RingTMP: Locally Distributed Machine Learning on Low Power Devices

Prifysgol Abertawe
Swansea University

Christopher Hopkins

Department of Computer Science
Swansea University

This dissertation is submitted for the degree of
Bachelor

April 22, 2021

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 40,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 100 figures.

Christopher Hopkins
October 2020

Abstract

More data than ever is being produced by low power devices such as smart phones and Internet of Things (IoT) devices at the network edge. The data being produced is so enormous it would be infeasible to send it to a centralised location. Instead models can be trained from data distributed across multiple edge nodes, with machine learning algorithms being performed locally. In this paper I explore training multiple low power devices using a new distributed machine learning paradigm *RingTMP*. This paradigm focuses on low power usage and power efficiency while having the capacity for larger models than comparative systems.

Contents

1 Introduction

1.1 Motivation and Context

Machine Learning has become an invisible but ubiquitous part of modern life, and is being used in a plethora of fields and industries. The uses of this technology range from dystopian facial recognition [1] to lifesaving diagnoses [2] and many more purposes besides. Machine Learning leverages existing data to train Machine Learning models in order to perform a task or find patterns, that previously only a human could. The key difference between Machine Learning and conventional programs is that the data itself is used to develop the model. Therefore the quality and quantity of the data can affect the effectiveness of a machine learning model.

As the amount and complexity of the data we are collecting increases, so does the size and complexity of the machine learning models we use to make sense of our data. For example the Internet Archive as of 2020 contains over 70 petabytes of data, while labeled datasets such as AViD have video data in the order of terabytes. [3] We are now reaching a point where the limiting factor of creating a machine learning model is not the data, but the machine learning algorithm itself.

This problem is two pronged. First machine learning models are getting very, very large. For instance GPT-3 the largest NLP model ever trained contains 175 billion parameters. [4] And efforts are being made to create models with trillions of parameters. [5] We have reached the point where its no longer possible to store some machine learning models on a single machine. [6] The second problem is that training a machine learning model is increasingly taking longer and longer. This is because we have more data and larger models, but the algorithms used to train models have fundamentally stayed the same and are inherently sequential and difficult to parallelise.

The popular current solution is to use a parameter server model. In brief the paradigm is made of two different types of components. The parameter server and the workers. The parameter server holds the global parameters of the model. Workers are given the model parameters by the parameter server. The workers then perform an iteration of whichever machine learning algorithm they are performing, modifying the parameters. Then the modified parameters are sent to the parameter server where they are aggregated, the global model parameters are updated and the cycle continues until the model has converged on an answer.

While this method has many benefits and is certainly faster than training using a single machine, it has two key limiting factors. First, every worker must communicate with a single parameter server, this limits scalability as eventually the network bandwidth becomes saturated severely impacting performance. [6] Secondly many parameter server models require the whole model to be replicated within each node. [7] This means that very large models simply cannot run on many machines.

In this paper I will outline an alternative distributed machine learning framework: *RingTMP*. RingTMP (Ring Topological Model Parallel) is a Ring Topological Model Parallel distributed machine learning framework focusing on optimising Distributed Gradient Descent. This is a novel design drawing in inspiration much research but particularly from the STRADS and DistBelief machine learning frameworks. [8, 9]

I believe my distributed framework may have some advantages over the current paradigm, these briefly are:

- There will be less communication between nodes
- A Potential for larger communication bandwidth between nodes
- Will be able to hold larger models
- Will be able to train neural networking models as faster or faster than a comparative parameter server, to the same level of accuracy

My aims more specifically for this project are to:

- Create a prototype RingTMP framework.
- Create a parameter server model framework.
- Demonstrate less communication between nodes
- Demonstrate that RingTMP is at least as scalable than a generic parameter server
- Demonstrate that RingTMP can hold larger models in comparison to a standard parameter server
- Demonstrate RingTMP can train neural networks to at least the same accuracy in at least the same amount of time

1.2 Overview

This document is split up into the following sections:

- **Introduction** Current section. Introduces the project and its aims.
- **Literature Review** Presents related research material in similar applications and areas.
- **Problem Description** Description of the problem aiming to be solved.
- **Solution Implementation** Details of the implementation of the solution.
- **Reflection** A small section reflecting on the project
- **Conclusion** Summary of the project and the paper.

2 Literature Review

2.1 Brief Introduction to Machine Learning and Neural Networks

To first understand Distributed Machine Learning you must first understand the fundamentals of machine learning and neural networks.

There are many machine learning methods some requiring training data which we call supervised and some being able to find patterns in data without being given solutions called unsupervised. [10] An example of a supervised system may be predicting house prices, using multiple factors about each house (market data, geographic area, square footage etc.) to come to a conclusion about what the house could sell for. This would be trained using data of previously sold houses to predict current ones. An example of an unsupervised system could be identification of new plant species. This could be done by taking as many features of a plant as possible, then apply a clustering algorithm to see if there are two distinct clusters in the data. If there were then that would suggest two different plant species. Neural networks tend to focus on supervised learning and use a form gradient descent called Stochastic Gradient Descent.

Many machine learning algorithms use a cost function to measure how well or badly they are solving a problem, these algorithms use parameters which are internal variables of a machine learning model and define how they solve the problem. If you map $costFunction(x)$, where x is the model parameter, for every x value. Then a graph will be produced $y = costFunction(x)$, the lowest point on the graph will be the global minimum. There may be other troughs higher than the global minimum these are called local minimums. A global minimum represents the lowest value of the cost function which indicates the parameter values produce the best solution for your problem. Initial model parameters are often randomised, which likely means they will start at a high point on the cost function graph, the goal is to get to the lowest point possible. To do this you must *descend* down the *gradient* to a local minima, the algorithm that does this is called gradient descent for that very reason. This often happens in little steps after the observation of each piece of data. However it is computationally expensive to step down the gradient after each example. It is more efficient to calculate the average step of a randomised selection of data. This is know as Stochastic Gradient Descent.

Neural networks are structures that can perform multi-variable gradient descent when

provided with training data. Neural networks are comprised of layers of interconnected neurons in a lattice like structure. Each neuron holds parameter information the adjusting of which through gradient descent leads to the solving of a problem through reaching the local minimum of the cost function. These structures can then be placed in a parameter server and run in a distributed fashion, if desired.

2.2 Limited History of Distributed Machine Learning

One of the first pieces of research into distributed machine learning was 'Distributed Inference for Latent Dirichlet Allocation' in 2008 [11] One of the first instances of distributed machine learning was used to categorise New York Times articles using Latent Dirichlet Allocation (LDA), which identifies the affiliations words have to certain topics. While the paper focused on parallelising the algorithm and running them over multiple artificially isolated cores the results showed that distributed machine learning could have scalability and didn't impact the rate of convergence of the model significantly. This was followed by a paper by Jia et al. [12] which produced much faster results than its predecessors by using memcache layer in every machine, every machine would message every other machine with updates of its local parameters to create an approximate global state, it was mentioned in passing that arranging the nodes in a star topology and caching the values that passed through it could make the system more scalable. After this followed a cambrian explosion of work in this area [9, 13–15] culminating in 2014 when the parameter server as it is known today [6] was produced. This parameter server is highly sophisticated and flexible, accommodating the difference in hardware components while spending more on computation and less time waiting.

2.3 Model and Data Parallelism

When creating distributed machine learning models there two different methods for distributing training, model parallelism and data parallelism. These two methods are not mutually exclusive and can be used in conjunction with one another, such as in DistBelief. [9]. Model parallelism is when model parameters are split between the nodes. As data parallelism is when the data is split between the nodes. [16] Often with model parallelism the whole set of training data is passed through each node. While in data parallelism its common for each node to hold the whole machine learning model.

The key advantage of model parallelism is that machine learning models can be far

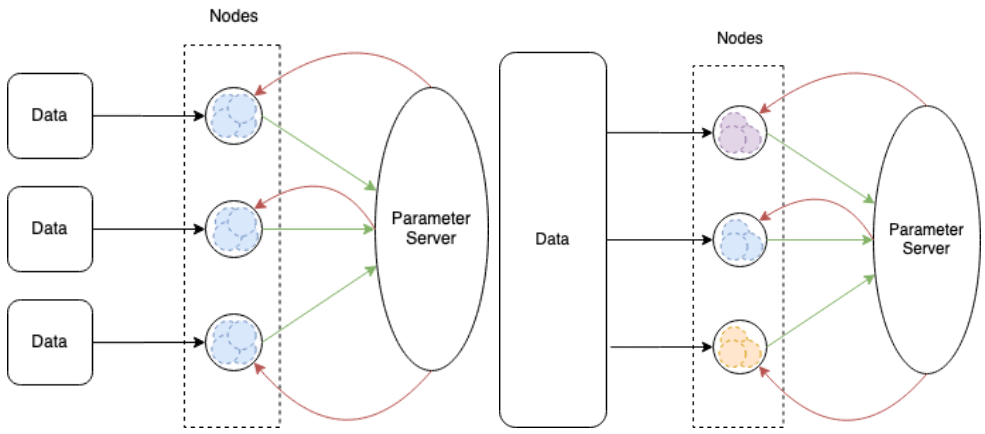

Figure 1: Left: Data Parallelism. Right: Model Parallelism. In both diagrams the green lines indicate local parameters being sent to the parameter server and red lines indicate parameters being sent to the worker nodes.

larger as they no longer have to sit on one machine. However this one great advantage comes with some disadvantages. Some parameters may take more time to converge than others, this means that at times some nodes may be idle while others are still converging, so the spread of computation is not equal or efficient. [9] Because some parameters converge at different rates a scheduler can be used, which does improve model convergence. However this requires more computational overhead and communication and reduces iteration throughput. [8]

Data parallelism has the benefit that data throughput can be very large, making processing using this method very fast. However with more nodes the communication overhead increases as the nodes must communicate the changes in their model parameters to each other. [17]

2.4 Model Consistency

Both of these parallel paradigms still function on the basis of a strictly iterative model, where communication is the limiting factor. Creating parameter server similar to how Googles map reduce algorithm is implemented [15] is still in some sense sequential. The next iteration can't continue until all workers have responded with their updated parameters, and each worker must send its results back no matter how significant its work is. Relaxing these restraints in specific ways have been shown to produce faster training times, while still converging on a model just as accurate. Especially when time, not data is at a premium. [14]

Partially relaxing the iterative restraint of the parameter server and instead using a bounded delay, dramatically decreases the wait times of workers to almost 0. A bounded delay allows workers to operate on slightly stale parameters. Unintuitive this allows for faster convergence than the sequential model, as it can iterate over the data faster and learn almost as much from each iteration. [14].

The other restraint that can be relaxed is the requirement to always send updated weights back to the parameter server. Many training examples don't dramatically change the parameters. Therefore only the training examples that cause the parameters to change significantly need to be sent back to the parameter server. This also allows for greater scalability, as each worker only communicates significant updates to the parameter server, meaning more nodes can be used before saturating the network. [14].

2.5 The Communication Issue

Distributed neural networks must communicate with each other in some way in order to work together. This needs to be formalised to be able to measure the efficiency of our machine learning system. Parts of this reasoning already appears in these papers too. [18, 19]

If we consider how a neural network operates if we were to run it on a single node, we could characterise its computation as such:

$$TIME = I_A(\epsilon) \times T_A \tag{1}$$

Where $I_A(\epsilon)$ is the number of iterations of the algorithm A it takes to reach accuracy ϵ and T_A is the time of each iteration of the algorithm. Here maximising the convergence per iteration or decreasing the time an iteration takes will decrease the runtime of the algorithm. In a distributed setting this equation changes to this:

$$TIME = I_A(\epsilon) \times (c + T_A) \tag{2}$$

In this equation we have the added variable c, this represents time taken for communication per iteration. In a distributed setting this will always remain above a non trivial amount of data. Unfortunately the majority of machine learning algorithms use a stochastic method which means a very large number of iterations ($I_A(\epsilon)$) that take a short amount of time (small T_A). You can see that no matter how small c is there will be a significant impact of the time taken. In fact with a naive approach of communication each iteration almost certainly $c > T_A$.

However this view doesn't take into account the possibility that communication could happen at the same time as an iteration. For example imagine a pipeline of nodes where each nodes performs an iteration but can communicate its previous iteration to the next node at the same time. Then the time taken could be described like so:

$$TIME = I_A(\epsilon) \times max(c, T_A) \tag{3}$$

Here you can see that if you can find a way of making the communication time equal to the time per iteration. Then c would have a negligible effect on the equation.

2.6 Low Power Hardware, IoT and Mobile Computation

Historically machine learning algorithms have been focused on high model accuracy through large models and vast amounts of training data, energy consumption and efficiency has rarely been taken into consideration. However with the rise of Internet of Things (IoT) devices and the established ubiquity of smart phones more data than ever is being produced. Soon this data generation with exceed the capacity of the internet, and experts estimate that over 90% of data will be stored and processed locally. [20] By extension this means machine learning algorithms will have to be performed locally too. This introduces some challenging issues. Modern machine learning algorithms require vast computational power and large amounts of data. Local devices don't have the capacity to hold large data sets or the power to compute large machine learning models in a viable amount of time, while many of them are also battery powered so power consumption becomes another issue.

A solution to this is to massively distribute the model over multiple decentralised nodes. The level of distribution is even greater than that of centralised compute clusters. In this solution each device computes a model using its own local data, infrequently (due to network constraints) the model is shared with a coordination server, which will then distribute the changes across all nodes in the network. [21] While this method is inefficient as the infrequent communications mean that many nodes may do much of the same work, and the merging of local models into a global one infrequently may cause loss of information. It still produces a model which converges in a relatively few rounds of communication. [18]

Efforts have been made in techniques to reduce the power memory and storage needed for machine learning algorithms to operate. One of these is Data Stream Mining. The idea is that a device can stream the analytics data directly into the model rather than

storing the data in storage for later use. [22] This means after the data has been read by the model it is lost. But that also means that no data needs to be stored, meaning resources are not spent reading and writing to storage. This has an application in mobile devices, as they produce data at a low rate through user interaction. Therefore the model can be build in real time as actions occur. The data produced on the mobile itself may not be enough to effectively train the model, but via communication with other users distributed over the network the model can converge. [18]

From the research available there seems to be research into distributed computing on local devices, investigation in to power measurements of machine learning algorithms [18,21] and power reduction of machine learning algorithms on local devices. [22] But there is a distinct lack of research into efficiency of *distributed algorithms* from the perspective of efficiency and power consumption on local devices. Having a more power efficient distributed machine learning algorithm, even if the optimisation was marginal on each device, would have an enormous effect on the output of the system, as many devices are connected.

3 Problem Description

In the most ideal world, communication between distributed network nodes would not to bottleneck performance. As many nodes as you liked could be added to your network and performance would scale linearly with each node added.

Currently when distributed neural networks become large enough, they become saturated. Adding more nodes no longer increases the speed of computation. To increase the speed of training further by adding more nodes at least one of two things must happen. Either the bandwidth between each node must be increased or the nodes must communicate less while still communicating enough information to continue training. [14] [23] This limit in the rate of computation means that training a neural network cost more time and money. We are at the point now where even if mistakes are found in the largest projects, its often 'infeasible' to retrain the model due to time and monetary cost. [4]

The impact of leaving this problem unsolved will mean the limiting factor of distributed neural networking will be not the processing power of the nodes but the bandwidth between them, bandwidth being one of the scarcest resources in data centres [24]. It will stifle the growth of machine learning models, especially for those that cannot get the funding for high bandwidth compute clusters.

This is why I designed a new paradigm for distributed neural networking that should reduce the communication between each node relative to a generic parameter server. The new paradigm also allows for higher bandwidth between adjacent nodes. As in this scenario nodes need only communicate with their adjacent partner. This results in a distributed learning model that isn't bound by bandwidth when scaling, and should be able to train models to the same level of accuracy in the same amount of time.

4 Implementation

4.1 Tooling

Implementing a distributed neural network is too large a task to be undertaken from scratch. Therefore its necessary to used existing tools, to make the development viable in the time given. This is difficult as not many languages lend themselves to both distributed systems and neural networks.

To ensure high performance the project could be implemented in C++. While C++ is often very performant and also has low level bindings for ML libraries such as TensorFlow. However even the creator of the language sees the need to improve its ability to improve its distributed performance. [25]

Python has great tooling for neural networking, such as TensorFlow [26], and PyTorch [27]. Moreover it has great support for numerical computing with NumPy [28]. These are performant too, by calling C functions or creating code which is optimised to run on GPUs to parallelise computation. However due to the Global Interpreter Lock (GIL) python is infamously bad at concurrency, while its distributed tooling is implemented in native python code, which lack of speed and could bottleneck the performance gained from using NumPy and TensorFlow.

Ultimately I decided to use Elixir as the programming language of implementation. This is because Elixir was designed for developing highly concurrent distributed systems. It does this by having a uniquely brilliant concurrency model. As opposed to OOP languages where 'everything is an object' in Elixir 'everything is a process'. This means the default way of writing the language enables it to be concurrent and scalable. Elixir also has the ability to communicate with other Elixir programs over the network using its own application protocol on top of TCP/IP. Meaning its as easy to communicate with local processes on your own machine as processes on another machine running an Elixir program. Its also been used by artificial intelligence researchers before as the process concurrency models effortlessly lends itself to modelling neurons. [29] Using Elixirs native float and arithmetic implementation would be slower than a C++ or a NumPy implementation, luckily there is a stable package which supports matrix calculations even faster than those in NumPy called Matrex. [30]

The only drawback of using Elixir is at the time of development it didn't have a strong machine learning library, which means implementing the mathematics of the neural

network myself. While this was a sizeable amount of work to do, it had the benefit that I didn't to wrestle with an opinionated API such as TensorFlow, I could create my own API to meet my ends.

4.2 Neural Network Implementation

In order to create a distributed neural network. I first needed to create a basic feed forward network that could operate on a single machine. This network doesn't need to be fully featured, its just a means to make an objective comparison between RingTMP and a generic parameter server. Therefore only 2 types of layers were implemented, the hidden layer and the output layer. The hidden layer is a generic dense layer similar to the kind you would find in any other neural network library. The output layer performs similarly to the hidden layer with the key difference that its always the final layer in a network and outputs the activations as probabilities. Within this section I will explain in more detail how the neural network was implemented from scratch and the mathematics behind its function.

4.2.1 Initialisation

Neural networks are composed of layers, while conceptually layers are composed of neurons, they're practically implemented with two components. A weights matrix and a bias vector. As I have already mentioned in this network there are two types of layers, hidden layers and output layers. We need to label each layer with its type, so we know how to perform forward and backpropagation. Therefore we can describe a layer as the tuple:

```
{layer_type, weights, bias}
```

Placing several of these tuples in a list creates a network:

```
network = [{:hidden_layer, weights_1, bias_1},
           {:hidden_layer, weights_2, bias_2},
           {:output_layer, weights_3, bias_3}]
```

The weights and biases have different dimensions depending on the layers input size and output size. A layer might take an activations vector with a size of m and output a size of n. The layer would hold a $n \times m$ matrix and the dimensions of the bias would be $n \times 1$. The output size of one layer must be the input size of the next layer. Initialising

the biases is simple, as biases have the function of an intercept in a linear equation, they can be initialised to 0. The trivial code is below:

```
defp initialise_bias(col) do
  Matrex.zeros(col, 1)
end
```

Weights are more complex. Each layer is initialised with random values from a uniform distribution, the shape of the uniform distribution is dependant upon the size of the input and output layers of the neural network. The type of initialisation used is dependant upon the activation function used.

In the hidden layer the ReLU function is used, common wisdom first established in this paper [31] states that for the fastest convergence He initialisation should be used. He initialisation is done by sampling random values from a normal distribution with a mean of 0 and a variance of $2/N$ where N represents the number of input values to a layer.

For the output layer, the softmax function is used. The best initialisation method in this case is using Xavier initialisation. [32] This also takes random samples from a normal distribution with a mean of 0 but has the variance of $1/N$ where N is $(inputSize + outputSize)/2$.

However in practice training the model often failed with He initialisations. This was because of what is known as the 'Dying ReLU Problem'. Which is when the elements of a z vector are negative, the ReLU activation function will return a zero meaning no learning is taking place. Once a neuron becomes dead its unlikely it will be revived as the function is piecewise and provides no slope for recovery such as a Leaky ReLU or a sigmoid function. To remedy this I trialed many distributions, settling on a mean of 0.5 with a variance of 0.25. While this is more simplistic, and may impact training times, its far more likely for the network to not be dead on arrival because of the initialisation parameters. This is part of the code which initialises the matrices in the layers:

```
defp random_val(_x, {n, :he, seed_state, acc}) do
  {val, new_state} = :rand.normal_s(0, 2 / col, seed_state)
  {col, :he, new_state, [val | acc]}
end

defp random_val(_x, {n, :pos, seed_state, acc}) do
  {val, new_state} = :rand.normal_s(0.5, 0.25, seed_state)
  {n, :pos, new_state, [val | acc]}
end
```

```
10
11  defp random_val(_x, {n, :xavier, seed_state, acc}) do
12    {val, new_state} = :rand.normal_s(0, 1 / col, seed_state)
13    {col, :xavier, new_state, [val | acc]}
14  end
```

You can find the wider context for this code snippet in code listing in the Appendix 1. 1

4.2.2 Forward Propagation

In forward propagation we give an input vector and an output vector is returned. In order for that to happen the input vector is passed through each layer, each time being transformed by the weights, bias and activation function of that layer. The input to the network could be measurements describing a flower (like in the setosa dataset), the output layer could describe which species. More broadly put, the input described the features, and the output predicts the categories to which those features belong.

On a layer level forward propagation is performed by taking the input vector multiplying it with the weights matrix, after which you add the bias and apply the activation function. The output of this is then passed to the next layer, or if its the last layer, output as the network prediction result. The generic mathematics of a forward function is described below in this linear equation where X is the input matrix, W is the weights, B is the bias, $activationFunc(x)$ is the activation function and finally A is the output activation:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad W = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \quad B = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{4}$$

$$A = activationFunc(W^T X + B)$$

In my implementation there are only two layer types. Hidden layers and output layers. The only difference between these two layers is the activation function that they use. Hidden layers use the ReLU activation function which is a simple piecewise non-linear function described as such:

$$relu(z) = max(0, z) \tag{5}$$

This function has become the de facto application function in dense hidden layers since its debut in 2011. [33]. In Elixir the forward action in the hidden layer is implemented

like so: [1]

```
defmodule HiddenLayer
 def forward(previous_activation, weights, bias) do
   weights |> Matrex.transpose()
   |> Matrex.dot(previous_activation)
   |> Matrex.add(bias)
   |> relu()
 end

 defp relu(z_vector) do
   z_vector|> Matrex.apply(
      fn value, _index -> if value > 0, do: value, else: 0 end
   )
 end

 ...
```

The output layer uses a more complex activation function, the softmax function, which transforms its inputs into probabilities. The sum of these probabilities is always 1. This is the softmax function, where z is a $i \times 1$ vector:

$$softmax(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{6}$$

This function is implemented the categorical output layer like so:

```
defmodule CategoricalOutputLayer do
 def forward(previous_activation, weights, bias) do
    weights |> Matrex.transpose()
    |> Matrex.dot(previous_activation)
    |> Matrex.add(bias)
    |> softmax()
 end

 defp softmax(z_vector) do
    stabilised_vec = Matrex.subtract(z_vector, Matrex.max(z_vector))
    exp = Matrex.apply(stabilised_vec, :exp)
    Matrex.divide(exp, Matrex.sum(exp))
 end

 ...
```

[1]The pipe operator `|>` transforms the function `val_a |> a_function(val_b)` into `a_function(val_a, val_b)`

Propagating forward through the layers happens like so, using the output of one layer as the input to the next, performing a slightly different forward action depending on the layer:

```
defmodule FeedForwardNetwork.Forward do
 def forward(network, input_vector) do
  Enum.reduce(network, input_vector, fn layer, acc ->
      forward_layer(layer, acc) end)
 end

 defp forward_layer({:hidden_layer, weights, bias}, prev_activation) do
  {_z_vec, activation} = HiddenLayer.forward(prev_activation, weights,
      bias, [])
  activation
 end

 defp forward_layer({:output_layer, weights, bias}, prev_activation) do
  CategoricalOutputLayer.forward(prev_activation, weights, bias, [])
 end

 ...
```

4.2.3 Cost Function

4.2.4 Back Propagation

References

[1] WIRED UK Matt Burgess. Some uk stores are using facial recognition to track shoppers.

[2] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey De Fauw, and Shravya Shetty. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, Jan 2020.

[3] AJ Piergiovanni and Michael S. Ryoo. Avid dataset: Anonymized videos from diverse countries, 2020.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[5] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models, 2020.

[6] Mu Li, David G. Andersen, Jun Woo Park, Alexander J. Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J. Shekita, and Bor-Yiing Su. Scaling distributed machine learning with the parameter server. In *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, pages 583–598, Broomfield, CO, October 2014. USENIX Association.

[7] Zhihao Jia, Matei Zaharia, and Alex Aiken. Beyond data and model parallelism for deep neural networks, 2018.

[8] Jin Kyu Kim, Qirong Ho, Seunghak Lee, Xun Zheng, Wei Dai, Garth A. Gibson, and Eric P. Xing. Strads: A distributed framework for scheduled model parallel machine

learning. In *Proceedings of the Eleventh European Conference on Computer Systems*, EuroSys '16, New York, NY, USA, 2016. Association for Computing Machinery.

[9] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Quoc V. Le, and Andrew Y. Ng. Large scale distributed deep networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1223–1231. Curran Associates, Inc., 2012.

[10] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[11] David Newman, Padhraic Smyth, Max Welling, and Arthur U Asuncion. Distributed inference for latent dirichlet allocation. In *Advances in neural information processing systems*, pages 1081–1088, 2008.

[12] Alexander Smola and Shravan Narayanamurthy. An architecture for parallel topic models. *Proc. VLDB Endow.*, 3(1–2):703–710, September 2010.

[13] Amr Ahmed, Moahmed Aly, Joseph Gonzalez, Shravan Narayanamurthy, and Alexander J. Smola. Scalable inference in latent variable models. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, page 123–132, New York, NY, USA, 2012. Association for Computing Machinery.

[14] Mu Li, David G Andersen, Alexander J Smola, and Kai Yu. Communication efficient distributed machine learning with the parameter server. In *Advances in Neural Information Processing Systems*, pages 19–27, 2014.

[15] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, January 2008.

[16] E. P. Xing, Q. Ho, W. Dai, J. K. Kim, J. Wei, S. Lee, X. Zheng, P. Xie, A. Kumar, and Y. Yu. Petuum: A new platform for distributed machine learning on big data. *IEEE Transactions on Big Data*, 1(2):49–67, 2015.

[17] Anis Elgabli, Jihong Park, Amrit S Bedi, Mehdi Bennis, and Vaneet Aggarwal. Gadmm: Fast and communication efficient framework for distributed machine learning. *Journal of Machine Learning Research*, 21(76):1–39, 2020.

[18] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

[19] Chenxin Ma, Jakub Konečný, Martin Jaggi, Virginia Smith, Michael I. Jordan, Peter Richtárik, and Martin Takáč. Distributed optimization with arbitrary local solvers. *Optimization Methods and Software*, 32(4):813–848, 2017.

[20] M. Chiang and T. Zhang. Fog and iot: An overview of research opportunities. *IEEE Internet of Things Journal*, 3(6):854–864, 2016.

[21] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan. When edge meets learning: Adaptive control for resource-constrained distributed machine learning. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 63–71, 2018.

[22] Eva García-Martín, Crefeda Faviola Rodrigues, Graham Riley, and Håkan Grahn. Estimation of energy consumption in machine learning. *Journal of Parallel and Distributed Computing*, 134:75 – 88, 2019.

[23] P. Sun, Y. Wen, T. N. Binh Duong, and S. Yan. Timed dataflow: Reducing communication overhead for distributed machine learning systems. In *2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS)*, pages 1110–1117, 2016.

[24] Luiz André Barroso, Jimmy Clidaras, and Urs Hölzle. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, Second Edition.* 2013.

[25] *Ask Me Anything with Bjarne Stroustrup, hosted by John Regehr.* YouTube, Jul 2020.

[26] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[28] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe,

Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with numpy. *Nature*, 585(7825):357–362, Sep 2020.

[29] Gene Sher. *Handbook of Neuroevolution Through Erlang*. 11 2012.

[30] Stas Versilov. Matrex. https://hexdocs.pm/matrex/Matrex.html.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[32] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[33] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

Appendices

A Appendix 1

A.0.1 Code Listings

Listing 1: Network Initialisation

```
defp initialise_parameters(definition, initial_seed) do
  definition
  |> Enum.reduce(
    {[], initial_seed},
    fn layer, {acc, seed_state} ->
      {layer, new_state} = initialise_layer(layer, seed_state)
      {acc ++ [layer], new_state}
    end
  )
end

defp initialise_layer({layer_type, input_size, output_size}, seed_state)
    do
  init_type = layer_to_init(layer_type)

  {initial_weights, seed_state} =
    initialise_weights(input_size, output_size, init_type, seed_state)

  initial_bias = initialise_bias(output_size)

  {{layer_type, initial_weights, initial_bias}, seed_state}
end

defp initialise_weights(col, row, init_type, seed_state) do
  {_, _, _, seed_state, list_of_lists} =
    Enum.reduce(
      0..(col - 1),
      {row, col, init_type, seed_state, []},
      &random_row/2
    )
  {Matrex.new(list_of_lists), seed_state}
end

defp random_row(_x, {row, col, init_type, seed_state, acc}) do
  {_, _, new_state, row_vals} =
    Enum.reduce(
      0..(row - 1),
      {col, init_type, seed_state, []},
```

```
38        &random_val/2
39      )
40    {row, col, init_type, new_state, [row_vals | acc]}
41  end
42
43  defp random_val(_x, {n, :he, seed_state, acc}) do
44    {val, new_state} = :rand.normal_s(0, 2 / n, seed_state)
45    {n, :he, new_state, [val | acc]}
46  end
47
48  defp random_val(_x, {col, :pos, seed_state, acc}) do
49    {val, new_state} = :rand.normal_s(0.5, 0.25, seed_state)
50    {col, :pos, new_state, [val | acc]}
51  end
52
53  defp random_val(_x, {n, :xavier, seed_state, acc}) do
54    {val, new_state} = :rand.normal_s(0, 1 / n, seed_state)
55    {n, :xavier, new_state, [val | acc]}
56  end
57
58  defp initialise_bias(col) do
59    Matrex.zeros(col, 1)
60  end
61
62  defp layer_to_init(:hidden_layer), do: :pos
63  defp layer_to_init(:output_layer), do: :xavier
```

B two