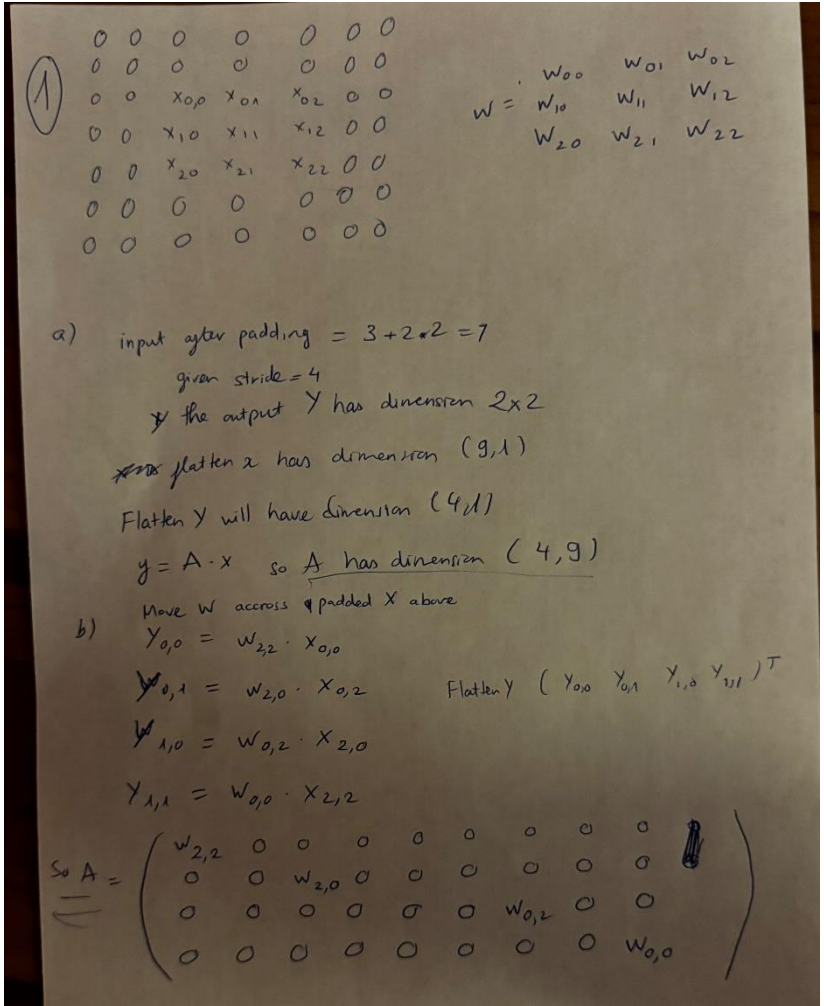
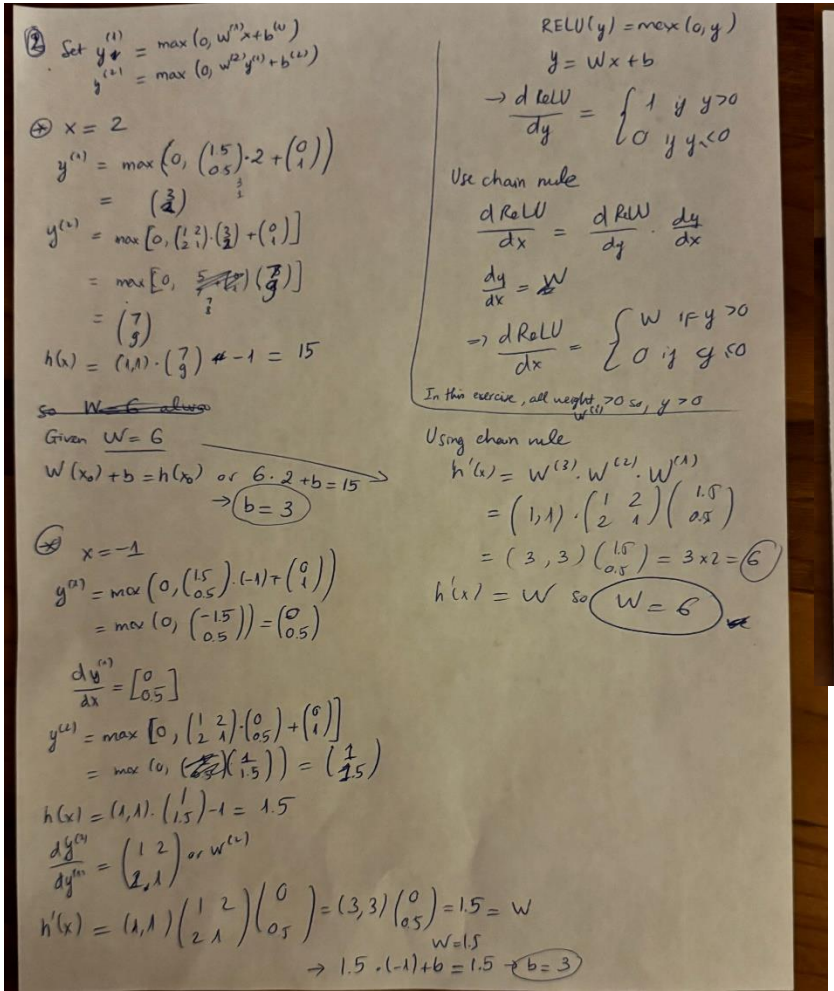
Assignment 2 Theory Problem Set

DO NOT TAG

Theory PS Q1. Must show your work for full credit. Feel free to add extra slides if needed.



Theory PS Q2. Must show your work for full credit. Feel free to add extra slides if needed.



Set $y^{(1)} = \max(0, W^{(1)}x + b^{(1)})$

$y^{(2)} = \max(0, W^{(2)}y^{(1)} + b^{(2)})$

⊛ $x = 2$

$y^{(1)} = \max\left(0, \begin{pmatrix}1.5\\0.5\end{pmatrix}\cdot 2 + \begin{pmatrix}0\\1\end{pmatrix}\right)$

$= \begin{pmatrix}3\\2\end{pmatrix}$

$y^{(2)} = \max\left[0, \begin{pmatrix}1&2\\2&1\end{pmatrix}\cdot\begin{pmatrix}3\\2\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix}\right]$

$= \max\left[0, \begin{pmatrix}7\\7\end{pmatrix}\right]$

$= \begin{pmatrix}7\\9\end{pmatrix}$

$h(x) = (1,1)\cdot\begin{pmatrix}7\\9\end{pmatrix} - 1 = 15$

~~So W = 6 also~~

Given $W = 6$

$W(x_0) + b = h(x_0)$ or $6 \cdot 2 + b = 15$

$\rightarrow b = 3$

⊛ $x = -1$

$y^{(1)} = \max\left(0, \begin{pmatrix}1.5\\0.5\end{pmatrix}\cdot(-1) + \begin{pmatrix}0\\1\end{pmatrix}\right)$

$= \max\left(0, \begin{pmatrix}-1.5\\0.5\end{pmatrix}\right) = \begin{pmatrix}0\\0.5\end{pmatrix}$

$\frac{dy^{(1)}}{dx} = \begin{bmatrix}0\\0.5\end{bmatrix}$

$y^{(2)} = \max\left[0, \begin{pmatrix}1&2\\2&1\end{pmatrix}\cdot\begin{pmatrix}0\\0.5\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix}\right]$

$= \max\left(0, \begin{pmatrix}1\\1.5\end{pmatrix}\right) = \begin{pmatrix}1\\1.5\end{pmatrix}$

$h(x) = (1,1)\cdot\begin{pmatrix}1\\1.5\end{pmatrix} - 1 = 1.5$

$\frac{dy^{(2)}}{dy^{(1)}} = \begin{pmatrix}1&2\\2&1\end{pmatrix}$ or $W^{(2)}$

$h'(x) = (1,1)\begin{pmatrix}1&2\\2&1\end{pmatrix}\begin{pmatrix}0\\0.5\end{pmatrix} = (3,3)\begin{pmatrix}0\\0.5\end{pmatrix} = 1.5 = W$

$W = 1.5$

$\rightarrow 1.5 \cdot (-1) + b = 1.5 \rightarrow b = 3$

$RELU(y) = \max(0, y)$

$y = Wx + b$

$\rightarrow \frac{d\,RELU}{dy} = \begin{cases}1 & \text{if } y > 0\\ 0 & \text{if } y < 0\end{cases}$

Use chain rule

$\frac{d\,RELU}{dx} = \frac{d\,RELU}{dy}\cdot\frac{dy}{dx}$

$\frac{dy}{dx} = W$

$\rightarrow \frac{d\,RELU}{dx} = \begin{cases}W & \text{if } y > 0\\ 0 & \text{if } y \le 0\end{cases}$

In this exercise, all weight $W^{(i)} > 0$ so, $y > 0$

Using chain rule

$h'(x) = W^{(3)}\cdot W^{(2)}\cdot W^{(1)}$

$= (1,1)\cdot\begin{pmatrix}1&2\\2&1\end{pmatrix}\begin{pmatrix}1.5\\0.5\end{pmatrix}$

$= (3,3)\begin{pmatrix}1.5\\0.5\end{pmatrix} = 3 \times 2 = 6$

$h'(x) = W$ so $W = 6$

2 - continue

⊛ $x = 1$

$y^{(1)} = \max\left(0, \begin{pmatrix}1.5\\0.5\end{pmatrix}(1) + \begin{pmatrix}0\\1\end{pmatrix}\right)$

$= \begin{pmatrix}1.5\\1.5\end{pmatrix}$

$y^{(2)} = \max\left(0, \begin{pmatrix}1&2\\2&1\end{pmatrix}\begin{pmatrix}1.5\\1.5\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix}\right)$

$= \max\left(0, \begin{pmatrix}4.5\\5.5\end{pmatrix}\right)$

$= \begin{pmatrix}4.5\\5.5\end{pmatrix}$

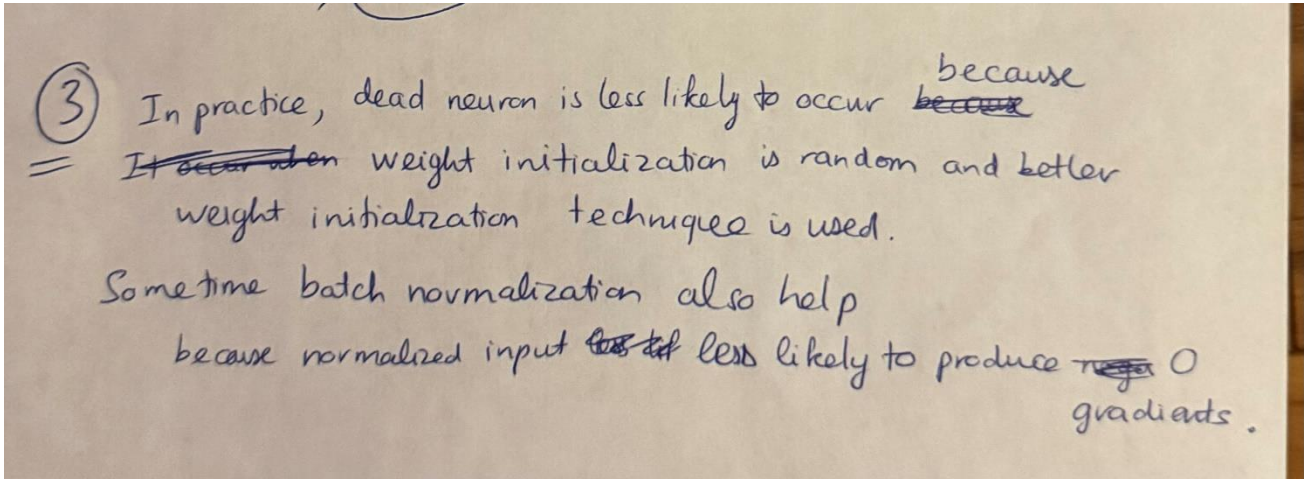
$h(x) = (1,1)\begin{pmatrix}4.5\\5.5\end{pmatrix} - 1 = 9$

$h'(x) = 6$ same as when $x_0 = 2$

$\rightarrow W = 6$, $6 \cdot (1) + b = 9$

$\rightarrow b = 3$

Theory PS Q3.


(3) In practice, dead neuron is less likely to occur because
= weight initialization is random and better
weight initialization techniques is used.
Sometime batch normalization also help
because normalized input less likely to produce 0
gradients.

Assignment 2 Paper Review

DO NOT TAG

Provide a short preview of the paper of your choice. (First paper – ImageNet Trained CNNs are bias toward texture…)

Convolutional Neural Network (CNN) trained on ImageNet tend to recognize objected based on texture rather than shapes. This contrast with human vision who recognize object based on shape more. This is a great paper that contribute to understand of CNN biases as well as provide practical solution to improve model robustness and accuracy.
The paper provide rigorous experiment to compare CNN performance with human perception using cue-conflict images.(“nine comprehensive and careful psychological experiments). It also provide experiments to conclude that CNN are lean toward texture due to the bias of data in ImageNet. By introducing a Stylized-ImageNet(SIN), a dataset that remove local cues through style transfer, the network are forced to train beyond texture recognition. In some case, ResNet-50 trained on SIN can reach or surpass human-level robustness. This highlighted the usefulness of shape-based representation data. Training CNNs on SIN enhances resistance to image distortion and improve reliability.
It raised very good point that CNNs maybe taking shortcut by focusing on textures, which is more common in ImageNet. By highlight the differences and potential it can be while training on shape-based image, we have potential to build more “plausible models of human visual object recognition”.

Still, the paper did not talk much about computational cost as well as if this approach generalize well outside ImageNet or SIN.

My personal take away from this paper is that CNNs and human visual perception still have different biases. It surprises me that ImageNet trained CNNs rely more on texture cues than shape which is different from what I thought it is. The result that show training CNNs on SIN reduces texture bias lead to more robustness and accuracy and can surpass human-level. I would love to see if this approach generalized well in other shape-based representation outside ImageNet, as well as a quantified research on computational cost of this approach.

Paper specific Q1. Feel free to add extra slides if needed.

Even if a neural network model can perform well on training dataset. It does not guarantee to perform well outside this dataset, in other word, might not generalize well on different dataset that is missing texture cues, distortion image. Therefore, understanding its biases are necessary to build a more robust model.
Should the network have same biases as humans, this is a debatable question. Do we want the model to see and perceive things as we are? If the purpose is interpretability or fairness, I think human-like bias can help, especially help align with our expectation. Help us see things as we do and understand things as we want.
Is human perception perfect? I do not think so, we have many biases as a human. So in different situation where human biases are not desired, having a model unbiased like human is better. It can help the model learn from human mistake of not making the same mistake due to biases.

Paper specific Q2. Feel free to add extra slides if needed.

Like the paper mentioned, the Stylized-ImageNet or SIN change the bias in the data, by removing texture cues. This forced the model training to focus more on shape. It helps the model recognize broader structural patterns which is more common in different dataset and more in line with human real-world condition.

A shape biased model is more reliable or robust toward image distortion or noise or corruption. This has been experimented in the paper. It also generalize better because shapes are more stable and resistant to corruptions. Shape biased model are more reliable and classify bad image or low quality image. The paper also showed result that SIN-trained models perform better than CNNs on ImageNet as well as are more reliable against noise, blur, corruption in image.

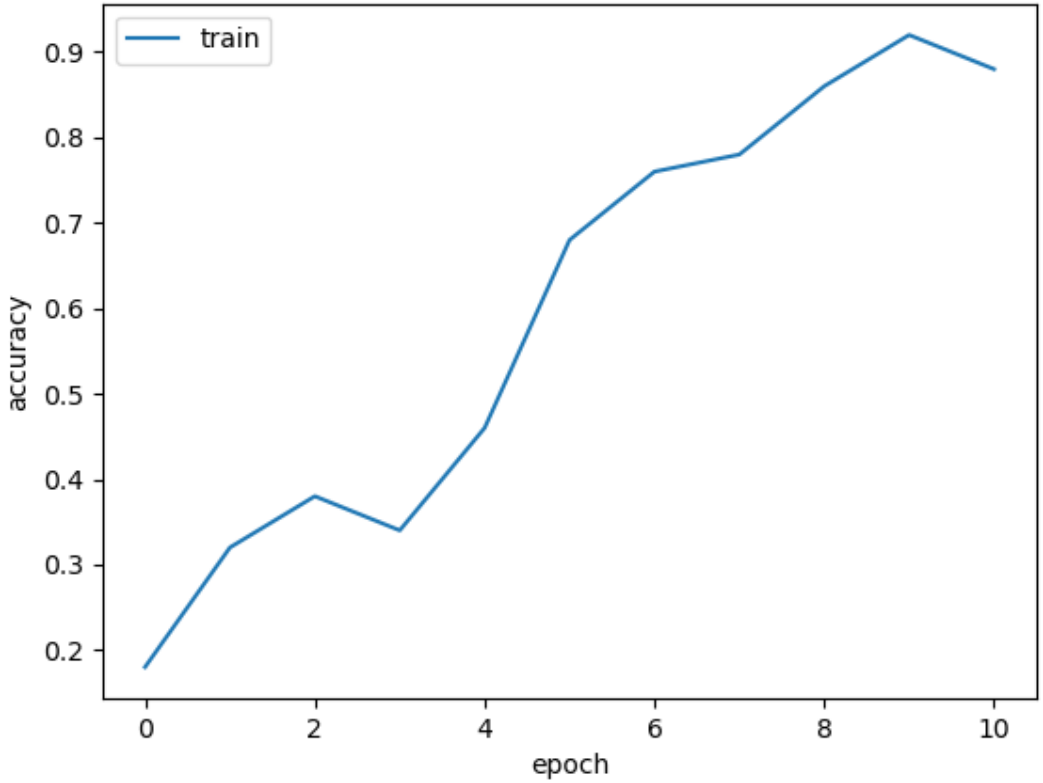
Assignment 2 Writeup

DO NOT TAG

Part-1 ConvNet

DO NOT TAG

Put your training curve here:



My CNN Model

DO NOT TAG

Describe and justify your model design in plain text here:

My models include 2 convolution blocks each block consists of 2 layers of convolution with MaxPooling. The purpose of convolutional is to extract feature, learning different spatial patterns. I start with 5x5 kernel to help capture broader spatial features and help train faster. Batch Norm is used in all convolution to improve stability and converge. Maxpool also help reduce computational cost, and keep more important information. I did increase the filters in subsequent convolutions to learn more complex patterns. The last is a fully connected layer to provide representation of learned feature that can convert to probability through softmax. I did have dropout feature to reduce overfitting.

Describe and justify your choice of hyper-parameters:

I did not have a lot of time and limited computational power so I keep epoch low but enough. Here I used 30.

Learning rate is one of the most important hyper parameters. I tuned and learn that around 0.005 is a great choice but due to low epoch. I need to increase the learning rate at the beginning. So above 0.1 is unstable, I choose to start at 0.01, and slowly decay them twice with decay rate 0.6. This help the the model learn fast from the beginning and slowly learn more detail and improving near the end of training.

Other choice for batch size, and momentum are pretty standard. I do not want to lower batch size because I want to train faster but the model require high accuracy above 0.8 so I choose a standard one 128.

What’s your final accuracy on validation set?

0.8209 (from 0.7 to 0.9 across 10 classes)

Data Wrangling

DO NOT TAG

What’s your result of training with regular CE loss on imbalanced CIFAR-10?

Tune appropriate parameters and fill in your best per-class accuracy in the table

	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
CE Loss Lr=0.005	0.891	0.858	0.612	0.579	0.150	0.078	0.072	0	0	0

What’s your result of training with CB-Focal loss on imbalanced CIFAR-10?

Additionally tune the hyper-parameter beta and fill in your per-class accuracy in the table; add more rows as needed

	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
beta=0.9999	0.629	0.584	0.258	0.216	0.431	0.456	0.497	0.476	0.391	0.452
beta=0.999	0.836	0.773	0.485	0.319	0.301	0.291	0.494	0.511	0.197	0.188
Beta = 0.99999	0.588	0.513	0.209	0.314	0.422	0.423	0.485	0.533	0.465	0.472

Put your results of CE loss and CB-Focal Loss(best) together:

	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
CE Loss	0.891	0.858	0.612	0.579	0.150	0.078	0.072	0	0	0
CB-Focal	0.629	0.584	0.258	0.216	0.431	0.456	0.497	0.476	0.391	0.452

Describe and explain your observation on the result: *Explanation should go into **WHY** things work the way they do in the context of Machine Learning theory/intuition, along with justification for your experimentation methodology. **DO NOT** just describe the results, you should explain the reasoning behind your choices and what behavior you expected. Also, be cognizant of the best way to mindful and show the results that best emphasizes your key observations. If you need more than one slide to answer the question, you are free to create new slides.*

First of all, the CE Loss perform wells on majority classes (0-2) but fails miserably on minority classes (6-9), because of imbalance data set, the model is biased toward majority classes. This help increase accuracy overall but not generalize well. So if 80% of the model is class 0, predicting class 0 will result in 80% chance correct but does not mean it can predict other class well. If using in a balance test set, it will fail.

CB Focal loss improves the minority class performance significantly, but lower the accuracy for majority classes. This is because it assigns higher loss weights to rare classes (7 to 9) . This forced the model to learn more from rare case and less from majority cases. That is why we see increase in accuracy of minority classes but decrease accuracy in majority classes.

Which one is better? I think CB focal loss is better in this case, because it can generalize well , when predicting other population that is balance it is guarantee to perform better while CE loss model can only predict certain classes (like from 0 to 3)

We also tune beta in CE focal loss. Lower beta results in smaller weight adjustment. Higher beta penalize majority classes and force model to learn more across all classes. We found the starting beta at 0.9999 have the best balance and trade off in this imbalanced CIFAR-10 dataset.