Project Report

On

Multi-objective Precision Optimization of Deep Neural Networks

Submitted by

Divyansh Choudhary
Yasasvi V Peruvemba

Computer Science and Engineering

2nd year

Under the Guidance of

Dr. Kapil Ahuja

Department of Computer Science and Engineering

Indian Institute of Technology Indore

Autumn 2019

Introduction

There are many obvious value propositions and use cases for the implementation of deep neural networks on resource constrained platforms. We aim to implement a method that strikes a middle path that efficiently allocates bit width at the level of layers, and take into consideration constraints on resources.

Prerequisites

We have used the MNIST dataset of images to train our self implemented neural network to obtain weights, in order to perform mathematical computations that will help us solve the problem of assigning bit width layer by layer.

The network consists of 5 layers, trained stochastically for 10 epochs, resulting in a final accuracy of ~ 96%

These weights and biases of the network were then stored, to perform the inject and extract algorithm.

Optimization Problem

ΔX_K = Error allowed in Layer X

$\sigma X_{K->L}$ = Standard Deviation of Outputs caused by error in Layer K

From [1],

We know that there exists a linear relation between the Error in Layer X and the Standard Deviation from unmanipulated network output.

For all K in the network,

$\Delta X_K = \lambda_K * \sigma X_{K->L} + \theta_K$

Hence, we use linear regression to decipher the values of lambda and theta pertaining to each layer.

$\sigma X_{K->L} \Rightarrow \sigma_Ł \sqrt{\xi_K}$, We replace these errors from a layer, by a distribution of overall error tolerance with the help of new variables ξ_K.

Now, we frame the optimization problem as given in [1],

$\Delta X_K = \lambda_K * \sigma_Ł \sqrt{\xi_K} + \theta_K$, For every layer, Number of bits allowed would by inverse logarithm of the error allowed for that layer.

Hence, the optimization problem finally presents itself as -

$\min \mathbf{F} = \sum \rho_K(-\log_2(\Delta X_K))$ subject to $\sum \xi_K = 1$

Determining Regression Coefficients

We began by amassing the weights and biases of the trained neural network. We then select 50 images for which we will create a dataset to apply regression to. We record the actual values of the outputs from the last layer of these 50 images. We iterate over every image 20 times, for K^{th} layer, whilst inducing an error of ΔX for each iteration, we then inject a uniform error from the range $(-\Delta X, \Delta X)$ to the input of the K^{th} layer. We then note down the standard deviation of the output from the last layer as σ_K. We have now generated the dataset for each layer, with 1000 entries each (50 images * 20 iterations per

image). Writing a simple Linear regression problem to solve this, we get the following results -

Layer No.	Lamda	Theta	Bits Assigned (After optimization)
1	0.612293131692 97480	-0.03371212188 158611	4
2	0.843734682196 4328	-0.04650270480 6702023	3
3	5.033100125970 339	-0.27769976431 01508	1
4	2.465840333955 5597	-0.13602926735 116322	2
5	3.103776965919 5202	-0.17124393186 657508	1

Implementation

Find the implementation of the Neural Network, the Dataset creation, the Linear Regression followed by the SQP octave implementation and BitAssignment in the GitHub repository below -

GitHub Repository

Solving the Optimization problem :

We optimized our objective function using the SQP ultility provided by octave. We took (0.30, 0.15, 0.18, 0.20, 0.17) (random values) as our initial guess for Ɛ and We considered a standard deviation of 0.30 as acceptable for the output. SQP took 11 iterations to minimize our objective function to a value of 13.219. The results of our optimization are below -

```
octave:255> function r = g (x)
  r = [x(1) + x(2) + x(3) + x(4) + x(5) - 1];
endfunction

function obj = fn(x)
        lam = [0.61229313169297480 0.8437346821964328 5.033100125970339 2.4658403339555597 3.1037769659195202];
        thet = [-0.03371212188158611 -0.046502704806702023 -0.2776997643101508 -0.13602926735116322 -0.17124393186657508];
        sigma = 0.32;

        obj=0;
        for i = 1:5
                obj = obj - log2(lam(i)*sigma*sqrt(x(i)) + thet(i));
        endfor
endfunction

x0 = [0.30; 0.15; 0.18; 0.20; 0.17];
[x, obj, info, iter, nf, lambda] = sqp (x0, @fn, @g, [], 0, +realmax)

x =

   0.19986
   0.19996
   0.20006
   0.20005
   0.20006

obj =  13.219
info =  104
iter =  11
nf =  24
lambda =

  -5.86733
   0.00000
   0.00000
   0.00000
   0.00000
   0.00000
   0.00000
   0.00000
   0.00000
   0.00000
   0.00000
```

As we can see, the minimizer for our problem is:

(0.19986, 0.19996, 0.20006, 0.20005, 0.20006)

When we use this minimization value to find the optimal number of bits, we get the following:

```
1  import numpy as np
2  lam = [0.61229313169297480, 0.8437346821964328, 5.033100125970339, 2.4658403339555597, 3.1037769659195202]
3  thet = [-0.03371212188158611, -0.046502704806702023, -0.2776997643101508, -0.13602926735116322, -0.17124393186657508]
4  sigma = 0.32
5  x = [0.19986, 0.19996, 0.20006, 0.20005, 0.20006]
6  for i in range(0,5):
7      print("Layer {} requires: {} bits".format(i, np.floor(-np.log2(lam[i]*sigma*np.sqrt(x[i]) + thet[i]))))
8

PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL

➜ ~ python code.py
Layer 0 requires: 4.0 bits
Layer 1 requires: 3.0 bits
Layer 2 requires: 1.0 bits
Layer 3 requires: 2.0 bits
Layer 4 requires: 1.0 bits
```

Here we can see that the max number of bits we require is 4 (for the first layer). To see how good this is, we must consider that the IEEE-754 format keeps 23 bits to represent the fractional part of number and we are getting a reasonable accuracy with only 4 bits. If we take this into account while designing hardware, we can end up saving a huge amount of space.

Conclusion :

This project has been instrumental in understanding Neural Networks and Error propagation through them. We also gained insight into the application of Optimization techniques (SQP) in problems of great scientific importance. We have optimized the number of fractional bits that need to be allocated to a neural network which would work on image classification (MNIST), which can be fruitful whilst designing minimal hardware solutions (Cost-effective) to such problems.

References

1. **https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8714785**
2. **https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7783722**
3. **https://arxiv.org/pdf/1511.06393.pdf**