Training Machine Learning Models

* Talk about Offline Training and its requirements
* Talk about the graph
  + High quality results with minimal effort
  + No LR tuning etc
* Limitations
  + So much real time data these days
  + Can be too expensive to store and use all of it
* Neural networks inspired by brains but continual learning overlooked
* Raises the question, why not just use Offline Training with the incoming data?

Catastrophic Forgetting

* Offline training causes catastrophic forgetting
* No context that the existing weights of the network is pre-existing knowledge
* Unable to preserve the existing knowledge

Example of Catastrophic Forgetting

* An experiment showcasing Catastrophic Forgetting
* Trained the same network sequentially on CIFAR-10 split into 5 tasks
* Each task contains 2 classes, each task is completely disjoint
* Attains high classification accuracy on the classes in the task
* Overwrites the knowledge already acquired
* Leads to very poor overall classification accuracy

Continual Learning

* CL is the field that attempts to overcome these issues
* Aims to preserve prior knowledge while training sequentially
* Most of the literature focuses on alternatives to Offline Training
* Smaller focus on network architecture
* Setup of the continual learning problem is important – will cover this later on
* Concept that links all techniques is not requiring the whole dataset at the same time
* Rose to prominence with Elastic Weight Consolidation
  + Maybe a quick description of this?

Applications of Continual Learning

* Real-world benefits
* Reduction in energy usage, computational resources, and time
* Retraining models from scratch compounds cost
* CL offers chance to reduce these costs
* Real-time training benefits
* Image Classification primary focus but NLP also possibility

Setup for Continual Learning

* As mentioned previously
* Vital to ensure results are real-world applicable
* Issues were highlighted in the literature by GDumb
  + Simplistic because it does not take a necessarily specialised approach to the problem but still achieves highly competitive performance
* Multiple factors affecting the setup, difficulty determined by their combination
* Online vs Offline
  + Online assumes the data is streamed continuously and not fixed
  + Offline, like Offline Training, assumes access to the data at all times
  + Offline mostly defeats the point of Continual Learning and voids the benefits
  + Online is preferred
* Disjoint vs Non-disjoint
  + Disjoint task formulations partition the problem into classes with distinct classes in
  + No overlap between tasks
  + Requires knowing all of the classes in advance so they can be partitioned
  + Non-disjoint is more real-world applicable
  + Many techniques do use the disjoint task formulation and it can have applications
* Class-IL vs Task-IL
  + Task-IL use the disjoint formulation but at inference time the model is told which task the sample to be classified belongs to, drastically reduces the difficultly since it limits the classes
  + Class-IL, can use disjoint or non-disjoint tasks, is the preferred formulation, at classification time the model is given no additional information – just the sample to classify and that’s all
* Resource consumption
  + Some techniques store samples from the data stream (this doesn’t violate Online assumption)
  + Necessary to impose constraints on how much can be stored etc

Part I: Literature Evaluation

* Implementing and comparing techniques
* Idea is a robust evaluation
* Baselines:
  + Offline Training – as previously covered, provides an upper bound and the goal is to match this performance
  + Finetuning – Similar to the example of catastrophic forgetting shown before. Simply train the model (using traditional techniques) on the incoming samples and then discard them. Represents the lower bound as this will cause catastrophic forgetting. If we did worse than this then no point!
* Historical techniques:
  + Important techniques that are often referenced in the literature and provide CL specific baselines as new techniques should beat these
  + Elastic Weight Consolidation – mentioned previously. Falls into a type known as regularisation techniques that penalise weight changes to preserve knowledge. Fallen out of favour in recent literature
* State of the Art:
  + Current best approaches, can be difficult to truly identify due to the issues in the literature outlined previously
  + GDumb – mentioned previously, stores a balanced subset of the data and then trains a classifier at inference time on these samples. Shown good performance.
  + Rainbow
  + Mnemonics
* Cutting Edge:
  + Present new approaches and ideas, will feed into the second part of the project
  + LNIDA – Dynamically changes the structure of the network in response to the data
  + L2P -

Part II: Novel Experimentation

* Focus on gaps in the literature