Simon Fraser University
Faculty of Statistics & Actuarial Science
Burnaby, British Columbia

A Gentle Introduction to Neural Differential Equations

Presented By:
Barinder Thind

Contents

1 Introduction

The ever-expanding umbrella that encompasses deep-learning methodologies welcomed another member earlier this year with the advent of Neural Ordinary Differential Equations (NeuralODEs) [3]. This approach expands on Residual Neural Networks [13] which circumvented the vanishing gradient problem [20] that traditional deep neural networks confronted with an increasing number of hidden layers. NeuralODEs transformed the above approach from a discrete to a continuous domain allowing for more efficient memory allocation, flexibility with respect to time-evaluation, and parameter efficiency.

In Section II, I introduce neural networks along with an example demonstrating their use. In Section III, I present residual neural networks and go into some *depth* as to why they were an improvement; an R example is provided. Then, in Section IV, I digress for the purposes of understanding NeuralODEs into a primer on differential equations. Lastly, Section V unifies together the previous sections by introducing neural differential equations as a gestalt of the aforementioned approaches. All code will be provided in the index and in a separate *.rmd* file.

2 Neural Networks

Neural networks have excelled at prediction problems over the last decade. In this section, I highlight the underlying methodology and present an R demonstration.

2.1 Methodology

2.1.1 Forward Pass

For simplicity sake, we will analyze a network with just a "singe hidden layer". Consider Figure 1 - the blue *"neurons"* contain values of the input data. For example, if the input was an image, then each neuron would hold a value corresponding to the gray scale of each pixel of the image. This is known as the *activation* value. The red neurons contained within the second row are referred to as neurons from the *hidden layer*. Each single layer is a transformation of a linear combination of each activation in the first layer. For example, z_1 would be defined as:

$$z_1 = \sigma(\alpha_{0_1} + \vec{\alpha}_1^T \vec{x}) \tag{1}$$

Where $\vec{\alpha}$ and α_{0_1} is the set of weights and biases[1] and $\sigma()$ is some *activation function* that transforms the resulting linear combination so that it can provide us with useful numbers (for example, we might want probabilities so we could use the sigmoid

[1]These are initialized randomly in this simple case

function[2] when that is the case). Note that the vector $\vec{x}$ corresponds to a single ”row” of our data set (or, a single image if that was the data type we were working with) and is therefore a p-dimensional vector where p is the number of covariates.

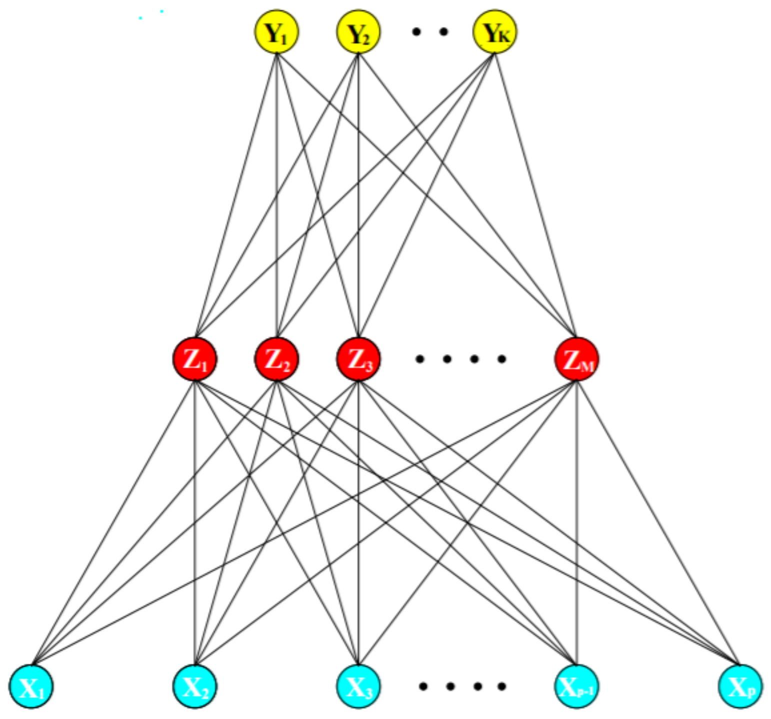

Figure 1: An overview of a single hidden layer network [19]

Once we have the m[3] neurons in the hidden layer, we have another set of activations! Using these, we can move onto the final layer. In the case of image classification, say for the purposes of number recognition, an image might correspond to a single number $y \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. In this case, $K = 10$ and is the number of possibilities for classification. The last transformation assigns a probability to each of the K classes effectively providing a likelihood that your observation (in our example, the single image) belongs to each of them, respectively. Often, the softmax function:

$$g_k(T) = \frac{e^{T_k}}{\sum_{i=1}^{K} e^{T_i}} \tag{2}$$

is used as it allows probability specification for a number of classes[4]. Here, there are k T values and these are just the linear combinations moving from the hidden layer to the output layer. You can imagine that this could be generalized to n number of hidden layers with some choice of m neurons in each of them resulting in a myriad of parameters to be estimated!

[2] $\frac{1}{1+e^{-x}}$
[3] The choice of m is left to the user
[4] In fact, letting $k \in \{0, 1\}$ returns the famous logistic transformation

2.1.2 Backpropogation

The "learning" of this approach takes place in a process called *backpropogation* - this is just jargon for gradient descent. In order to do so, we must first define a loss function. This function is some measure of the aggregated residual between our prediction and the true value. One approach is to use squared-error:

$$R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2 \tag{3}$$

This function is a sum of the difference between the prediction for every observation for every output. For example, the outer sum would be of the difference between the predicted probability of an image belonging to a particular class, over all classes and the inner sum would aggregate over every image (or observation) you have[5].

Now that we have a measure of error, we can look to minimize it! This function takes in $(p+1) \cdot M \cdot (M+1) \cdot K$ parameters[6] so we will have a very high dimensional gradient. This results in an inordinate amount of peaks and valleys on the optimization landscape. It is also very likely that the global minimizer of $R(\theta)$ will overfit the data so any local minimizer may serve us even better; in fact, we will take small steps towards the optimum specified as a "learning rate", γ.

For simplicity sake, we consider a network with a single neuron it its 2 hidden layers and only look at a 1-dimensional observation [1].

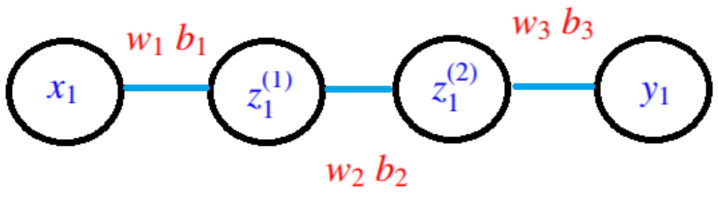

Figure 2: Schematic representing toy example

Then, the cost function, R will have 6 parameters outlined in red in Figure 2. Using (3), we see that in this simple case, the cost function reduces to $R_1 = (a^{(l)} - y)^2$ where $a^{(l)}$ is the activation in the final layer[7] defined as: $a^{(l)} = \sigma(w^{(l)}a^{(l-1)} + b^{(l)})$. For convenience, define $w^{(l)}a^{(l-1)} + b^{(l)}$ as $z^{(l)}$. Remember, we want to minimize the cost function; we can note that a change in the weight w^l cause some change to the cost function, R_1 - we want to know this change: $\frac{\delta R_1}{\delta}$ as our goal is to minimize

[5]There are a plethora of loss functions to pick from for example, cross-entropy and log-loss
[6]The set θ encompasses these parameters
[7]Let l be indicative of the last layer i.e. $w_3 = w^{(1)}$ and $w_2 = w^{(l-1)}$

this. Using chain rule, we can observe that this derivative can be broken down to a number of sub derivatives:

$$\frac{\delta R_1}{\delta w^{(l)}} = \frac{\delta z^{(l)}}{\delta w^{(l)}} \cdot \frac{\delta a^{(l)}}{\delta z^{(l)}} \cdot \frac{\delta R_1}{\delta a^{(l)}} \tag{4}$$

$$= a^{(l-1)} \cdot \sigma'(z^{(l)}) \cdot 2 \cdot (a^{(l)} - y) \tag{5}$$

We want to find the roots of this derivative but, not ONLY this derivative. First, we note that (3) is defined for every observation:

$$\frac{\delta R}{\delta w^{(l)}} = \frac{1}{n} \sum_{i=0}^{n-1} \frac{\delta R_i}{\delta w^{(l)}} \tag{6}$$

And note that (6) is only one of the 6 derivatives making up the gradient of the cost function:

$$\nabla R = \left(\frac{\delta R}{\delta w^{(l)}} \quad \frac{\delta R}{\delta w^{(l-1)}} \quad \frac{\delta R}{\delta w^{(l-2)}} \quad \frac{\delta R}{\delta b^{(l)}} \quad \frac{\delta R}{\delta b^{(l-1)}} \quad \frac{\delta R}{\delta b^{(l-2)}} \right)^T = \vec{0}$$

The parameter values that satisfy the above equation are the changes we need to make to the current weights. The change is done proportional to the aforementioned learning rate. This approach is taken for computational efficiency - finding the full gradients is nearly impossible so the optimal values are found in *mini batches*[8]; these are subsets of observations for which the optimization takes place as opposed to the entire data set. The process repeats for some number of *epochs*[9].

In summary, you begin with a set of weights, train the model, and get predictions. You run these predictions through a loss function and attempt to minimize it by updating the parameters of the function according to the gradient. You do this at some learning rate and the evaluations are done on subsets of data. And, you do this for some number of epochs.

2.2 Results

2.2.1 Data Description

The data set is taken from a 2017 Kaggle competition [18] in which participants were asked to classify satellite images as either icebergs or ships. There are two variables corresponding to the pixel values of the images (x, y coordinates) and a unique *ID* variable which corresponds to the i index in the theory above. There's a final binary

[8]Definitions for some of this jargon are provided in the Appendix

[9]A single epoch is one full forward and backward pass

(output) variable which classifies the image as an iceberg (or not). In Figure X1, some of the images are visualized.

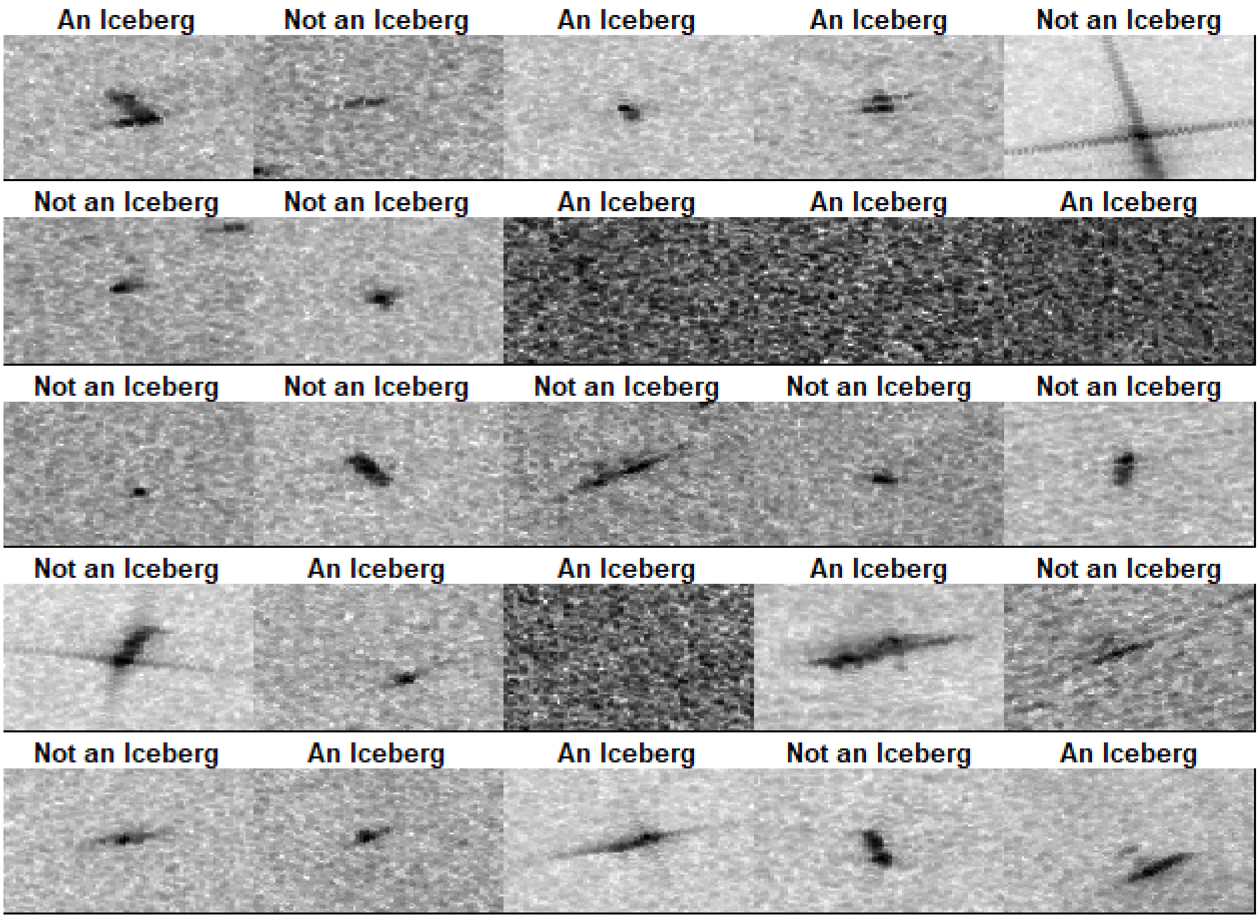

Figure 3: A snapshot of the grayscale iceberg/not iceberg images

2.2.2 Model Specifics

The model was trained on 1300 images and used to predict 304 [2]. The activation function used in the 4 *dense* hidden layers was *relu*[10] and the sigmoid function was used in the output. The optimization landscape was explored by *stochastic gradient descent* (sgd) and loss was characterized by *binary crossentropy*.

2.2.3 Performance

The model performed with exceptional mediocrity after being run through 150 epochs. Figure X2 provides loss and accuracy results for the model as it worked its way through the epochs. The final accuracy on the test images was: 54%.

[10] $\sigma(z) = max\{0, z\}$

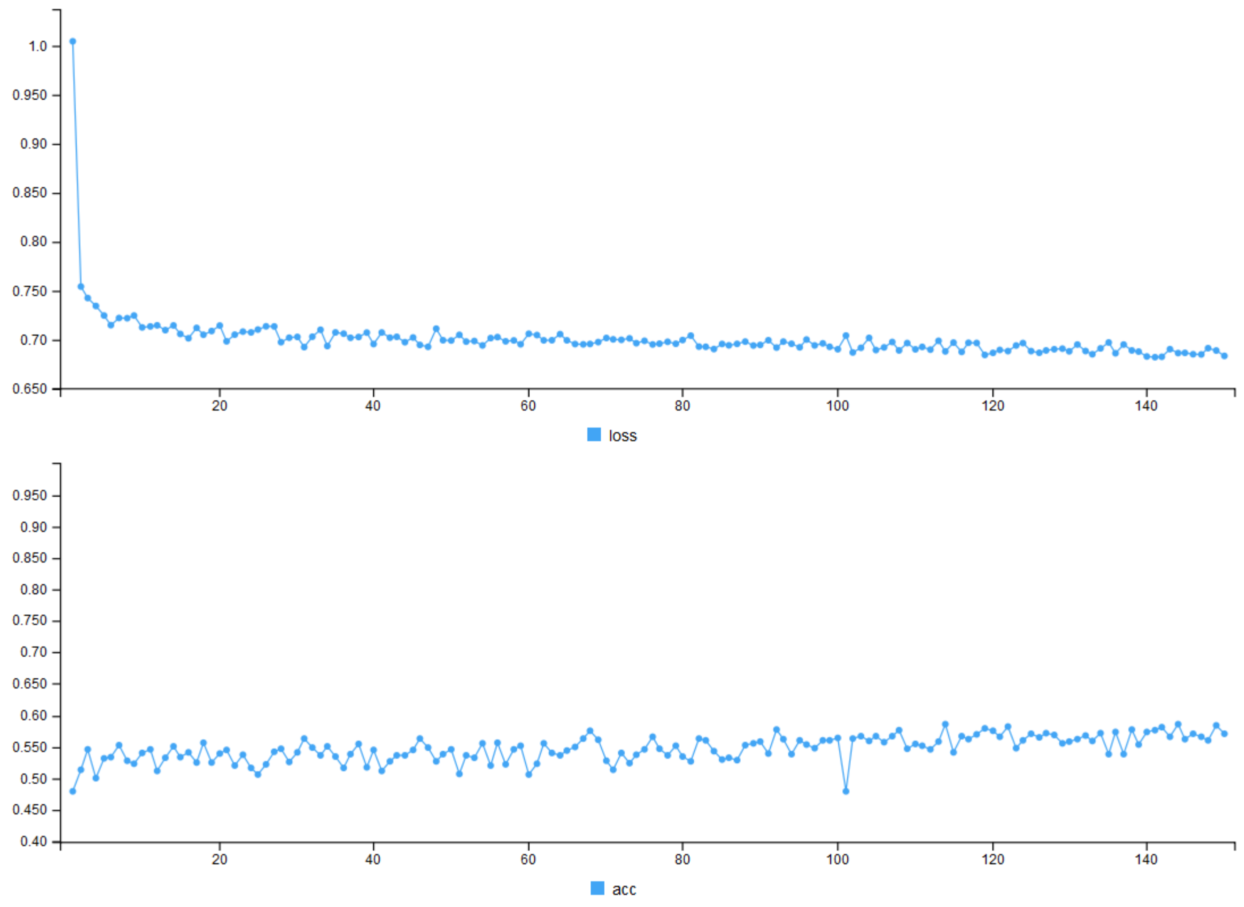

Figure 4: Loss and accuracy results for neural network model

3 Residual Neural Networks

In this section, I introduce an extension to deep neural networks developed by researchers at Microsoft [13].

3.1 Methodology

3.1.1 Residual Blocks

A common problem with plain neural networks is their inability to be trained on a large number of hidden layers. This problem arises due to vanishing (and exploding) gradients. A vanishing gradient occurs for weights and biases earlier in the network. Recall that, during back propogation, we use chain rule to find gradient values and that, the further back we are, the more terms there are that are used to compute the gradient. Since there are more terms, their exists a higher probability that some of those terms will be small and hence, there becomes a tendency for those earlier weights to hardly even move during the update portion of the iteration[11] [5].

[11]The update is: $w_i = w_{i-1} - \gamma \cdot \frac{\delta R}{\delta w_{i-1}}$. That is to say, the second term in this equation can become very small

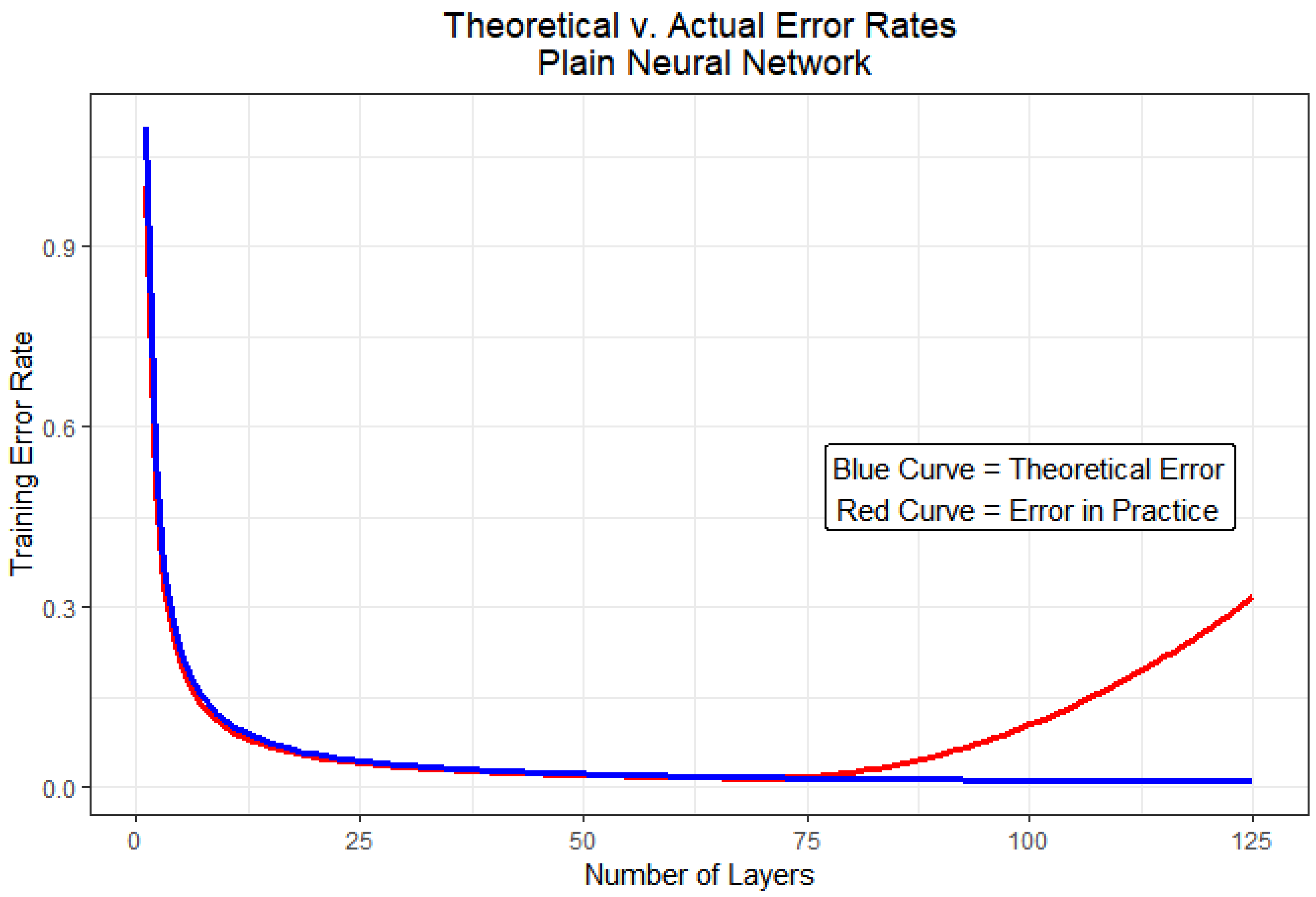

Figure 5: An overview of the training set error rates for recurrent (plain) neural networks. The vanishing gradient problem is theorized to be responsible for the blue curve

A solution to this problem comes in the form of *residual blocks*. These are modifications to the linear part of the nerual network in between layers. Consider an activation in classic nerual networks:

$$z_1 = \sigma(\alpha_{0_1} + \vec{\alpha}_1^T \vec{x}) \tag{7}$$

Letting z_2 and z_3 be the activations in the second and third layer[12]. Then, normally, we would have the following:

$$\begin{aligned} z_1 &= \sigma(\alpha_{0_1} + \vec{\alpha}_1^T \vec{x}) && (8)\\ z_2 &= \sigma(\alpha_{0_2} + \alpha_2 z_1) && (9)\\ z_3 &= \sigma(\alpha_{0_3} + \alpha_3 z_2) && (10) \end{aligned}$$

However, in a residual block we ad just say, z_3 so that we get:

$$z_3 = \sigma(\alpha_{0_3} + \alpha_3 z_2 + z_1) \tag{11}$$

[12]These are single dimensional i.e. only a single neuron in each layer. This generalizes easily an m-dimensional case where these would be vectors instead

The key insight here is that as the weight α_3 and the bias α_{0_3} vanish, the input into the activation function tends toward the identity transformation rather than 0. This means that, instead of having a degradation in learning as we increase the number of layers, the neural network will instead have, at *worst*, an identity transformation layer to layer (that is, the activation function will just take you back to the activation value of the $((i-2)+1)^{th}$ layer and allow the optimization to flourish in other elements of the gradient that are not (yet) experiencing the problem.

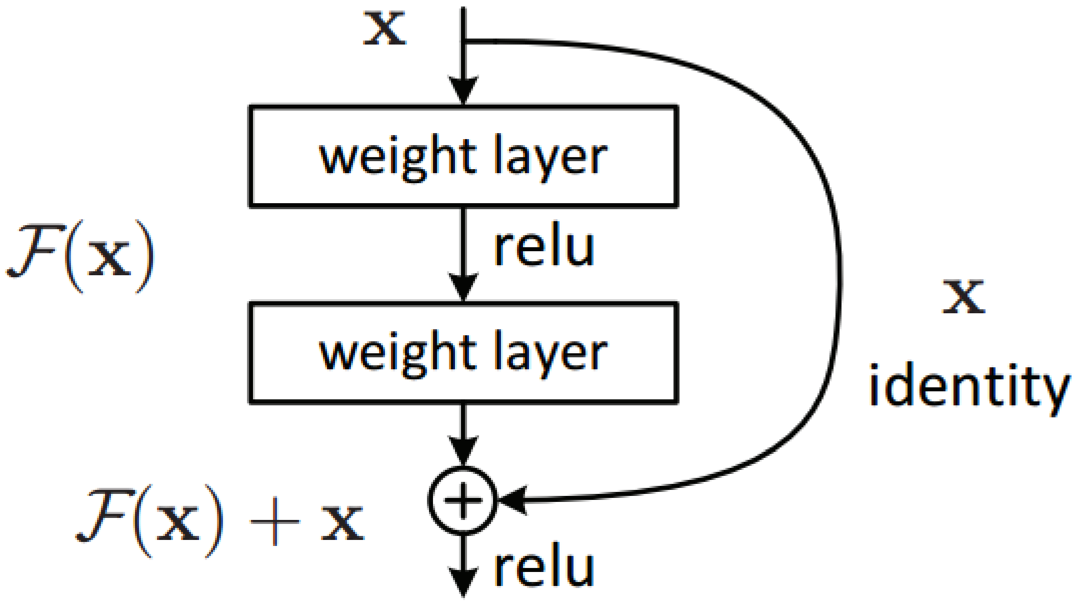
x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 6: A residual block. The $F(x)$ here is analagous to the z_3 in the notation used here. [10]

The algorithm for backpropagation remains the same. The additional derivative is computed with respect to the added term but the only process follows the same logic.

3.2 Results

The titanic data set was used once again here for the implementation. The residual blocks were used in conjunction with a convolutional neural network[13] (as opposed to be added onto regular neural networks) [8]. The relevant code is found in section 7.2.2 of the Appendix.

The model used the same number of epochs as the previous neural network and was trained on the same number of images (1300). Batch normalization was applied along with a number of other sub-layers relevant to a convolutional neural network[14]. In Figure 21312, we can see the relative superiority of this approach:
The prediction accuracy for this model on the same set of images was over 98% using mean squared error as the measure.

[13]This choice was made due to the nature of the data
[14]Definitions are provided in Section 7.2.1

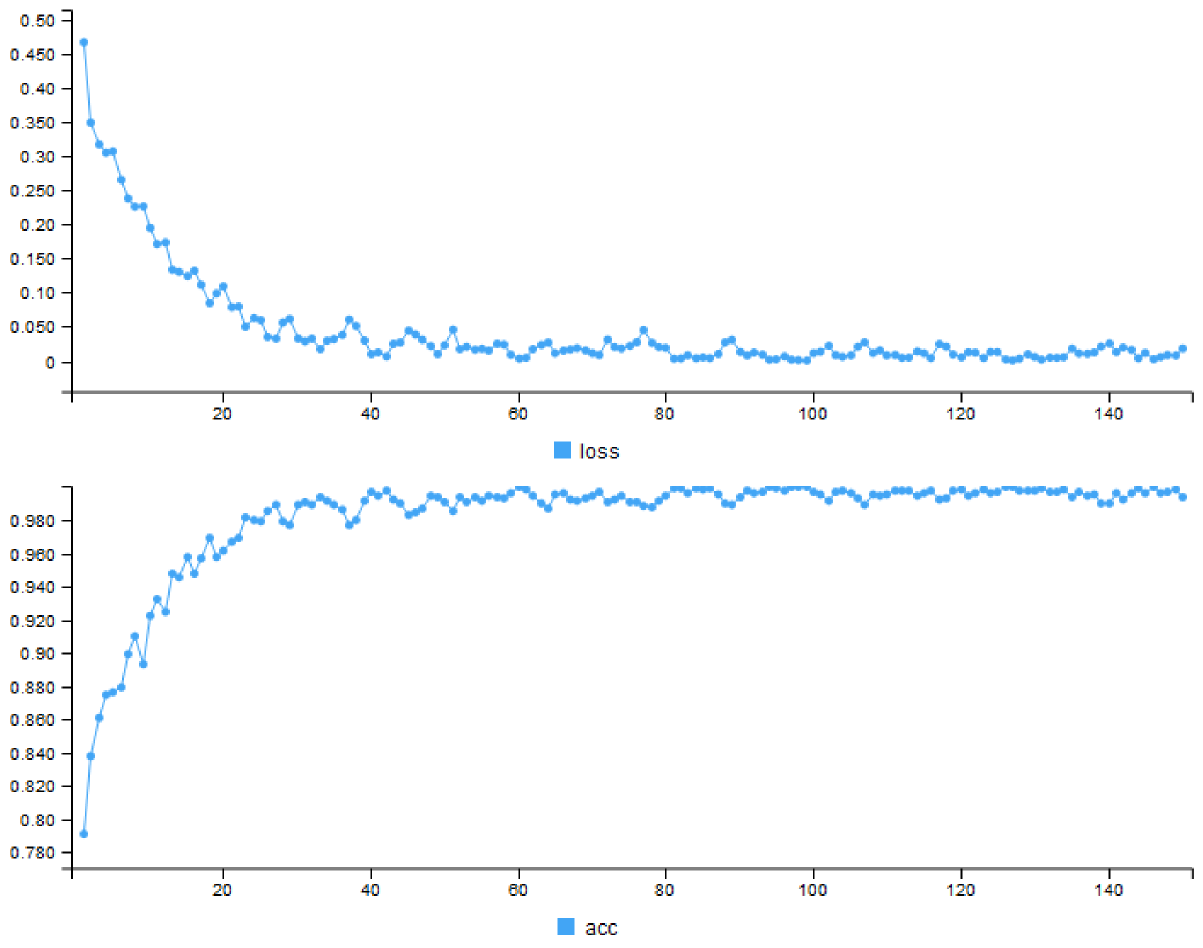

Figure 7: Accuracy and loss results for the CNN with residual blocks.

4 Neural Differential Equations

4.1 Methodology

4.1.1 An Overview

The main idea underlying the use of differential equations is that they require a fewer number of parameters which contributes to efficiency. To see why, consider a simple linear regression problem where the goal is to estimate optimal values of a and b for $f(x) = ax + b$. Observe that we make an implicit assumption here - the function $f(x)$ is differentiable and so, we can find $f(x)$ directly or we can estimate its derivative, $f'(x)$. The derivative of $f(x)$ is a; in the differential equation approach, we only have one parameter to estimate! And, in fact, differential-equation solver approaches don't provide analytic forms of $f(x)$ but rather, numerical values that are dependent on the initial inputs (the data).

Remember that a neural network, more than anything else, is a high dimensional function, $f(\vec(x); \theta)$ where θ is the set of weights and biases. Instead of estimating this function, we can model the derivative instead - i.e. the change in the function from layer to layer. Consider some vector [11] of hidden activations[15]:

$$h_{t+1} = f(h_t, \theta_t) \tag{12}$$

More importantly, inn the case of residual networks, the functional form becomes:

$$h_{t+1} = h_t + f(h_t, \theta_t) \tag{13}$$

An important insight is the striking resemblance of (xxw) to Euler's method[16] and, recall that Euler's method is a discretization of a continuous relationship between x and y (inputs and outputs). A neural network then, similarly, is also a discretization characterized by the hidden layers. ResNets, while discrete, effectively work as ODE solvers by measuring the rate of differnce in their hidden layers. Let t, the depth, go to infinite - then the entire set of layers of a neural network can be written as a differential equation:

$$\frac{dh}{dt} = f(h(t), t; \theta) \tag{14}$$

Intuitively, we have taken a step back in the ODE solving process to where we now have an option on which direction to go to solve the problem. In ResNets, Euler's method is the specified direction however, we aren't limited to that approach here and could use more sophisticated and efficient estimators. The authors use a "black-box differential equation solver".

The trajectory of Euler's method attempts to model the dynamic of the output over the continuum, x; analogously, the hidden layers in a neural network represent the dynamics of the hidden activations with respect to the depth of the network. The limit allows us to smooth out this trajectory so that we can evaluate a hidden activation at any depth $\in \mathbf{R}$. Note that the differential equation trajectories will differ depending on the weights and biases (think of these as initial conditions). In Figure 23213, I present one such trajectory[17].

The hidden state is evaluated by the following integral:

$$h(t) = \int f(t, h(t), \theta_t) dt \tag{15}$$

Where θ_t is the set of parameters at some layer, t. Lastly, note that the initial conditions (that is, at t_0) are given by the observations, $\vec{x}$ with the output being evaluated

[15]Moving forward, we will consider the depth (or the hidden layer we are at by t

[16]The appendix provides more details

[17]It's the plot on the right hand side

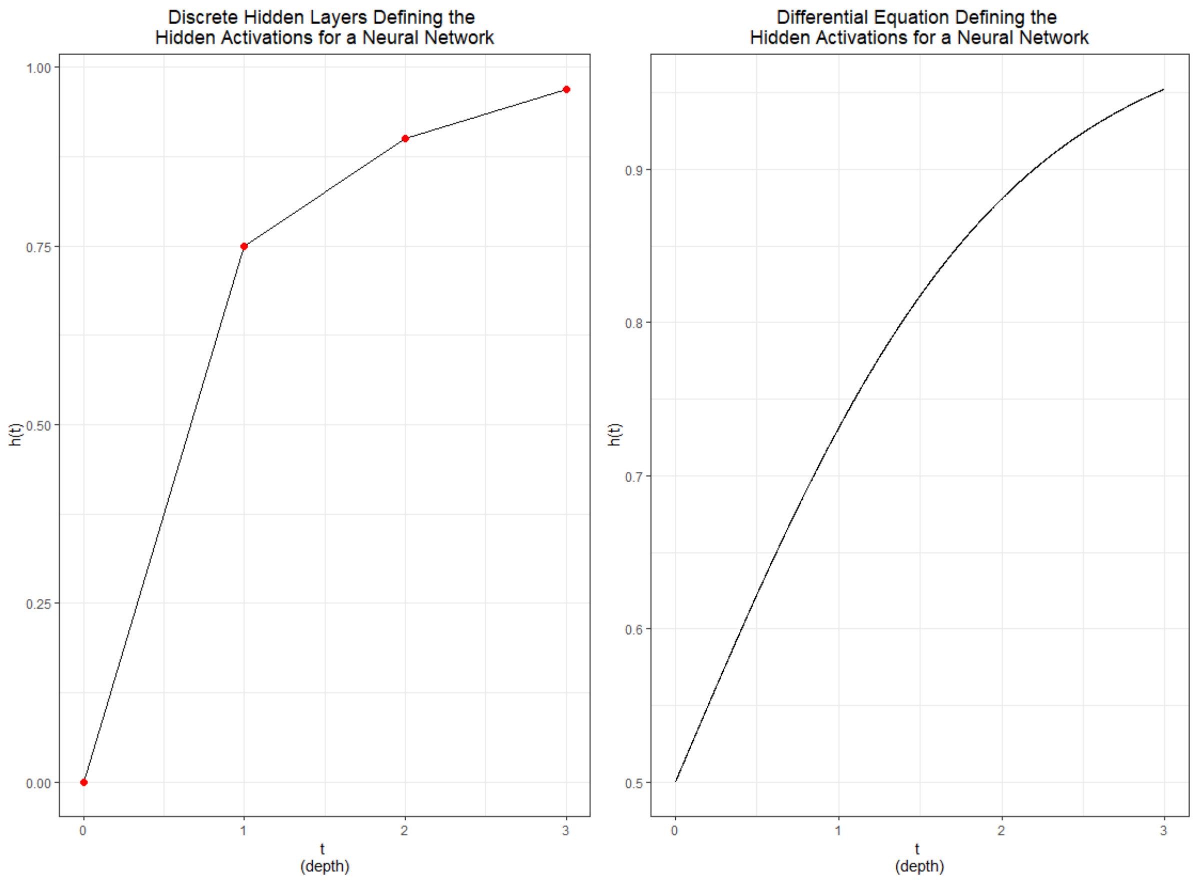

Figure 8: Trajectory comparisons of the two hidden state approaches. Note that the red dots in the left plot are the only evaluations we can do with classic neural networks whereas the dynamics are modeled at any depth in the NeuralODE approach

at some t_j where j is the + 1 iteration from the last hidden state. Deciding on t_0 and t_j is a problem left best to the optimization process; therefore, the final predictions can be summarized as [16]:

$$\hat{y} = h(t_j) = ODESolve(h(t_0, t_0, t_1, \theta, f)) \tag{16}$$

4.1.2 Backpropagation

Now that we have a functional form of the hidden states, we can begin to formulate the backpropagation process. As before, we begin with a loss function:

$$R(t_0, t_1, \theta_t) = R(ODESolve(h(t_0, t_0, t_1, \theta, f))) \tag{17}$$

Beginning with the hidden states, we can compute the gradient: $\frac{\delta R}{\delta h(t)}$. We implement the chain rule here because the hidden states themselves are dependent on t - essentially, we are working backwards along the path taken to get to the output, $h(t_j)$. In the paper, they use the adjoint method [12]. This is a numerical technique used to compute derivatives. An adjoint state is defined as:

$$a(t) = -\frac{\delta R}{\delta h(t)} \tag{18}$$

This is the change in the loss at any point t in the hidden state interval. Note that the loss function and the neural network are differentiable. We observe then that:

$$\frac{\delta a(t)}{\delta t} = -a(t)\frac{\delta f(t, h(t), \theta_t)}{\delta h(t)} \tag{19}$$

Which we note is also a differential equation. Using the Fundamental Theorem of Calculus, we can integrate both sides to find a solution for $a(t)$ and, recalling (xxx), we derive:

$$\frac{\delta R}{\delta h(t)} = -a(T) = \int a(t)^T \frac{\delta f(t, h(t), \theta_t)}{\delta h(t)} dt \tag{20}$$

And finally, we can solve this integral with the black-box ODE solve that was alluded to earlier. Computing this integral from $t1$ to t_0[18], we can get the gradient at t_0. Lastly, the θ gradient is computed by:

$$\frac{\delta R}{\delta h(t)} = \int_{t_1}^{t_0} a(t)^T \frac{\delta f(t, h(t), \theta_t)}{\delta \theta} dt \tag{21}$$

All of these derivatives can be computed simultaneously as the results do not depend on one another; this parallelization leads to computational efficiency.

4.2 Results

5 Conclusions & Future Considerations

In this report, I worked through a number of different machine learning techniques that have been significant with respect to machine learning. The recurrent neural networks were revolutionary in their ability to model non-linear relationships but suffered from problems in the realm of numerical analysis. The residual neural networks provided a reasonable solution to the vanishing gradient problem and allowed the training of over 150 layers without detrimental effects.

Neural ordinary differential equations recognized the similarity between the ResNet algorithm and Euler's method and took a step back in terms of the algorithmic process; they allowed the training of an infinite number of hidden layers and the flexibility of modelling using a differential equation. That is, there was great parameter

[18]Remember, this is a reverse traversal of the hidden states

and time efficiency that was otherwise not available.

It seems that the examples given in the paper were limited to an equal number of dimensions between layers - this can be expanded upon. A different dimensionality may contribute to the need of more sophisticated models that are defined for some different numbers of neurons, layer to layer. This is an open area of research. Expansions could also be made to the realm of functional data analysis where the inputs of the neural network would be sets of functions rather than scalar values. All in all, the approach seems to show promise as a new and useful family of the neural network world as is evident by the results provided in the paper. It remains to be seen the expansions that could be made on such a topic!

6 References

[1] 3Blue1Brown. *Backpropagation calculus — Deep learning, chapter 4*. Nov. 2017. URL: https://www.youtube.com/watch?v=tIeHLnjs5U8&index=4&list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi.

[2] J.J. Allaire. *TensorFlow for R*. URL: https://tensorflow.rstudio.com/keras/articles/tutorial_basic_classification.html.

[3] Tian Qi Chen et al. «Neural Ordinary Differential Equations». In: *CoRR* abs/1806.07366 (2018). arXiv: 1806.07366. URL: http://arxiv.org/abs/1806.07366.

[4] colesbury. *Batch Normalization Momentum? Issue #695 torchnn*. URL: https://github.com/torch/nn/issues/695.

[5] deeplizard. *Vanishing and Exploding Gradient explained*. Mar. 2018. URL: https://www.youtube.com/watch?v=qO_NLVjD6zE.

[6] Firdaouss Doukkali and Firdaouss Doukkali. *Batch normalization in Neural Networks*. Oct. 2017. URL: https://towardsdatascience.com/batch-normalization-in-neural-networks-1ac91516821c.

[7] William E Boyce and Richard C DiPrima. «Elementary differential equations and boundary value problems / William E. Boyce, Richard C. DiPrima». In: *SERBIULA (sistema Librum 2.0)* (Mar. 2019).

[8] Dimitri F. *keras with data augmentation (LB: 0.1826)*. URL: https://www.kaggle.com/dimitrif/keras-with-data-augmentation-lb-0-1826.

[9] Lima Fonseca and Lima Fonseca. *What's happening inside the Convolutional Neural Network? The answer is Convolution*. Nov. 2017. URL: https://buzzrobot.com/whats-happening-inside-the-convolutional-neural-network-the-answer-is-convolution-2c22075dc68d.

[10] Vincent Fung. *An Overview of ResNet and its Variants*. July 2017. URL: https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035.

[11] Kevin Gibson. *Neural networks as Ordinary Differential Equations*. Dec. 2018. URL: https://rkevingibson.github.io/blog/neural-networks-as-ordinary-differential-equations/.

[12] Pontryagin Mishchenko Boltyanskii Gramkrelidize. *The mathematical theory of optimal processes*.

[13] Kaiming He et al. «Deep Residual Learning for Image Recognition». In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[14] Ayeshmantha Perera and Ayeshmantha Perera. *What is Padding in Convolutional Neural Network's(CNN's) padding*. Sept. 2018. URL: https://medium.com/@ayeshmanthaperera/what-is-padding-in-cnns-71b21fb0dd7.

[15] Rajat. *Neural Ordinary Differential Equations and Adversarial Attacks*. URL: https://rajatvd.github.io/Neural-ODE-Adversarial/.

[16] Jonty Sinai. *Understanding Neural ODE's*. Jan. 2019. URL: https://jontysinai.github.io/jekyll/update/2019/01/18/understanding-neural-odes.html.

[17] Nitish Srivastava et al. «Dropout: A Simple Way to Prevent Neural Networks from Overfitting». In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[18] *Statoil/C-CORE Iceberg Classifier Challenge*. URL: https://www.kaggle.com/c/statoil-iceberg-classifier-challenge/data.

[19] Jerome Friedman Trevor Hastie Robert Tibshirani. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[20] *Vanishing gradient problem*. Feb. 2019. URL: https://en.wikipedia.org/wiki/Vanishing_gradient_problem.

[21] Wikipedia contributors. *Euler method — Wikipedia, The Free Encyclopedia*. [Online; accessed 26-March-2019]. 2004. URL: https://en.wikipedia.org/w/index.php?title=Plagiarism&oldid=5139350.

7 Appendix

7.1 Neural Network Information

7.1.1 Definitions

Definition 7.1. *Batch Noramlization: is the process of normalizing activations layer to layer in an effort to increase stability and avoid covariate shift.*

Definition 7.2. *Covariate Shift: is a significant change in distribution of new input data. For example, consider an animal classifier trained on black and white images and used on colored images. This color difference is the cause of a covariate shift in the example.* [6]

Definition 7.3. *Mini Batching: An initial step in stochastic gradient descent where the roots of the gradient are found for only a subset of the data for computational efficiency (and feasibility).*

Definition 7.4. *Momentum: A smoothing factor for the moving mean and variance of the batch normalization process.* [4]

Definition 7.5. *Dense Layer: A dense layer is one that takes a linear combination of all activations from the previous layer for each neuron in its layer.*

Definition 7.6. *Units: The number of neurons present in some layer i of the neural network. The input layer will have units == p where p is the number of input variables.*

7.1.2 Neural Network Code

```r
### Final NN Code Using Iceberg Dataset

## Libraries
library(RJSONIO)
library(keras)
library(abind)
library(kohonen)
library(tidyr)
library(ggplot2)

## Setting seed
set.seed(1)

## Reading in dataset

# Iceberg data
train = fromJSON("train.json")

# Getting relevant information
x <- train %>%
  lapply(function(x){c(x$band_1, x$band_2)}) %>%
  unlist %>%
  array(dim=c(75,75,1604)) %>%
  aperm(c(3,1,2))

# Values for Output
```

```r
y <- classvec2classmat(unlist(lapply (train, function(x) {x$is_iceberg})))

# Training Set list
nums <- sample(1:1604, 1300)

# Organizing
train_iceberg <- x[nums, , ]
train_truth <- y[nums, 2]
test_iceberg <- x[-nums, , ]
test_truth <- y[-nums, 2]

# Class Names
iceberg_name <- c("Not an Iceberg", "An Iceberg")

## Need to scale data
train_iceberg <- train_iceberg/max(abs(train_iceberg))
test_iceberg <- test_iceberg/max(abs(train_iceberg))

## Looking at the first 25 images
par(mfcol=c(5,5))
par(mar=c(0, 0, 1.5, 0), xaxs='i', yaxs='i')
for (i in 1:25) {
  img <- train_iceberg[i, , ]
  img <- t(apply(img, 2, rev))
  image(1:75, 1:75, img, col = gray((-44:0)/-44), xaxt = 'n', yaxt = 'n',
        main = paste(iceberg_name[train_truth[i] + 1]))
}

#### Creating model

# Initialization
iceberg_nn <- keras_model_sequential()

# Adding Layers
iceberg_nn %>%
  layer_flatten(input_shape = c(75, 75)) %>% # Turning image into 784 input variables
  layer_dense(units = 128, activation = 'relu') %>% # 128 neurons with relu activation,
      HL1
  layer_dense(units = 128, activation = 'relu') %>% # 128 neurons with relu activation,
      HL2
  layer_dense(units = 128, activation = 'relu') %>% # 128 neurons with relu activation,
      HL3
  layer_dense(units = 128, activation = 'relu') %>% # 128 neurons with relu activation,
      HL4
  layer_dense(units = 1, activation = 'sigmoid') # Output layer: 1 of 10 things with
      softmax
# activation function

## Densely connected means FULLY-CONNECTED (EACH NEURON IS INVOLVED IN THE CALCULATION OF
# EVERY SINGLE NEURON IN THE NEXT LAYER)

## Adding loss function and optimizer
iceberg_nn %>% compile(
  optimizer = 'sgd', # Using stochastic gradient descent as backprop method
  loss = 'binary_crossentropy', # Using cross-entropy as loss evaluator
  metrics = c('accuracy') # Looking at accuracy
)

## Fitting the model
iceberg_nn %>% fit(train_iceberg, train_truth, epochs = 150)

## Seeing the accuracy
score <- iceberg_nn %>% evaluate(test_iceberg, test_truth)

cat('Test loss:', score$loss, "\n")
cat('Test accuracy:', score$acc, "\n")
```

7.2 Residual Neural Network Information

7.2.1 Definitions

Definition 7.7. *Pooling: Role of the pooling layer is to reduce the resolution of the feature map but retain features of the map required for classification through translational and rotational invariants.*

Definition 7.8. *Dropout: is a regularization technique developed by google to prevent overfitting. The process involves the prevention of "learning" complex patterns within training data.* [17]

Definition 7.9. *Activation-Elu: An exponential linear unit: $f(x) = \alpha \cdot (exp(x) - 1.0)$.*

Definition 7.10. *Padding: is an additional layer added to act on the border of an image suppressing pixels with less information.* [14]

Definition 7.11. *Kernel: A matrix transformation that changes the input image to blur, blacken, sharpen, etc.* [9]

7.2.2 Neural Network Code

```r
### RESNET Final Code

## Libraries
library(RJSONIO)
library(keras)
library(abind)
library(kohonen)
library(tidyr)
library(ggplot2)

## Loading data
train = fromJSON("train.json")

# Getting relevant information
x = train %>% lapply(function(x){
  c(x$band_1,
    x$band_2,
    apply(cbind(x$band_1,x$band_2), 1, mean))}) %>%
  unlist %>%
  array(dim=c(75,75,3,1604)) %>%
  aperm(c(4,1,2,3))

# Values for Output
y <- classvec2classmat(unlist(lapply (train, function(x) {x$is_iceberg})))

# Training Set list
nums <- sample(1:1604, 1300)

# Organizing
train_iceberg <- x[nums, , , ]
train_truth <- y[nums, ]
test_iceberg <- x[-nums, , , ]
test_truth <- y[-nums, ]

## Prepare model
kernel_size = c(5,5)
input_img = layer_input(shape = c(75, 75, 3), name="img")
```

```

## Normalizing data
input_img_norm = input_img %>%
  layer_batch_normalization(momentum = 0.99)

## input CNN
input_CNN = input_img_norm %>%
  layer_conv_2d(32, kernel_size = kernel_size, padding = "same") %>%
  layer_batch_normalization(momentum = 0.99) %>%
  layer_activation_elu() %>%
  layer_max_pooling_2d(c(2,2)) %>%
  layer_dropout(0.25) %>%
  layer_conv_2d(64, kernel_size = kernel_size,padding = "same") %>%
  layer_batch_normalization(momentum = 0.99) %>%
  layer_activation_elu() %>%
  layer_max_pooling_2d(c(2,2)) %>%
  layer_dropout(0.25)

## first residual
input_CNN_residual = input_CNN %>%
  layer_batch_normalization(momentum = 0.99) %>%
  layer_conv_2d(128, kernel_size = kernel_size,padding = "same") %>%
  layer_batch_normalization(momentum = 0.99) %>%
  layer_activation_elu() %>%
  layer_dropout(0.25) %>%
  layer_conv_2d(64, kernel_size = kernel_size,padding = "same") %>%
  layer_batch_normalization(momentum = 0.99) %>%
  layer_activation_elu()

input_CNN_residual = layer_add(list(input_CNN_residual,input_CNN))

# ## second residual
input_CNN_residual = input_CNN_residual %>%
  layer_batch_normalization(momentum = 0.99) %>%
  layer_conv_2d(128, kernel_size = kernel_size,padding = "same") %>%
  layer_batch_normalization(momentum = 0.99) %>%
  layer_activation_elu() %>%
  layer_dropout(0.25) %>%
  layer_conv_2d(64, kernel_size = kernel_size,padding = "same") %>%
  layer_batch_normalization(momentum = 0.99) %>%
  layer_activation_elu()

input_CNN_residual = layer_add(list(input_CNN_residual,input_CNN))

## final CNN
top_CNN = input_CNN_residual %>%
  layer_conv_2d(128, kernel_size = kernel_size,padding = "same") %>%
  layer_batch_normalization(momentum = 0.99) %>%
  layer_activation_elu() %>%
  layer_max_pooling_2d(c(2,2)) %>%
  layer_conv_2d(256, kernel_size = kernel_size,padding = "same") %>%
  layer_batch_normalization(momentum = 0.99) %>%
  layer_activation_elu() %>%
  layer_dropout(0.25) %>%
  layer_max_pooling_2d(c(2,2)) %>%
  layer_conv_2d(512, kernel_size = kernel_size,padding = "same") %>%
  layer_batch_normalization(momentum = 0.99) %>%
  layer_activation_elu() %>%
  layer_dropout(0.25) %>%
  layer_max_pooling_2d(c(2,2)) %>%
  layer_global_max_pooling_2d()

## Output layer
outputs = top_CNN %>%
  layer_dense(512,activation = NULL) %>%
  layer_batch_normalization(momentum = 0.99) %>%
  layer_activation_elu() %>%
  layer_dropout(0.5) %>%
  layer_dense(256,activation = NULL) %>%
  layer_batch_normalization(momentum = 0.99) %>%
  layer_activation_elu() %>%
  layer_dropout(0.5) %>%
  layer_dense(2,activation = "softmax") ## not sure using softmax is the right thing to
     do...

## Setting up model
model_resNN <- keras_model(inputs = list(input_img), outputs = list(outputs))

## Setting up functions for model evaluation and passes
model_resNN %>% compile(optimizer = optimizer_adam(lr = 0.001),
                  loss="binary_crossentropy",
                  metrics = c("accuracy"))

## Fitting the model
model_resNN %>% fit(train_iceberg, train_truth, epochs = 150)

## Trying on test data
predictions_resnet <-  mean(round(predict(model_resNN, test_iceberg))[,2] == test_truth
    [,2])
paste("Test Accuracy (ResNN):", predictions_resnet)
```

7.3 Differential Equations Primer

Discussion in this section will be limited to first order ordinary differential equations. The purpose is to instill enough understanding so that their relevance in Section V is apparent and clear.

7.3.1 General Methodology

Generally, a differential equation relates the values of some function to the values of its derivatives. A first order differential equation is limited to the relationship between a single derivative of a single variable. They are of the form:

$$\frac{dy}{dx} = f(x, y) \tag{22}$$

The function $f(x, y)$ is any of the set of functions which is defined for x (the independent variable) and y (the dependent variable). Accompanying the equation is usually an initial condition which defines the behaviour of the function at some point, x_0[19]. It is sometimes possible to find analytic solutions to differential equations provided they are of a particular form, for example:

$$g(y)\frac{dy}{dx} = f(x), \quad y(x_0) = y_0 \tag{23}$$

[19]The value here is sometimes apparent from the context; for example, consider half-life models in which you know the amount present at time, $t = 0$

But, in general, differential equations are solved numerically[20]. A method falling under the umbrella of numeric methods is presented in the next section.

Let's consider the following ODE:

$$\frac{dy}{dx} + \frac{y}{2} = \frac{3}{2}, \quad y(0) = 2 \tag{24}$$

This differential equation can be solved analytically as follows:

$$\frac{dy}{3-y} = \frac{dx}{2} \tag{25}$$

$$\int_2^y \frac{dy}{3-y} = \frac{1}{2}\int_0^x dx \tag{26}$$

$$\ln(3-y) = -\frac{1}{2} \tag{27}$$

$$3 - y = \exp\{-\frac{x}{2}\} \tag{28}$$

$$y = 3 - \exp\{-\frac{x}{2}\} \tag{29}$$

The solution of the differential equation depends on the initial conditions provided however, the critical points of the function will be clear in any of them. In the above example, when $y = 3$, the derivative is 0 and hence we would expect a horizontal asymptote for any provided initial condition at this value. This behaviour is presented in Figure XX.

Another important visualization tool for differential equations is the phase portrait. The phase portrait allows us to discern important information about the original function, $f(x)$ without actually solving the differential equation. The phase portrait involves computing the roots of the function (the 0's of the derivative) and plotting the behaviour of the derivative for various values of the dependent variable, y. Consider the following differential equation [7]:

$$\frac{dy}{dt} = r(1 - \frac{y}{K})y \tag{30}$$

Where $K = \frac{r}{a}$ and r is known as an 'intrinsic growth rate'[21]. The first step in identifying the phase portrait is to find the 0's; in this case, if we let $y = f(x) = \{0 \cap K\}$, then the value of $\frac{dy}{dt}$ in (13) is 0 - these are known as the **equilibrium solutions**. This is when there is no change in the *variation* of y as t changes. From there, we

[20]The equation in (6) is known as a separable differential equation because you can split the $f(x, y)$ in (5) into two separate functions

[21]This equation is an extension on the exponential growth function and is commonly referred to as the Verhulst or logistic equation

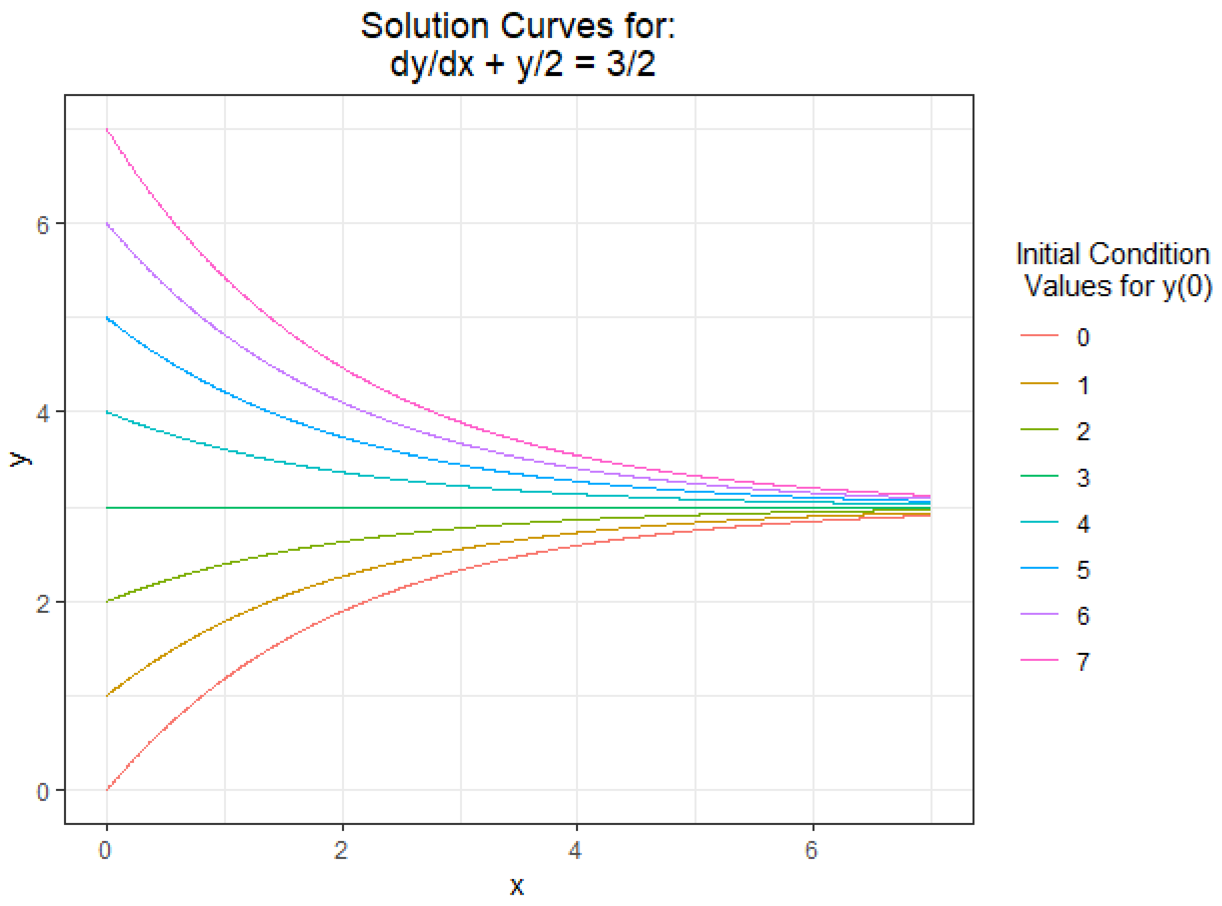

Figure 9: Solution curves for the ODE given in (7). The particular initial condition that solved for is given by the sea green curve

can observe the behaviour of the derivative around these equilibrium points.

In Figure XXX, the arrows point to the right for positive derivative values and to the left otherwise. You can imagine these arrows as directions for convergence. The **stable equilibrium** is some value that the system being modelled tends to (also known as a saturation level). Intuitively, imagine you were modelling population growth; then, you would expect quicker growth the greater the population size however, at some point resources become limited. This results in a decrease (or stabilization) of the population at some level which, in Figure XXX, is the second equilibrium point. The unstable equilibrium in this example, occurs when the population is 0 - we expect no growth if no one is around however, as soon as it is possible to move away from this position, we tend to do so.

7.3.2 A Numeric Approximation: Euler Method

In the previous section, I introduced differential equations, an example of an analytic solution, and visualization tool to glean information about the function $y(x)$ without actually solving the equation; here, I introduce a numerical approach to solving dif-

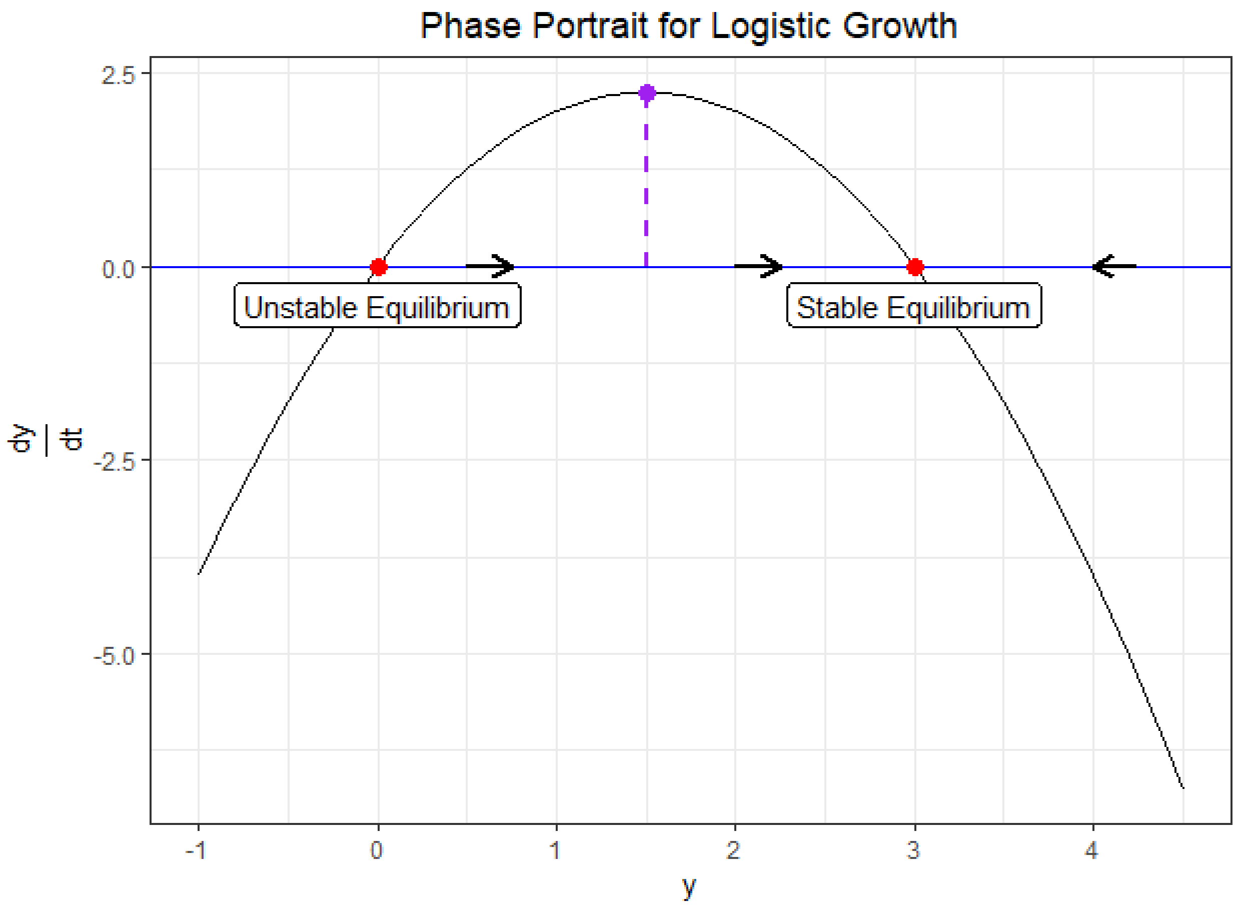

Figure 10: A phase portrait for the Verhulst Equation defined in (13)

ferential equations: Euler method[22]

Euler's method is an iterative approach that, for some change in x, provides an estimate of the function, $f(x)$ using the derivative in the interval, $\Delta(x)$. In order to make this more concrete, consider the differential equation [21]:

$$\frac{dy}{dx} = y, \quad y(0) = 1 \tag{31}$$

The solution, $f(x)$ to this equation is $y = \exp\{x\}$. However, let's assume that we didn't have the means to find the analytic solution and instead, use Euler's method to approximate $f(x)$. Let $\Delta(x) = 1$ be the iterative interval and consider $x = [0, 4]$. At $x = 0$, we have $y = 1$. The derivative at this point is also 0. Then, using the iterative process: $y_{i+1} = y_i + \Delta(x)\frac{dy}{dx}$, we see that $y_1 = 1 + 1 \cdot 0$ and, redoing this process, we get the following:

[22]This approach is a specification of a more general approach known as Runge-Kutta

Iteration	x	y	$\frac{dy}{dx}$
0	0	1	1
1	1	2	2
2	2	4	4
3	3	8	8
4	4	16	16

Essentially, we are figuring out the tangent lines for intervals and connecting them - this is our approximation! Note that, as $\Delta(x) \to 0$, our approximation approaches the exact solution. A visualization is provided in Figure XXXX.

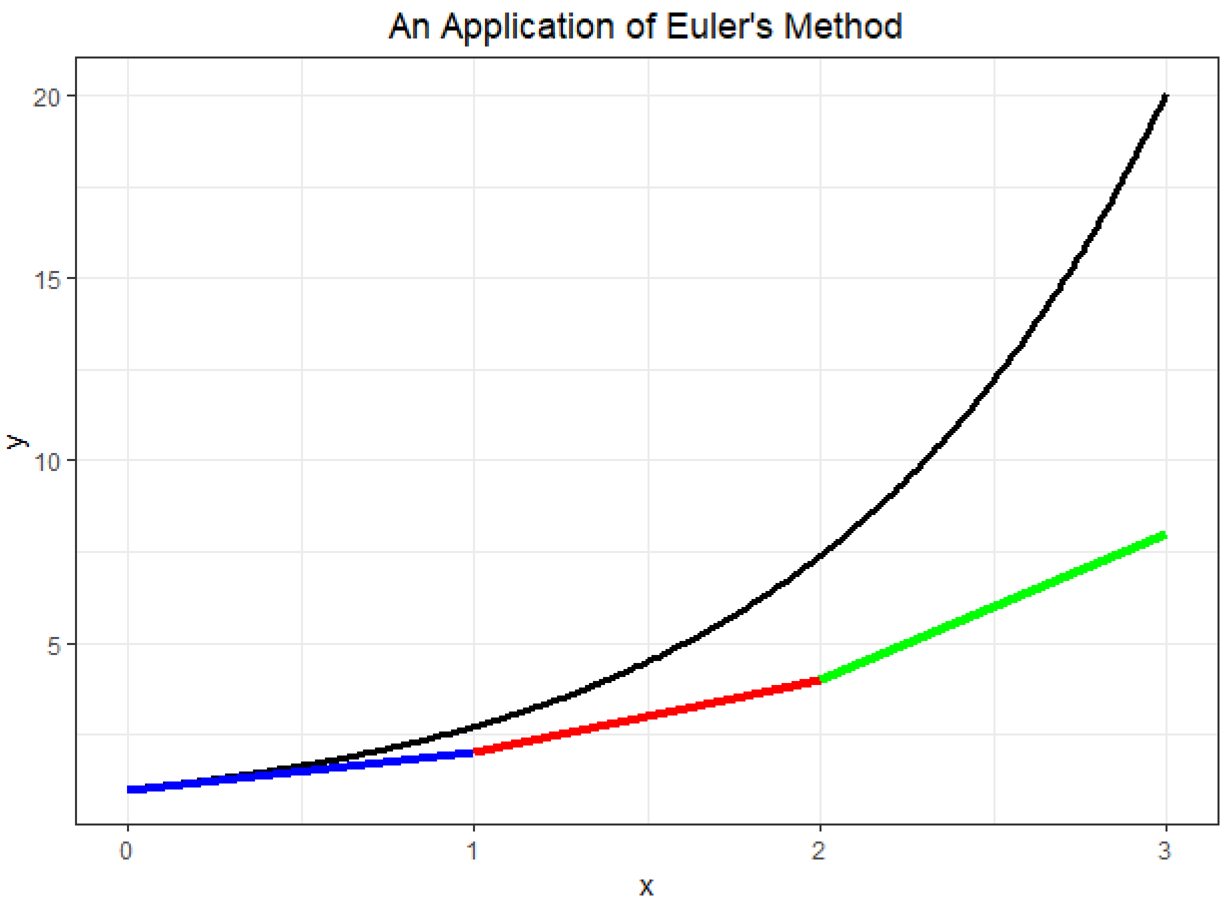

Figure 11: Euler method for the differential equation in (XXX). The colored segments represent the Euler approximation with a step size of 1. The black curve is the true function: $y = \exp\{x\}$

This brings an end to the primer. The approach in Euler's method is particularly important when comparing ResNets to NeuralODEs in the next section!

7.4 On the relationship between ResNets and Euler's Method

Consider again the general transformation performed in ResNets[23] [15]:

[23]In the main section, $z_1 = a_1$

$$a_1 = f_0(x) + x \tag{32}$$
$$a_2 = f_1(a_1) + a_1 \tag{33}$$
$$a_3 = f_2(a_2) + a_2 \tag{34}$$

Rearranging these, we can observe that this, is almost exactly the form of Euler's method! Letting $a(t = 0) = x$,

$$a(1) - a(0) = f(a(0), t = 0) \tag{35}$$
$$a(2) - a(1) = f(a(1), t = 1) \tag{36}$$
$$... \tag{37}$$

Recall that Euler's method is a descritization dependent on the step size, Δh. In essence, the key insight of the ResNets approach is characterized by a rearranging of Euler's method. If that is the case, then we are essentially solving a differential equation. A Neural ODE takes another step backwards in the process by appealing to the fundamental equation underlying the neural net rather than looking at the intermediary step that ResNets do.