before, the setup and comparison of techniques has been an issue within the literature.

This will start with implementation of the baselines: Offline Training and Finetuning, which is similar to the catastrophic forgetting example shown earlier and represents the lower bound. The goal is to match Offline Trainings results.

After this, I will implement historical techniques such as Elastic Weight Consolidation.

Following this, I will implement the current state of the art such as GDumb, a method that stores a subset of data and trains a classifier at inference time, Rainbow, another method that samples data that is representative of its own class and discriminative against others, and Mnemonics which is a similar sampling technique.

Finally, I will implement cutting edge techniques that are presenting new ideas such as Learning to Prompt, a technique that learns small inputs to be prepended to the input prior to classification, and meta-learning techniques that learn how to preserve knowledge during training.

**Part II: Novel Experimentation**

After this, I will focus on experimenting with gaps in the literature with the aim to contribute something novel to the field. This will be influenced by the cutting-edge techniques as well as potentially experimenting with the architecture of the model. This will involve explaining and understanding the theoretical underpinning, implementing these ideas and comparing them to existing techniques.