Reimplementation of the Classical Neural Ordinary Differential Equation Using Modern Computational Tools

Wasif Ul Islam [1]

1. Introduction

Residual Neural Networks (RNN) provided a significant improvement on performance regarding deep layered neural networks. The performance gains were obtained from introducing residual elements applied to the ReLu operations within the hidden layer. This introduced robustness against the vanishing/exploding gradient problem that most classical neural networks face (Shorten, 2019).

However, there are some application of RNNs that still did not provide the ideal flexibility, in terms of training models for data involving time-series data. Since RNNs have discrete hidden layers and residual components to the ReLu operations performed to the hidden layers, it can interfere with learning dependencies which are sensitive to time (Walther et al., 2023).

A different class of neural networks look to satisfy the drawbacks that the previously mentioned network contain: Neural Ordinary Differential Equations (NeuralODE). The prime benefit that NeuralODEs provide is it is naturally suited for modelling continuos time data. In addition, it can provide computational flexibility in terms of configurability in forward-pass and backpropagation.

This paper's objective is to provide a brief literature review into works that have enabled the use of NeuralODEs and discuss the advantages and disadvantages related to the use of NeuralODEs. As many pieces of work focus on performance of NeuralODEs on time-series datasets, a short experiment involving the classification of the MNIST dataset using a proof-of-concept NeuralODE model is presented along with comparisons of performance with a typical proof-of-concept implementation of a residual neural network.

[1]School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana USA.

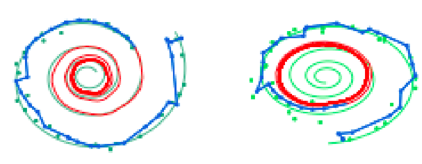
Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

2. Related Works

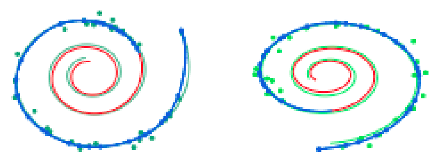
2.1. Article: Neural Ordinary Differential Equations

The concept of NeuralODEs have been discussed before the prescribed implementation by the paper *Neural Ordinary Differential Equations*. However, the primitive approach to the implementation of NeuralODEs had very large computational resource consumption, to the point where training the model on a substantial scale was not feasable. Since the NeuralODE consist of one continous layer with specific time steps for numerical integration (Derivation of Neural ODE vs. Classical NN is explained further below), performing backpropagation on NeuralODEs would be taking the gradient across each of the time steps. Since solving ODEs require small timesteps for accurate representation of the dynamics that is being learned, ordinary back-propagation become as resource dependant as a ResNet with hundreds of layers.

The major contribution for the respective paper is the adjoint sensitvity method, which allows for the back-propagation of NeuralODE block at near constant space-complexity (He et al., 2015). A comparison between a Recurrent Neural Network with 25 hidden-layers and the proposed NeuralODE architecture was presented with time-series data.



(a) Recurrent Neural Network



(b) Latent Neural Ordinary Differential Equation

Figure 1. Comparison Between Recurrent Neural Network and NeuralODE on Time-Series Data (He et al., 2015)

2.2. Article: Augmented Neural ODE

The family of NeuralODEs have been extended from the previous contribution to make it more generalizable for existing framworks. The paper *Augmented Neural ODEs* proposes several postulates. The first assertion is that the classical implementation of NeuralODEs with the use of adjoint sensitivity method has limitations regarding representation of functions containing intersecting vector flows. The second assertion involves the existance of output space being rigid to the input space. The proposed solution to the problem was to introduce higher-dimensionality to the output space. This allows the ODE learning algorithm to lift points into additional dimension without having differential collisions (Dupont et al., 2019). The higher dimensionality mapping allowed for functions that were not representable by the classical NeuralODEs, available for training.

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathbf{h}(t)\\ \mathbf{a}(t)\end{bmatrix} = \mathbf{f}(\begin{bmatrix}\mathbf{h}(t)\\ \mathbf{a}(t)\end{bmatrix}, t), \qquad \begin{bmatrix}\mathbf{h}(0)\\ \mathbf{a}(0)\end{bmatrix} = \begin{bmatrix}\mathbf{x}\\ \mathbf{0}\end{bmatrix}$$

Figure 2. Activation Function Augmented with Higher Dimensionality (Dupont et al., 2019)

2.3. Article: How to Train Your Neural ODE

Beside the adjoint sensitivity method, developed by the authors of the paper *Neural Ordinary Differential Equations*, parallel efforts were also taken to address concerns of high resource usage of back-propagation of NeuralODEs. The paper *How To Train Your Neural ODE* proposes a method to reduce the computational complexity of back-propagation in time-series data predictions, by introducing regularization terms to the loss-function (Finlay et al., 2020). This allows for simpler solutions that can be sufficient while limiting NFEs (Number of Function Evaluation) to a reasonable number. The first regularization term penalty enourages particles to travel with one-dimensional translation by using optimal transport mapping, while the second regularization term enforces limitations of the vector-field Jacobian to provide force experienced to be as constant as possible. Such constraint of dynamics allow for simpler solutions with less computational complexities (Finlay et al., 2020).

2.4. A Comparative Derivation of Neural ODE

Conceptially, NeuralODEs are an extension of Residual Neural Networks. As mentioned previously, the difference between the respective family of neural networks is the continuos nature of NeuralODEs vs. the discrete nature of Residual Neural Networks. To show the difference, we will go from one forward-pass calculation of a Residual Neural Network to a NeuralODE.

Residual Neural Network

$$h_{t+1} = ReLu(W_t h_t + b_t) + h_t$$

The additional hidden layer term added after the activation function is the residual component of the Residual Neural Network.

Containing the activation function with standard notations, we can rewrite the equation as:

$$h_t = f(h_t, \theta_t) + h_t$$

We can convert the following equation to its differential form by expressing it in terms of its limits

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

If we reduce the change between one hidden layer to the next to be infinitesimally small, we can express the equation as:

$$\frac{h_{t+\delta} - h_t}{\delta} = f(h_t, \theta_t)$$

Taking the limit of δ and setting it to zero provides us with the differential form of the equation:

$$\lim_{\delta \to 0} \frac{h_{t+\delta} - h_t}{\delta} = f(h_t, \theta_t)$$

$$\frac{\mathbf{dh_t}}{\mathbf{dt}} = \mathbf{f}(\mathbf{h_t}, \theta_\mathbf{t})$$

- Legend:
  - h_t: Hidden Layer t
  - W_t: Weight Matrix t
  - b_t: Bias Vector t
  - θ_t: Layer Parameters t
  - δ: Change in Hidden Layer (Infinitesimally Small)

3. Experiment

3.1. Defining the Experiment

An implementation of the Proof-of-Concept NeuralODE is developed. The workflow is modified from a typical NeuralODE implementation to accomodate visual classification of the MNIST dataset

The paper body should be set in 10 point type with a vertical spacing of 11 points. Please use Times typeface throughout the text.

3.2. Title

The paper title should be set in 14 point bold type and centered between two horizontal rules that are 1 point thick, with 1.0 inch between the top rule and the top edge of the page. Capitalize the first letter of content words and put the rest of the title in lower case.

3.3. Author Information for Submission

ICML uses double-blind review, so author information must not appear. If you are using LaTeX and the `icml2021.sty` file, use `\icmlauthor{...}` to specify authors and `\icmlaffiliation{...}` to specify affiliations. (Read the TeX code used to produce this document for an example usage.) The author information will not be printed unless `accepted` is passed as an argument to the style file. Submissions that include the author information will not be reviewed.

3.3.1. Self-Citations

If you are citing published papers for which you are an author, refer to yourself in the third person. In particular, do not use phrases that reveal your identity (e.g., "in previous work (Langley, 2000), we have shown ...").

Do not anonymize citations in the reference section. The only exception are manuscripts that are not yet published (e.g., under submission). If you choose to refer to such unpublished manuscripts (Author, 2021), anonymized copies have to be submitted as Supplementary Material via CMT. However, keep in mind that an ICML paper should be self contained and should contain sufficient detail for the reviewers to evaluate the work. In particular, reviewers are not required to look at the Supplementary Material when writing their review.

3.3.2. Camera-Ready Author Information

If a paper is accepted, a final camera-ready copy must be prepared. For camera-ready papers, author information should start 0.3 inches below the bottom rule surrounding the title. The authors' names should appear in 10 point bold type, in a row, separated by white space, and centered. Author names should not be broken across lines. Unbolded superscripted numbers, starting 1, should be used to refer to affiliations.

Affiliations should be numbered in the order of appearance. A single footnote block of text should be used to list all the affiliations. (Academic affiliations should list Department, University, City, State/Region, Country. Similarly for industrial affiliations.)

Each distinct affiliations should be listed once. If an author has multiple affiliations, multiple superscripts should be placed after the name, separated by thin spaces. If the authors would like to highlight equal contribution by multiple first authors, those authors should have an asterisk placed after their name in superscript, and the term "*Equal contribution" should be placed in the footnote block ahead of the list of affiliations. A list of corresponding authors and their emails (in the format Full Name <email@domain.com>) can follow the list of affiliations. Ideally only one or two names should be listed.

A sample file with author names is included in the ICML2021 style file package. Turn on the `[accepted]` option to the stylefile to see the names rendered. All of the guidelines above are implemented by the LaTeX style file.

3.4. Abstract

The paper abstract should begin in the left column, 0.4 inches below the final address. The heading 'Abstract' should be centered, bold, and in 11 point type. The abstract body should use 10 point type, with a vertical spacing of 11 points, and should be indented 0.25 inches more than normal on left-hand and right-hand margins. Insert 0.4 inches of blank space after the body. Keep your abstract brief and self-contained, limiting it to one paragraph and roughly 4–6 sentences. Gross violations will require correction at the camera-ready phase.

3.5. Partitioning the Text

You should organize your paper into sections and paragraphs to help readers place a structure on the material and understand its contributions.

3.5.1. Sections and Subsections

Section headings should be numbered, flush left, and set in 11 pt bold type with the content words capitalized. Leave 0.25 inches of space before the heading and 0.15 inches after the heading.

Similarly, subsection headings should be numbered, flush left, and set in 10 pt bold type with the content words capitalized. Leave 0.2 inches of space before the heading and 0.13 inches afterward.

Finally, subsubsection headings should be numbered, flush left, and set in 10 pt small caps with the content words capitalized. Leave 0.18 inches of space before the heading and 0.1 inches after the heading.

Please use no more than three levels of headings.

3.5.2. Paragraphs and Footnotes

Within each section or subsection, you should further partition the paper into paragraphs. Do not indent the first line of a given paragraph, but insert a blank line between succeeding ones.

You can use footnotes[1] to provide readers with additional information about a topic without interrupting the flow of the paper. Indicate footnotes with a number in the text where the point is most relevant. Place the footnote in 9 point type at the bottom of the column in which it appears. Precede the first footnote in a column with a horizontal rule of 0.8 inches.[2]

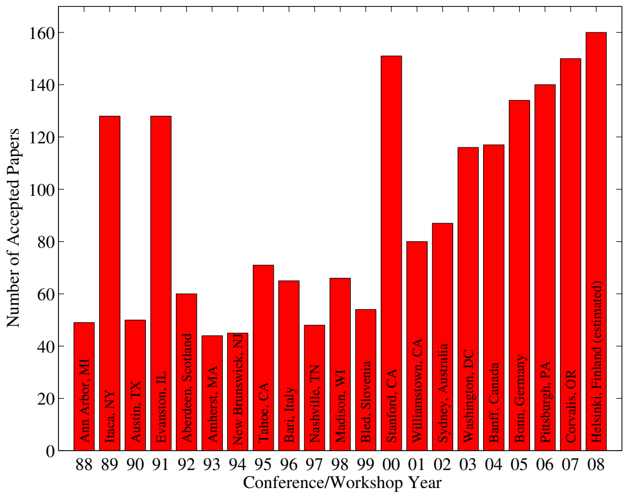


Figure 3. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

3.6. Figures

You may want to include figures in the paper to illustrate your approach and results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction, and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should serve this function.

Number figures sequentially, placing the figure number and caption *after* the graphics, with at least 0.1 inches of space

[1] Footnotes should be complete sentences.

[2] Multiple footnotes can appear in each column, in the same order as they appear in the text, but spread them across columns and pages if possible.

Algorithm 1 Bubble Sort

Input: data x_i, size m
repeat
  Initialize $noChange = true$.
  for $i = 1$ **to** $m - 1$ **do**
    if $x_i > x_{i+1}$ **then**
      Swap x_i and x_{i+1}
      $noChange = false$
    end if
  end for
until $noChange$ is $true$

Table 1. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

Data set	Naive	Flexible	Better?
Breast	95.9± 0.2	96.7± 0.2	$\surd$
Cleveland	83.3± 0.6	80.0± 0.6	$\times$
Glass2	61.9± 1.4	83.8± 0.7	$\surd$
Credit	74.8± 0.5	78.3± 0.6	
Horse	73.3± 0.9	69.7± 1.0	$\times$
Meta	67.1± 0.6	76.5± 0.5	$\surd$
Pima	75.1± 0.6	73.9± 0.5	
Vehicle	44.9± 0.6	61.5± 0.4	$\surd$

before the caption and 0.1 inches after it, as in Figure 3. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left. You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX). Always place two-column figures at the top or bottom of the page.

3.7. Algorithms

If you are using LaTeX, please use the "algorithm" and "algorithmic" environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

3.8. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title *above* the table with at least 0.1 inches of space before the title and the same after it, as in Table 1. The table title should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left.

Tables contain textual material, whereas figures contain graphical material. Specify the contents of each row and column in the table's topmost row. Again, you may float tables to a column's top or bottom, and set wide tables across both columns. Place two-column tables at the top or bottom of the page.

3.9. Citations and References

Please use APA reference format regardless of your formatter or word processor. If you rely on the LaTeX bibliographic facility, use `natbib.sty` and `icml2021.bst` included in the style-file package to obtain this format.

Citations within the text should include the authors' last names and year. If the authors' names are included in the sentence, place only the year in parentheses, for example when referencing Arthur Samuel's pioneering work (1959). Otherwise place the entire reference in parentheses with the authors and year separated by a comma (Samuel, 1959). List multiple references separated by semicolons (Kearns, 1989; Samuel, 1959; Mitchell, 1980). Use the 'et al.' construct only for citations with three or more authors or after listing all authors to a publication in an earlier reference (Michalski et al., 1983).

Authors should cite their own work in the third person in the initial version of their paper submitted for blind review. Please refer to Section 3.3 for detailed instructions on how to cite your own papers.

Use an unnumbered first-level section heading for the references, and use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 10 points. The references at the end of this document give examples for journal articles (Samuel, 1959), conference publications (Langley, 2000), book chapters (Newell & Rosenbloom, 1981), books (Duda et al., 2000), edited volumes (Michalski et al., 1983), technical reports (Mitchell, 1980), and dissertations (Kearns, 1989).

Alphabetize references by the surnames of the first authors, with single author entries preceding multiple author entries. Order references for the same authors by year of publication, with the earliest first. Make sure that each reference includes all relevant information (e.g., page numbers).

Please put some effort into making references complete, presentable, and consistent. If using bibtex, please protect capital letters of names and abbreviations in titles, for example, use {B}ayesian or {L}ipschitz in your .bib file.

Software and Data

If a paper is accepted, we strongly encourage the publication of software and data with the camera-ready version of the paper whenever appropriate. This can be done by including a URL in the camera-ready copy. However, **do not** include URLs that reveal your institution or identity in your submission for review. Instead, provide an anonymous URL or upload the material as "Supplementary Material" into the CMT reviewing system. Note that reviewers are not required to look at this material when writing their review.

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and probably should) include acknowledgements. In this case, please place such acknowledgements in an unnumbered section at the end of the paper. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

References

Author, N. N. Suppressed for anonymity, 2021.

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.

Dupont, E., Doucet, A., and Teh, W. Y. Augmented neural odes. Number 32, 2019.

Finlay, C., Jacobsen, J.-H., Nurbekyan, L., and Oberman, A. M. How to train your neural ode. 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. pp. 770–778. IEEE, 2015.

Kearns, M. J. *Computational Complexity of Machine Learning*. PhD thesis, Department of Computer Science, Harvard University, 1989.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (eds.). *Machine Learning: An Artificial Intelligence Approach, Vol. I*. Tioga, Palo Alto, CA, 1983.

Mitchell, T. M. The need for biases in learning generalizations. Technical report, Computer Science Department, Rutgers University, New Brunswick, MA, 1980.

Newell, A. and Rosenbloom, P. S. Mechanisms of skill acquisition and the law of practice. In Anderson, J. R. (ed.), *Cognitive Skills and Their Acquisition*, chapter 1, pp. 1–51. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1981.

Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):211–229, 1959.

Shorten, C. Introduction to resnets, 2019. URL https://towardsdatascience.com/introduction-to-resnets-c0a830a288a4. Accessed: April 10th, 2023.

Walther, D., Viehweg, J., Haueisen, J., and Mader, P. A systematic comparison of deep learning methods for eeg time series analysis. *Frontiers in Neuroinformatics*, 17, 2023.

A. Do *not* have an appendix here

Do not put content after the references. Put anything that you might normally include after the references in a separate supplementary file.

We recommend that you build supplementary material in a separate document. If you must create one PDF and cut it up, please be careful to use a tool that doesn't alter the margins, and that doesn't aggressively rewrite the PDF file. pdftk usually works fine.

Please do not use Apple's preview to cut off supplementary material. In previous years it has altered margins, and created headaches at the camera-ready stage.