Promising Hyperparameter Configurations for Deep Fully Connected Neural Networks to Improve Image Reconstruction in Proton Radiotherapy

S. A. York[1], A. M. Ali[2], D. C. Lashbrooke Jr.[3], R. Yepez-Lopex[4], C. A. Barajas[5], M. K. Gobbert[5], J. C. Polf[6]

[1]Center for Data, Mathematics, and Computational Sciences, Goucher College
[2]Dept. of Mathematics and Statistics, University of Houston–Downtown
[3]Dept. of Mathematics and Department of Statistics, Purdue University
[4]Department of Computer Science, American University
[5]Dept. of Mathematics and Statistics, University of Maryland, Baltimore County
[6]Dept. of Radiation Oncology, University of Maryland School of Medicine

REU 2021 Symposium, collocated at the IEEE BigData 2021 Conference
December 15, 2021

Acknowledgments: NSF (Big Data REU Site , MRI), NIH, UMBC

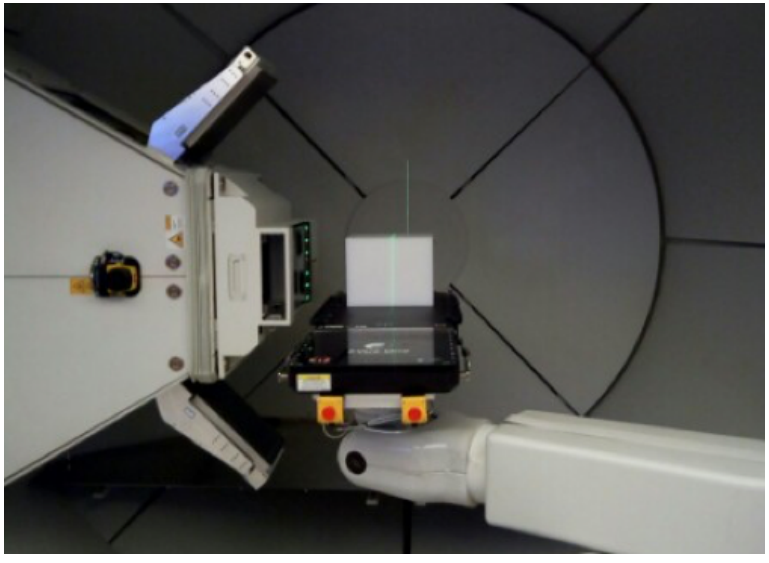
Table of Contents

Maryland Proton Treatment Center

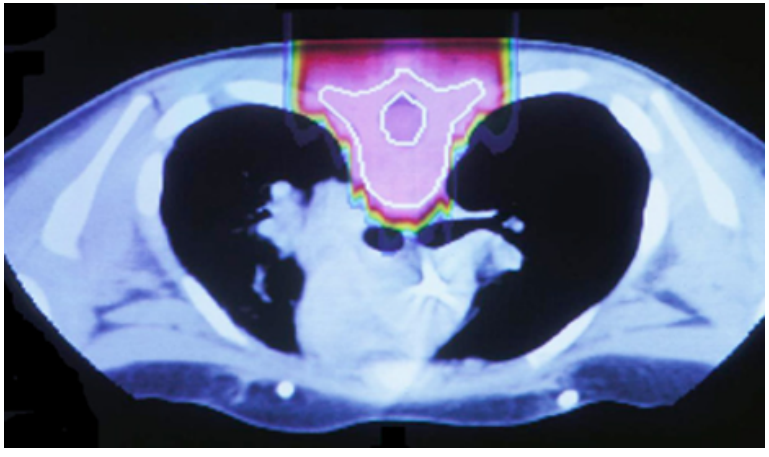
Maryland Proton Treatment Center

This work was done in collaborations with the Maryland Proton Treatment Center located in Baltimore, Maryland. Opened in 2016, the center was the first in the Maryland/DC region to offer proton therapy for cancer treatment. In the past four years it has trained more than 200 health care professionals in proton therapy and, with its state of the art facilities and four treatment rooms, has been able to treat over 2,000 patients (`www.mdproton.com`).

Proton Beam Therapy



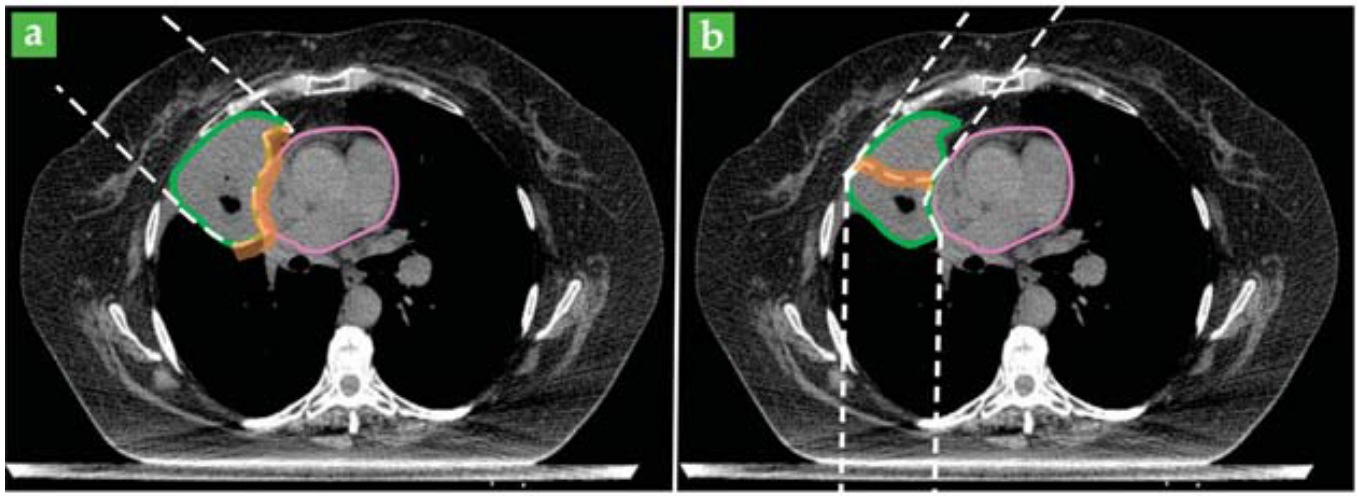
Treatment table in MD Proton Treatment Center (Maggi 2019)



Radiation levels in proton beam therapy

Proton beams' advantage in cancer research is their finite range. they reach their highest dose just before they stop, at what is called the Bragg peak. Little to no radiation is delivered beyond this point.

The Need for Real-Time Imaging

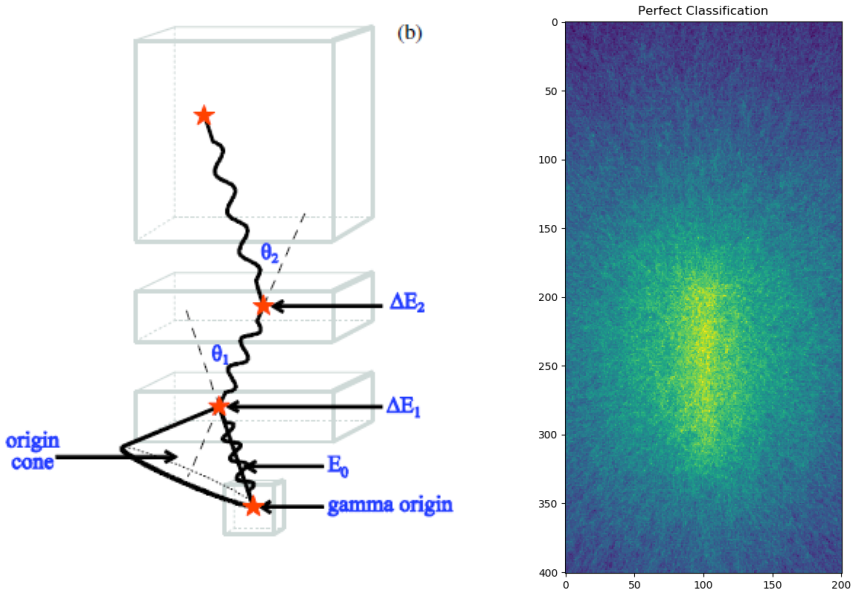


(a) Optimal trajectory. Notice the orange bar, which represents uncertainty, intersects the heart (magenta) but completely covers the tumor (green).

(b) Suboptimal trajectory necessary to protect heart (magenta). A low dosage irradiates healthy lung tissue (black) while still covering the tumor (green). (Polf, *Physics Today*, 2015).

- Uncertainties in the beam's position limit proton beam therapy's advantages.
- Imaging the beam in near real time would reduce uncertainties and allow the advantages of the Bragg peak to be fully exploited.

Image Reconstruction Using Compton Camera



(Phys. Med. Biol., 2010.)

(a) (b)

- (a) Nuclear reactions between beam and tissue produce prompt gamma rays. A Compton camera records the position and energies of each interaction.
- (b) By analysing how prompt gammas scatter through the Compton camera we can reconstruction their origin, thereby imaging the beam.

Limitations of the Compton Camera

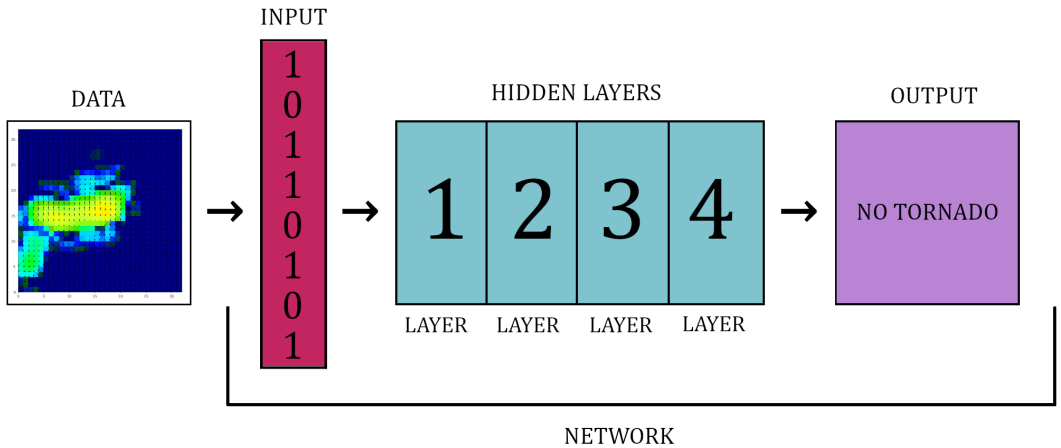
- The Compton camera simply records events as a **single**, **double**, or **triple** scatters.
- The Compton camera cannot determine the correct orderings of camera events.
- The Compton camera cannot determine if a **double** or **triple** scatter event was triggered by prompt gammas originating from different physics events that just happened to enter the camera at the same time.
  1. This lack of distinction means that **single** events can end up paired together as a **double** or **triple** event.
  2. Coupled **single**s are referred to as **false** events.
  3. A double to triple or, **dtot**, is when a true **double** is incorrectly paired with a **single** which appears as a **triple**.

The Data Layout

	Interaction 1				Interaction 2				Interaction 3			
event 1	e1	x1	y1	z1	e2	x2	y2	z2	e3	x3	y3	z3
event 2	e1	x1	y1	z1	e2	x2	y2	z2	e3	x3	y3	z3

- An **interaction** consists of all of the data for a gamma ray's specific collision.
- An **event** is all three **interaction**s together.
- Since the Compton camera cannot determine the correct ordering, the **interaction**s are mixed up causing noise and corruption.
- If we say that an event is a 123 event that means that the **interaction**s are correctly ordered. If an event is a 231 event then we know that the 2nd **interaction** is actually the 1st, the 3rd is actually the 2nd, and the 1st is actually the 3rd.
- Doubles Data uses simplified 2-digit labels 12, 21, 44.

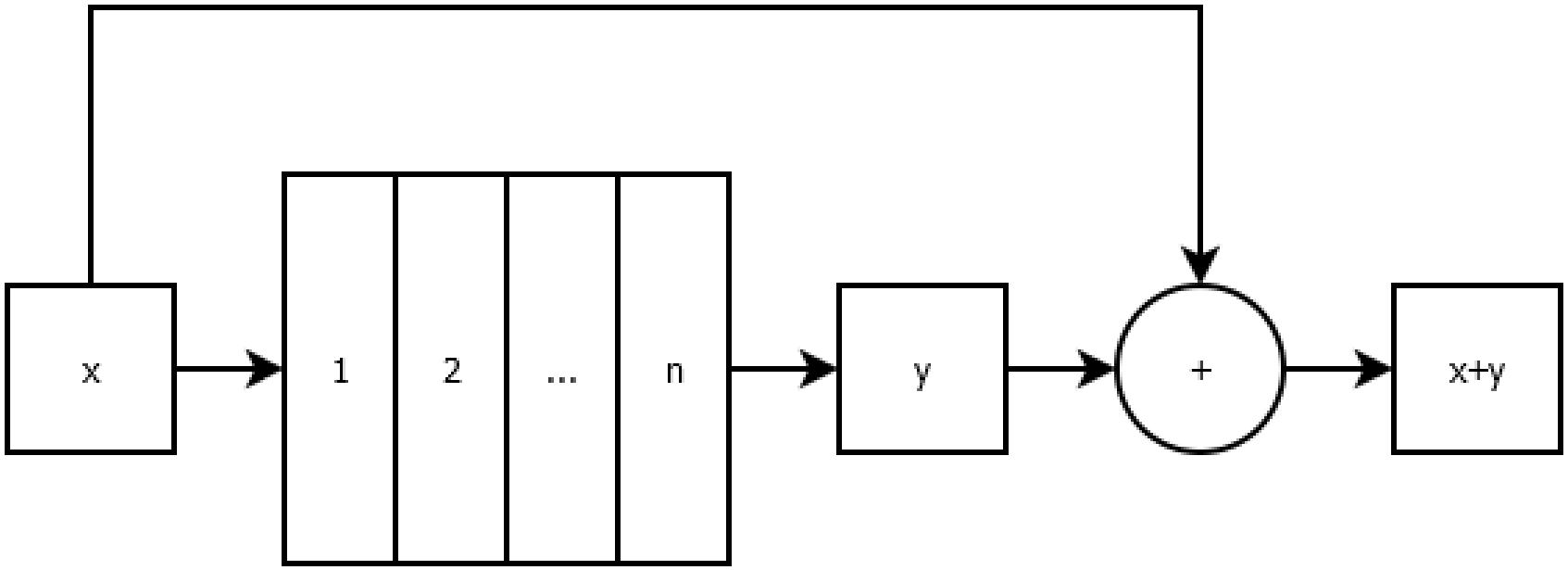
Deep Fully Connected Neural Network at a High Level



- The general mode of operation in the feed-forward process is:
  1. A layer takes in some input vector.
  2. The layer multiplies a matrix of weights with the input vector then adds bias to the result.
  3. We then use a non-linear transformation, called an activation function, on the result.
- This process is repeated for every layer until the final output is obtained.
- Check how close the output vector is to the desired vector using a loss function.

Chollet, *Deep Learning with Python*, 2018

Residual Blocks



- Our network uses residual blocks with our fully connected layers instead of simply stacking them.
- A block consists of some number of layers.
- The input to the first layer is concatenated to output of the block's last layer.
- That concatenated data is passed as the input to the next block.
- This helps remove problem typically associated with deep networks and fully connected networks.

Our Network Constants

Hyperparameter	Value
Inter-layer activation	Leaky ReLU
Final activation	Softmax
Output Layer	13 Neurons
Optimizer	Adam
Loss Function	Categorical Crossentropy
Block size	8
Learning Rate	10^{-3}

- These are the hyperparamters that we hold constant during our hyperparameter study to create our neural network that we will use for training and testing.
- We also transform our energy deposition data before use via sklearn's `PowerTransformer` Yeo-Johnson method.
- We transform our spatial data using sklearn's `MaxAbsScaler`.
- The above parameters are determined by a hyperparameter experimentation done previously.

Hardware and Software Used

Hardware: UMBC High Performance Computing Facility (`hpcf.umbc.edu`).

- HPCF2018 GPU Cluster
  - 1 node contains four NVIDIA Tesla V100 GPUs (5120 computational cores over 84 SMs, 16 GB onboard memory), connected by NVLink, and two 18-core Intel Skylake CPUs.
  - The node has 384 GB of memory and a 120 GB SSD disk.

Software packages used:

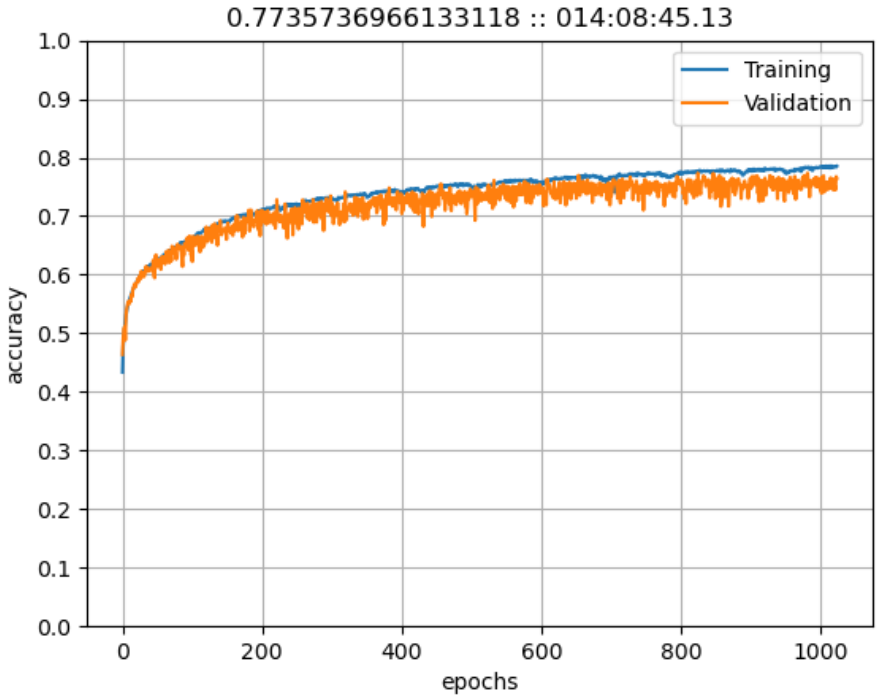
- Python 3.7.6
- Tensorflow 2.4.0
- Keras TF2.4.0
- Numpy 1.16.0
- sklearn 0.23.dev0

Hyperparameter Study Set

Hyperparameter	Possible Values
Dropout rate	0, 0.1, 0.4
Neurons per layer	32, 64, 128, 256
Total Layers	8, 16, 32, 64, 128, 256
Batch size	1024, 2048, 4096, 8192

- Our hyperparameter study was done using the grid search method.
- With the grid search method we have $(3)(6)(4)(4) = 288$ total studies to run.

Best Performing Network



- All of our networks trained for 1024 epochs.
- The network had a dropout rate of 0, 64 neurons per layer, 256 layers, and a 2048 batch size.
- We see that our training accuracy and validation accuracy are between 70% and 80% for most of the training process.
- Our peak validation accuracy is approximately 77%.

Best Performing Network's Confusion Matrix – 20kMU

	123	132	213	231	312	321	124	214	134	314	234	324	444
123	66.3	8.1	2.1	3.4	3.0	2.7	8.1	0.6	0.1	0.1	3.5	1.4	0.6
132	3.8	71.2	2.6	2.2	2.9	3.1	0.3	0.1	7.8	0.8	1.4	3.3	0.5
213	3.3	3.5	70.6	3.4	2.0	2.9	1.1	6.7	4.4	1.2	0.3	0.0	0.6
231	1.5	3.2	4.8	71.1	3.0	5.4	0.1	0.4	1.6	2.5	4.8	1.1	0.3
312	2.7	2.7	2.3	2.7	74.6	4.0	3.0	1.5	0.9	5.1	0.1	0.2	0.3
321	2.5	3.3	2.8	2.2	5.9	72.1	1.4	2.8	0.0	0.3	0.7	5.6	0.3
124	3.5	0.4	0.6	0.1	3.2	2.1	71.5	9.3	0.9	0.4	0.4	1.7	5.8
214	0.8	0.3	4.5	0.3	2.3	3.3	13.6	66.8	0.5	1.2	0.8	0.5	5.0
134	0.4	4.0	3.7	2.5	0.4	0.1	1.0	0.6	71.3	8.8	1.8	0.5	5.0
314	0.1	0.8	2.1	5.1	6.8	0.4	0.3	1.4	8.9	66.5	0.6	0.9	6.1
234	2.6	2.4	0.3	7.6	0.1	1.5	0.8	1.1	1.3	0.7	62.2	13.8	5.7
324	1.3	4.6	0.2	0.8	0.2	8.0	1.1	0.6	0.9	0.6	8.9	67.6	5.2
444	0.6	0.3	0.9	0.6	0.6	0.3	6.3	6.6	4.1	5.0	8.5	6.3	59.9

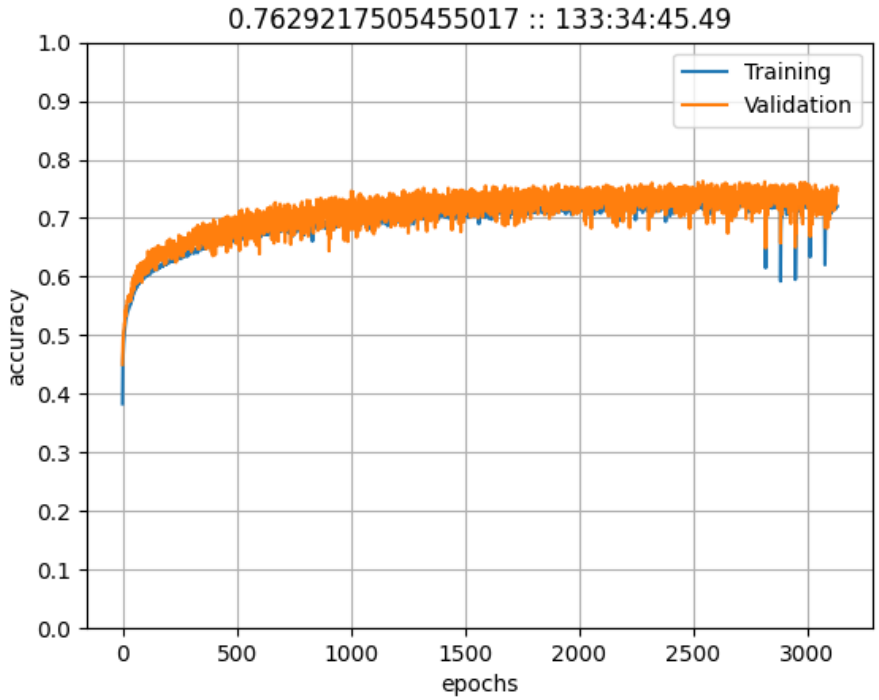
- The dominant classification is the correct class for all classes.
- For **triple**s, the second highest prediction is the corresponding **dtot**.
- For **dtot**, the second highest prediction is a reversed double ordering.
- For **false triples**, we see it is often mistaken as a **dtot**.
- **Triples** often become valid doubles which keeps the data usable.
- Reverse **dtot** orderings can be recovered using a network trained on **doubles**.
- The testing accuracy for each class is considerably lower than the average validation accuracy seen in our previous plot. This is usually due to no dropout rate.

Long Running Studies

- We wanted to see if training our best performing studies yielded even better better results.
- We extend the studies from 1024 epochs to 4096 epochs.
- We opt for a dropout rate greater than 0 to help combat the training/validation testing accuracy discrepancy.
- We used the following parameters sets for our longer running, reduced set, of hyperparameter studies:

Hyperparameter	Values
Dropout Rate	0.1
Neurons per layer	64, 128
Number of layers	64, 128
Batch size	2048, 4096, 8192

Best Long Running Network



- The network had a dropout rate of 0.1, 128 neurons per layer, 128 layers, and a 8192 batch size.
- We see that our training accuracy and validation accuracy are between 70% and 80% for most of the training process.
- Our peak validation accuracy is approximately 77%.

Best Long Running Network's Confusion Matrix – 20kMU

	123	132	213	231	312	321	124	214	134	314	234	324	444
123	77.0	6.4	1.4	1.9	1.9	2.0	5.0	0.4	0.1	0.0	2.5	1.0	0.3
132	3.8	72.8	1.9	2.4	6.4	2.7	0.2	0.0	5.1	1.0	1.0	2.6	0.2
213	3.1	2.5	71.3	9.1	1.8	2.5	0.5	5.0	2.1	1.7	0.1	0.0	0.2
231	2.5	2.6	2.9	73.5	2.4	8.0	0.0	0.1	0.9	2.1	3.4	1.4	0.2
312	3.4	1.8	1.9	2.9	77.1	3.2	2.3	1.3	0.6	5.3	0.0	0.1	0.1
321	2.6	3.0	3.1	2.1	3.4	77.8	0.6	2.1	0.0	0.1	0.6	4.3	0.3
124	5.0	0.4	1.0	0.1	3.7	2.1	68.9	11.1	0.6	0.6	0.4	1.3	4.7
214	0.8	0.3	5.4	0.6	1.8	4.0	7.0	74.3	0.3	1.0	0.3	0.4	3.8
134	0.7	4.5	2.6	2.8	1.1	0.2	0.6	0.2	63.8	18.1	1.3	0.8	3.3
314	0.1	0.7	1.8	5.0	6.1	0.4	0.6	1.2	6.8	72.1	0.2	0.8	4.2
234	3.0	2.3	0.1	6.5	0.1	1.5	0.5	0.8	1.1	0.6	62.3	16.7	4.5
324	1.5	4.5	0.1	0.6	0.3	7.2	0.9	0.4	0.5	0.8	7.8	72.0	3.5
444	0.6	0.6	0.3	0.9	0.0	0.6	4.1	5.3	4.7	3.8	5.6	7.5	65.8

- The dominant classification is the correct class for all classes.
- For **triple**s, the second highest prediction is the corresponding **dtot**.
- For **dtot**, the second highest prediction is a reversed double ordering.
- For **false triples**, we see it is often mistaken as a **dtot**.
- **Triples** often become valid doubles which keeps the data usable.
- Reverse **dtot** orderings can be recovered using a network trained on **doubles**.
- The testing accuracy for each class is much closer to the average validation accuracy seen in our associated plot than the shorter study. Increasing dropout rate could bring them even closer.

Conclusions and Future Work

Conclusions

- Note that the fewer neurons the network has, the more computationally cheap it is to use, and the faster the network can classify records.
- We see that the usage of even a small amount of dropout has brought our confusion matrices' accuracies much closer to our validation accuracy seen in the training and validation plots.
- The triples now have comparable accuracy to a more complex network but the doubles-to-triples and false data are still 1% to 7% worse than the more complex network seen in previous works.
- Our networks fall short in classification accuracies, but if we can tackle the long training times then, with more hyperparameter tuning, we may be able to create a network that is easier to train and cheaper to use than previous networks.

Future Work

- Particular studies, if given considerably more training time, could yield competitive, if not superior, testing accuracy to existing architectures while maintaining a simpler structure.
- We are currently experimenting with recurrent neural networks to test the viability of this type of architecture for this application.