COVID-19 Risk Neural Network

Hrishikesh Bagalkote

Gwinnett School of Science, Mathematics, and Technology

COVID-19 Risk Neural Network

I. Purpose- The purpose of this project is to create a neural network that is trained on data encompassing various factors in large areas, such as states and cities which creates a finished machine learning model that outlines the impact and significance of each selected factor in the dataset on COVID-19 cases, transmissibility, and deaths. By optimizing a neural network to perform this task, the problem of not being able to make reasonable predictions about a smaller area, such as a district or portion of a county due to insufficient data is solved. The goal of the produced model is to be able to extrapolate on factors that are given in larger areas in order to determine the effect of each factor respectively and make reasonable predictions about COVID-19 related risk of smaller areas with less publicly available, usable data. If the correct inferences are drawn by the neural network and the output model scales to smaller areas properly, conclusions about COVID-19 concentration and risk in targeted areas such as school districts and neighborhoods can be drawn without the need for large amounts of processing power (CPUs and GPGPUs) and fast storage (Enterprise-grade NVME/PCIe SSDs).

A. Documentation

- Current COVID-19 forecasting methods entail extensive statistical modeling or complicated neural network training processes.
- Neural networks related to COVID-19 use image classification algorithms to diagnose COVID, not to forecast it.

- No software exists to predict the risk of exposure over time of a target area.
- Using publicly available statistics for states and cities and determining correlation to COVID-19 spread is more efficient than formatting existing datasets and executing time consuming and hardware demanding training cycles.

B. Research (paraphrased)

- In order to achieve nuanced answers to seemingly unsolvable issues relating to COVID-19, a large amount of public data must be available (Vestal, 2020).
- Even when a vaccine for COVID-19 emerges, usable public data must exist to track and measure its effectiveness (Vestal, 2020).
- Private corporations are collecting this valuable data instead of public institutions and organizations such as the CDC (Vestal, 2020).
- A large scale reporting system with easily accessible data about outbreaks must be put in place in order to lessen the effects of the next epidemic (Meyer, 2020).
- There is a bottleneck of data transfer between hospitals and public health agencies (Meyer, 2020).
- Different training methods for a COVID-19 forecasting neural network had their drawbacks, such as being too slow or inaccurate (Wieczorek et al., 2020, 2-4)

- It is a difficult and time-consuming process to format existing data for the purpose of training a neural network (Wieczorek et al., 2020, 2-4).

II. Hypothesis (engineering)- The most prevalent existing method of COVID-19 forecasting is statistical modeling. Based on current factors of a location, a rate of spread and increased risk is assumed over time. Making a neural network to analyze the correlation of these factors to COVID-19 risk by transmissibility, deaths, and confirmed cases is a superior method of modeling because it allows for the analysis of each respective factor in relation to other factors. Independent variables for this experiment are the number of epochs (cycles in the network) trained and the different factors trained on and incorporated into the model. The general dependent variable is the accuracy of the model's predictions, which can be measured by the percentage risk difference from the actual output value in the dataset. It is hypothesized that a greater amount of training epochs will result in a higher prediction accuracy of the network. Approaches to training can be varied by experimentation with the incorporation of different input factors. The fewer and more relevant the factors are to a specific area, the greater the accuracy of the predictions will be.

A. Documentation

- The learning rate of the model will be tweaked to have a balance between speed and precision of learning increment.
- By creating respective models according to population density and existing confirmed cases of an area, the correct model can be chosen for the right situation for various prediction scenarios.

- Similarly, creating different models trained on different features (input factors) is helpful for a situation in which a certain factor is not relevant to a target area.
- One potential method of including more input factors into the dataset is by using a mixed data training algorithm (qualitative, quantitative, and image data).

B. Research (paraphrased)

- Overfitting occurs when the network's learning rate is too high and does not recognize a trend in the data. Underfitting is when the network's learning rate is too low and does not correctly identify the trend in data quickly (Sharma, 2017).
- Creating a neural network / machine learning model that is capable of handling mixed data is extremely challenging but can be essential to certain use cases (Rosebrock, 2019).

III. Literature Review

A. Weight Correction and Back Propagation

- Traditional neural networks require an activation function, hypothesis function, and loss function (Al-Masri, 2019).
- The activation function determines the value at every node in the neural network and is what starts the training process.
- The hypothesis function is what predicts the input to the activation function (Al-Masri, 2019).
- The loss function scales the parameter values in the neural network. Forward propagation takes place with the existing weights to predict an output, and back propagation follows, correcting the weights according to the error.

- Differentiation is used to find an instantaneous rate of change of a function, which is also used to operate on concentration gradients in neural networks (Mayo, 2017).
- Weights are associated with neuron connections, and the error correction formula adjusts each weight according to the error margin. Weights are then fed back into the network for the next epoch (Mayo, 2017).

B. Unsupervised Machine Learning

- Parametric Unsupervised Learning is parametrized by mean and standard deviation and allows for the recognition of variability by the model (Shukla, 2018).
- Clustering deals with finding patterns and categorizing data.
- cluster can be defined as “similar values and dissimilar values to another cluster” (Shukla, 2018).
- Unsupervised machine learning allows for computers to recognize patterns that humans were not aware of previously (Shukla, 2018).
- Supervised machine learning has seen huge adoptions in recent years in applications such as computer vision (Metz et al., 2016, 1).
- For a task like computer vision, unsupervised machine learning is viable, as the neural network is allowed to understand training data and produce an output similar to that in the training/testing dataset.
- The GAN (generative adversarial network) was able to generate both bedrooms and human faces as outputs somewhat accurately without being supervised during the training process (Metz et al., 2016, 5-10).

C. Bias in Machine Learning and AI

- AI inherits the flaws of its underlying dependencies (Kilpatrick, 2019).
- If a dataset trends towards human biases, the model will train to those same biases (Kilpatrick, 2019).
- Representative datasets along with the right algorithm for the scope of the problem should be used to avoid bias in the model (Kilpatrick, 2019).
- The behavior of the neural network should be monitored and tested (Kilpatrick, 2019).
- Biases in datasets can unfairly advantage certain groups during machine learning training (Jiang & Nachum, 2020, 1).
- Data distribution must be considered when selecting data values to avoid inherited bias in the network (Jiang & Nachum, 2020, 2).

D. Hardware Costs of Neural Networks and Machine Learning Models

- By running segments of code together on high performance servers, debugging autonomy and ability is drastically reduced (Lipton, 2015).
- If one segment/part of code has a flaw, it affects the performance of other software components running on the same server (Lipton, 2015).
- On top of this, it is also often difficult to obtain usable data to train these high performance machine learning algorithms on.
- GPUs are more consistent and reliable for faster machine learning because of the speed at which they are able to perform matrix operations and floating point calculations (Baltazar, 2018).

- High performance GPGPU arrays are required to process this data, and high speed storage is required to keep it readily accessible.
- High performance general purpose ML GPUs are not at a stage of advancement or production yet to be cheaply available to the public.
- In the scope of the purpose of this project, one must have this high performance hardware readily available if it is desired to constantly monitor and forecast the apparent COVID-19 related risk of a targeted area.
- A more refactored neural network structure must be considered that can train for millions of epochs at a time on lower end hardware without relying on a GPU for optimized matrix operations.

E. Existing Methods of Disease Forecasting

- The CDC's methods of forecasting future COVID-19 cases and deaths in an area are based on statistical inferences and assumptions.
- Models make various assumptions about the levels of social distancing and other interventions, which may not reflect recent changes in behavior (CDC, 2020).
- Different universities researching COVID-19 and modeling the spread of the disease predict a percentage growth rate over an interval of time based on factors (such as levels of social distancing as stated before).
- This method of modeling is not entirely optimal as it takes longer to observe the trends in data and determine which factors were accurately weighted in the forecasting model.
- One use of mathematical modeling and forecasting of COVID-19 is determining which methods of intervention are most effective (Jewell et al., 2020, 1).

F. Contributing COVID-19 Factors

- Age, race, gender, pre-existing medical conditions, medications, poverty, occupation, and pregnancy are all severe risk factors for COVID-19 (CDC, 2020).
- COVID-19 is likely more severe to older adults, those with heart or lung disease, and diabetes (Maragakis, 2020).
- By analyzing these demographic factors, along with geographic, social, and economic factors quantitatively, a neural network can determine the relative risk related to COVID-19 in a target area, Along with this, the magnitude of each circumstance and factor on the computed risk can be analyzed and compared with one another.

Bibliography

Al-Masri, A. (2019, January 19). *How Does Back-Propagation in Artificial Neural Networks Work?* Towards Data Science. Retrieved November 12, 2020, from https://towardsdatascience.com/how-does-back-propagation-in-artificial-neural-networks-work-c7cad873ea7

Baltazar, G. (2018, September 13). *GPU vs CPU in Machine Learning*. Oracle. Retrieved November 13, 2020, from https://blogs.oracle.com/datascience/cpu-vs-gpu-in-machine-learning#:~:text=As%20a%20general%20rule%2C%20GPUs,be%20carried%20out%20in%20parallel.

CDC. (2020, August 10). *Assessing Risk Factors for Severe COVID-19 Illness*. CDC. Retrieved November 13, 2020, from https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/assessing-risk-factors.html

CDC. (2020, November 12). *COVID-19 Forecasts: Deaths*. CDC. Retrieved November 13, 2020, from https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html

Jewell, N. P., Lewnard, J. A., & Jewell, B. L. (2020, April 16). Predictive Mathematical Models of the COVID-19 Pandemic Underlying Principles and Value of Projections. *JAMA*, *323*(19), 2. doi:10.1001/jama.2020.6585

Jiang, H., & Nachum, O. (2020). Identifying and Correcting Label Bias in Machine Learning. *Proceedings of Machine Learning Research*, *108*, 10. http://proceedings.mlr.press/v108/jiang20a.html

Kilpatrick, S. (2019, January 22). *How to Prevent Machine Bias in AI*. Logikk. Retrieved November 13, 2020, from https://www.logikk.com/articles/prevent-machine-bias-in-ai/

Lipton, Z. C. (2015, January). *The High Cost of Maintaining Machine Learning Systems*. KDNuggets. Retrieved November 13, 2020, from https://www.kdnuggets.com/2015/01/high-cost-machine-learning-technical-debt.html

Maragakis, L. (2020, June 25). *Coronavirus and COVID-19: Who is at higher risk?* Hopkins Medicine. Retrieved November 13, 2020, from https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/coronavirus-and-covid19-who-is-at-higher-risk

Mayo, M. (2017, October). *Neural Network Foundations, Explained: Updating Weights with Gradient Descent & Backpropagation*. KDNuggets. Retrieved November 12, 2020, from https://www.kdnuggets.com/2017/10/neural-network-foundations-explained-gradient-descent.html

Metz, L., Radford, A., & Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv*, 16. https://arxiv.org/abs/1511.06434

Meyer, H. (2020, August 14). *COVID-19 data often slowed by paper forms and fax machines, thanks to creaky US public health data system*. USA Today. Retrieved November 12, 2020, from https://www.usatoday.com/story/news/health/2020/08/14/covid-19-data-slowed-us-lack-national-health-data-network/3369859001/

Rosebrock, A. (2019, February 4). *Keras: Multiple Inputs and Mixed Data*. Py Image Search. Retrieved November 13, 2020, from https://www.pyimagesearch.com/2019/02/04/keras-multiple-inputs-and-mixed-data/

Sharma, S. (2017, September 23). *Epoch vs Batch Size vs Iterations*. Towards Data Science. Retrieved November 13, 2020, from https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9

Shukla, S. (2018, June 12). *How Does Unsupervised Machine Learning Work?* Upgrad. Retrieved November 12, 2020, from https://www.upgrad.com/blog/how-does-unsupervised-machine-learning-work/

Vestal, C. (2020, August 3). *Lack of Public Data Hampers COVID-19 Fight*. Pew. Retrieved November 12, 2020, from https://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2020/08/03/lack-of-public-data-hampers-covid-19-fight

Wieczorek, M., Silka, J., & Wozniak, M. (2020, November). Neural network powered COVID-19 spread forecasting model. *Science Direct*, *140*(Chaos, Solitons, and Fractals), 5. https://doi.org/10.1016/j.chaos.2020.110203