An Analysis of “Deep learning Methods for Forecasting COVID-19 Time-Series Data: A Comparative Study”

Gregory Peterson
pet03356@morris.umn.edu

Outline

- Introduce Paper
- Neural Networks
    - Variational Autoencoder (VAE)
    - Recurrent Neural Networks (RNN)
- Performance Metrics
- Results Analysis
- Conclusion
- Q&A

Deep learning Methods for Forecasting COVID-19 Time-Series Data: A Comparative Study

Abdelhafid Zeroual, Fouzi Harrou, Abdelkader Dairi, and Ying Sun

Publisher: Elsevier

Journal: Chaos, Solitons, and Fractals

Special COVID-19 issue, November 2020.

Citations: 459

To my knowledge results table contains errors.

The specifics about their “Deep Learning Methods” were not described.

Testing the Accuracy of Neural Networks at COVID-19 Forecasting.

Variational Autoencoder (VAE)

Recurrent Neural Network (RNN)

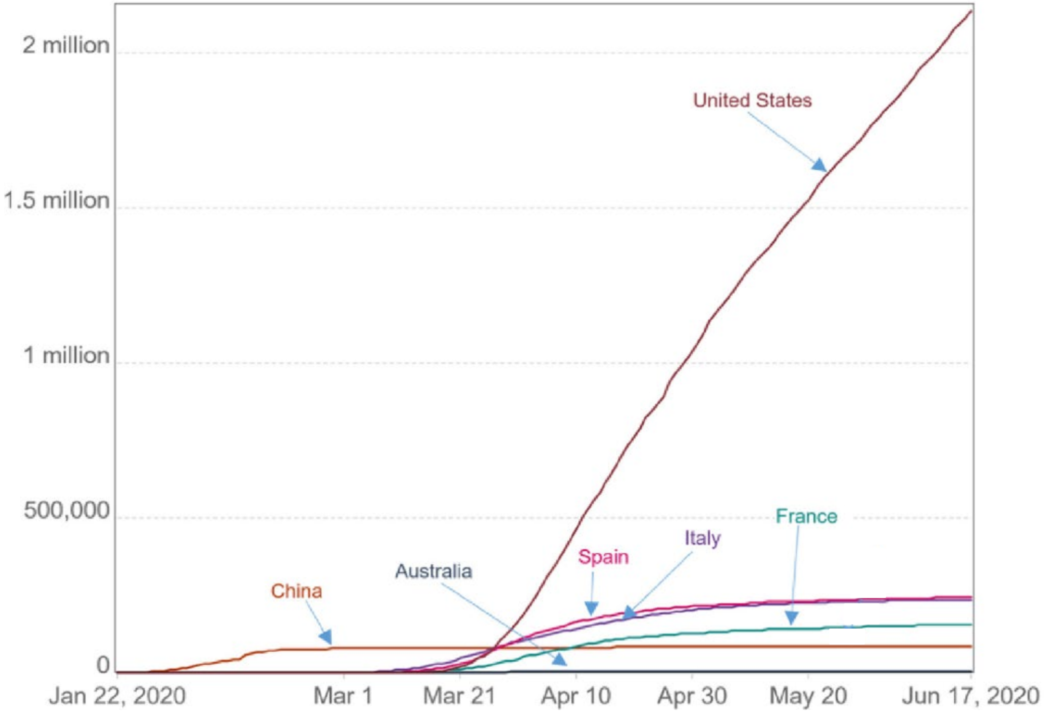
Gated Recurrent Unit (GRU)

Long Short-Term Memory (LSTM)

Bidirectional LSTM (BiLSTM)

Cumulative Number of COVID-19 Cases at Each Date

- Forecasts for six countries.
- A perfect forecast would match a countries curve exactly.
- January 22nd - June 1st, 2020 used to train (131 days).
- June 1st-17th, 2020 used to test (17 days).

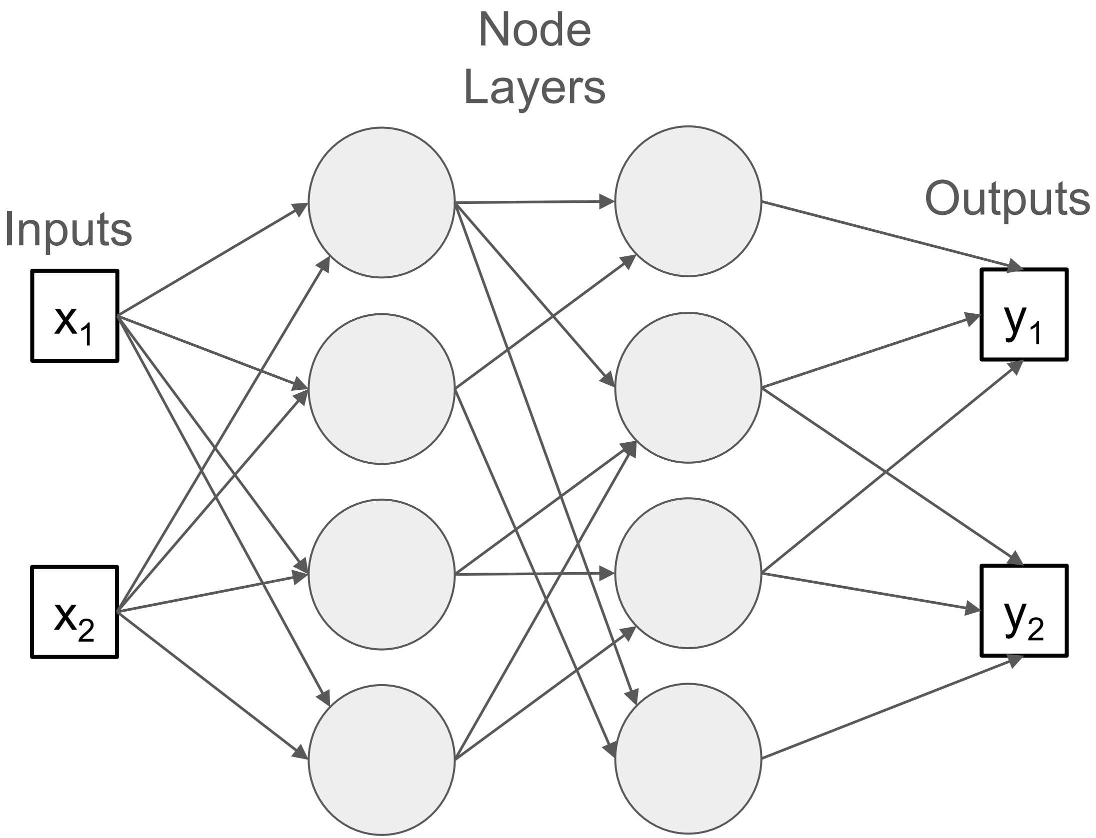


Dataset from Johns Hopkins University.

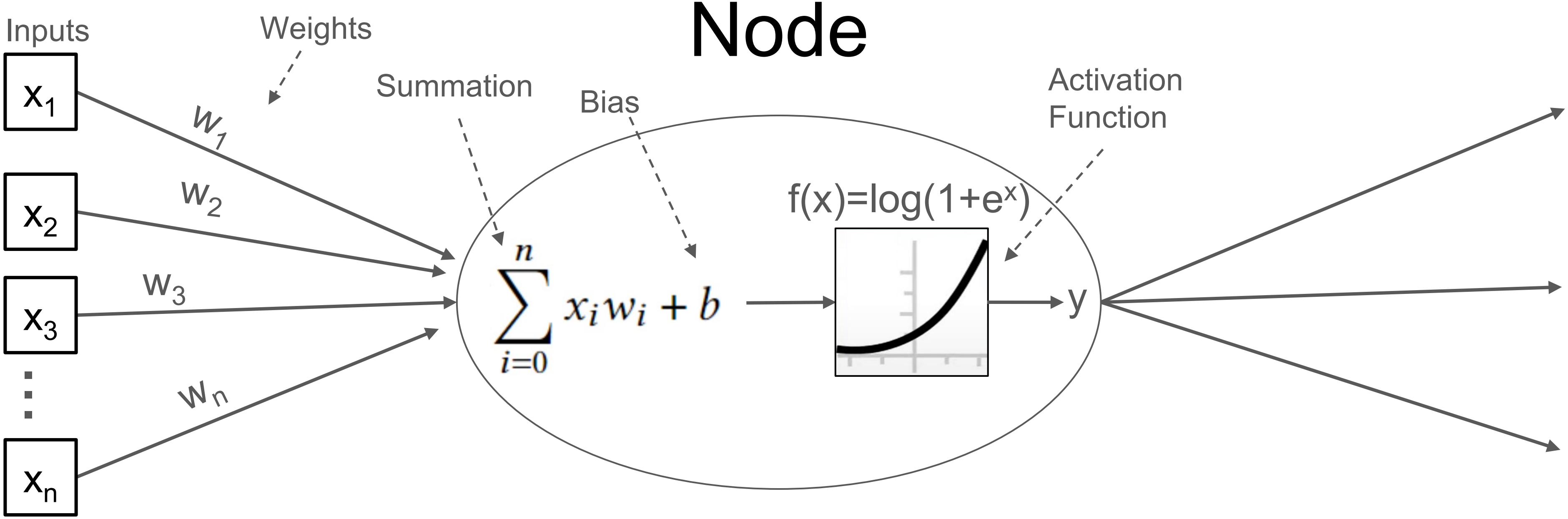
Neural Networks

The authors did not describe their data preprocessing.

We can conceptualize each input as the number of cases that day.



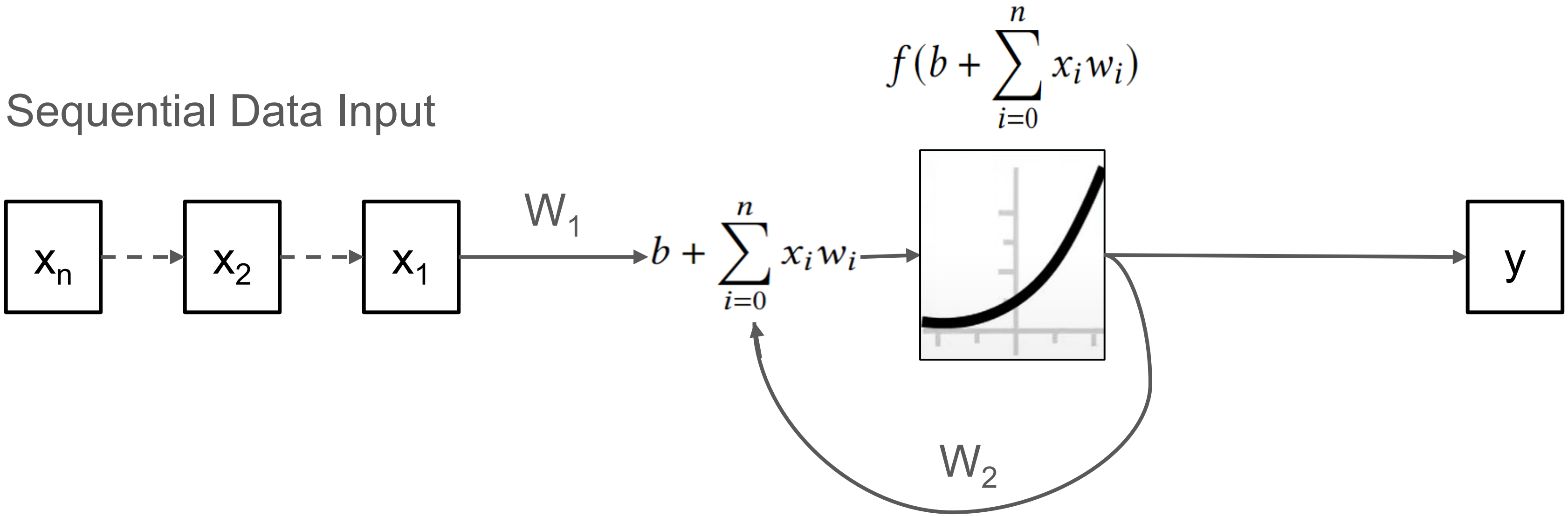
Neural Network Node



How are weights and biases decided?

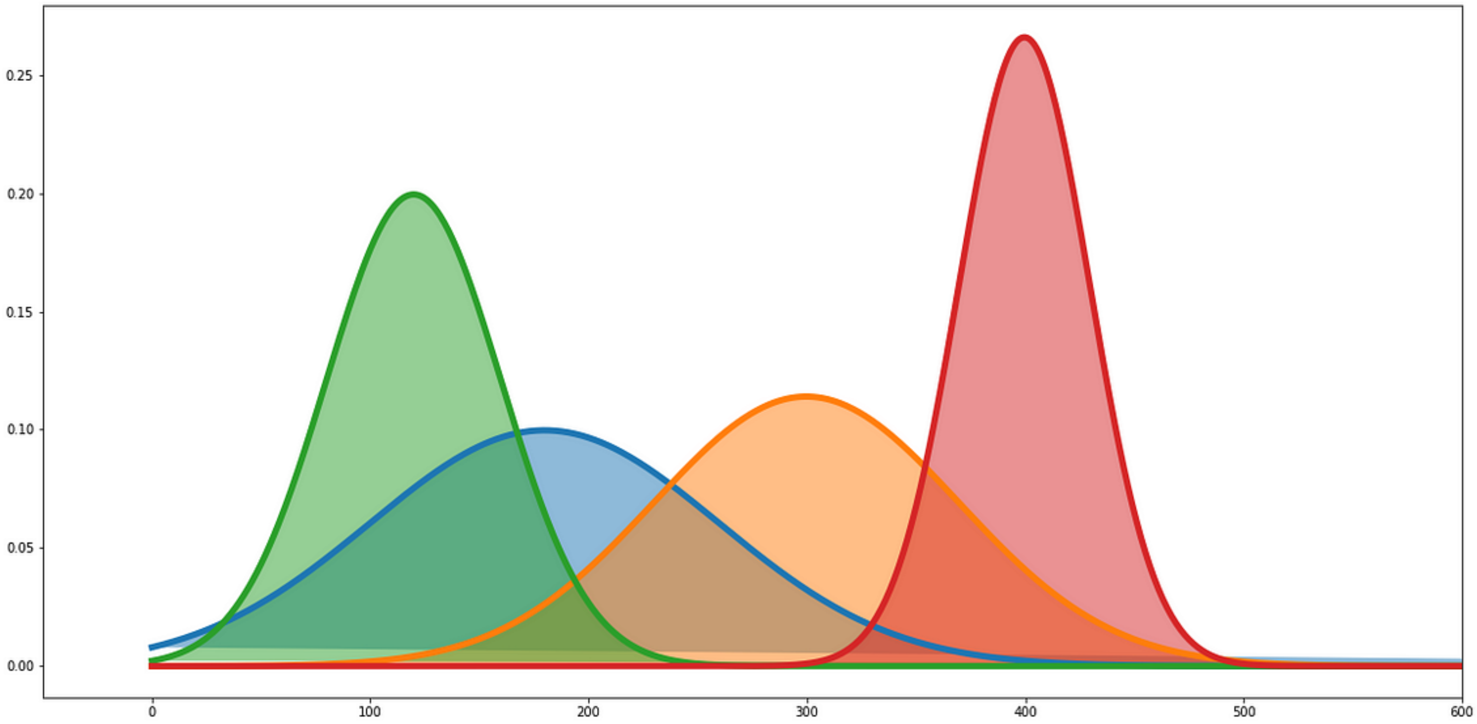
Inspiration

Recurrent Neural Network Node



Inspiration

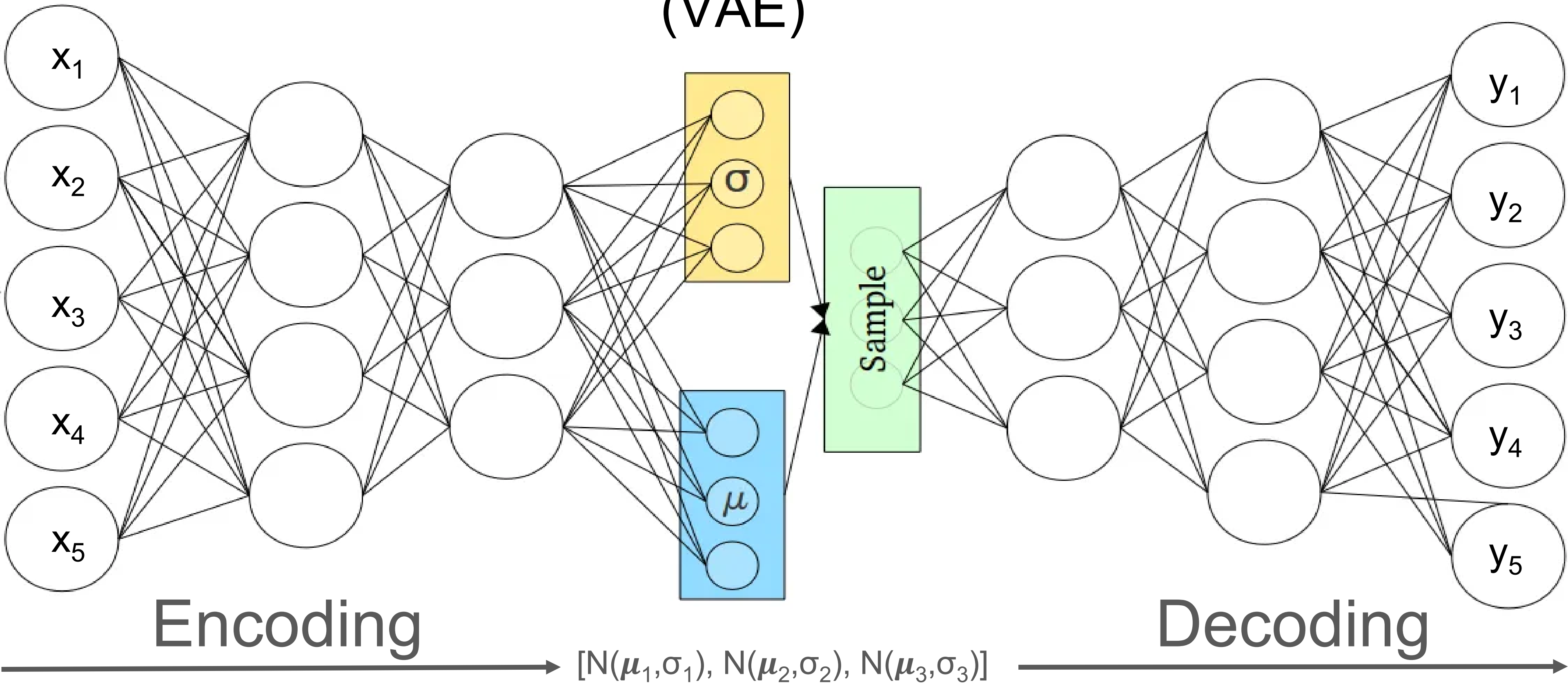
Normal Distribution



A normal distribution describes likelihood of a particular value occuring.
Standard deviation $\boldsymbol{\sigma}$ and mean μ are the parameters of a normal distribution.

Image Credit

Variational Autoencoder (VAE)



$[N(\mu_1,\sigma_1), N(\mu_2,\sigma_2), N(\mu_3,\sigma_3)]$

Credit: [Irhum Shafkat. 2018. Intuitively Understanding Variational Autoencoders.]

Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Absolute Percentage Error

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

y = true values
ŷ = forecasted values

RMSE and MAE:

- Average difference between true and forecasted values.
- The closer a score is to zero the more accurate the forecast.
- RMSE is by definition greater than or equal to MAE.

MAPE:

- Average difference as a percentage of each true value.
- The closer a score is to 0% the more accurate the forecast.

Root Mean Squared Logarithmic Error

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log(y+1) - \log(\hat{y}+1)\right)^2}$$

Explained Variance

$$EV = 1 - \frac{\mathrm{Var}(y - \hat{y})}{\mathrm{Var}(y)}$$

y = true values
ŷ = forecasted values

RMSLE:

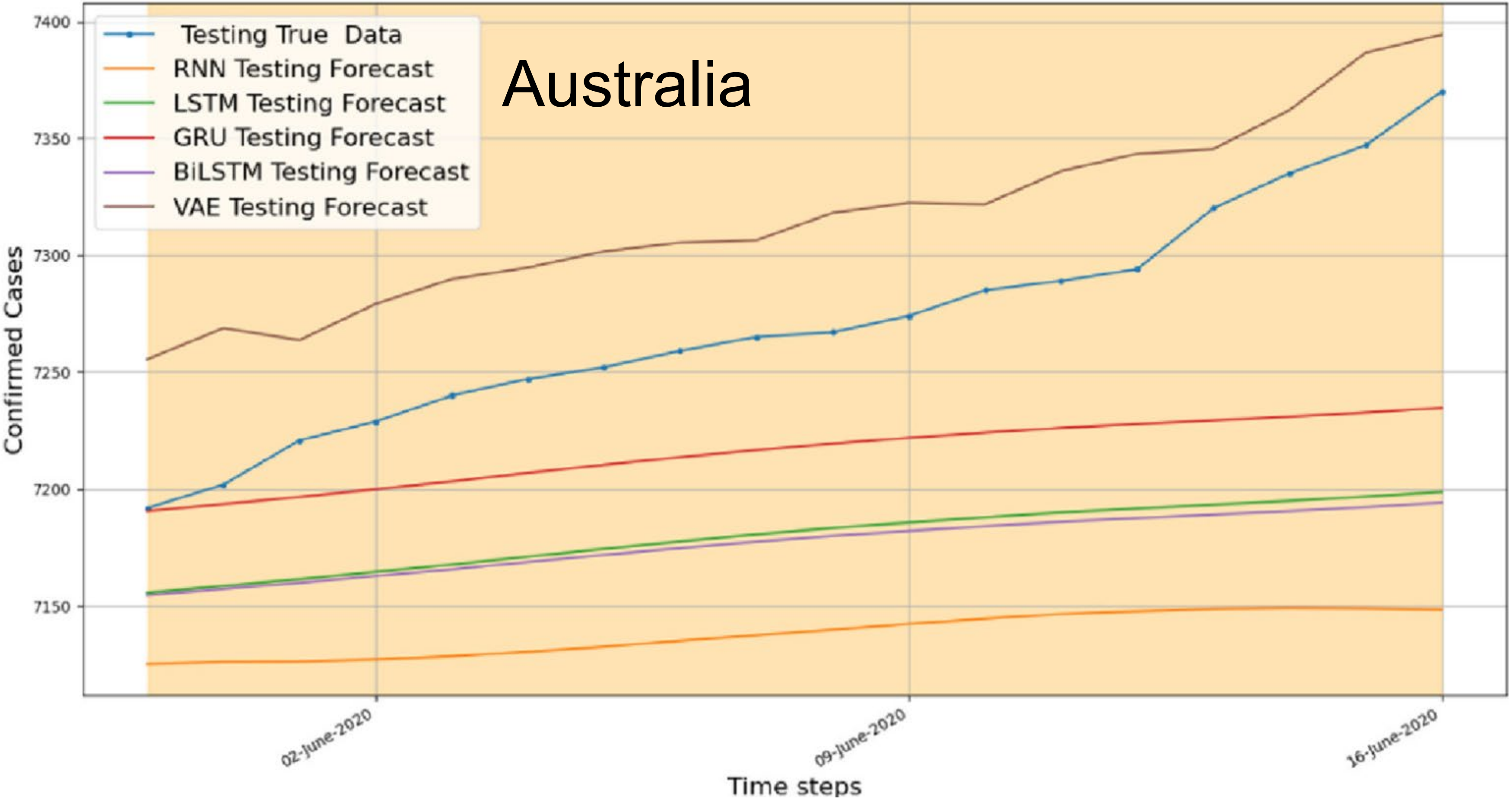
- Average difference between true and forecasted values on a logarithmic scale.
- Penalizes underestimations more.
- The closer a score is to zero the more accurate the forecast.

EV:

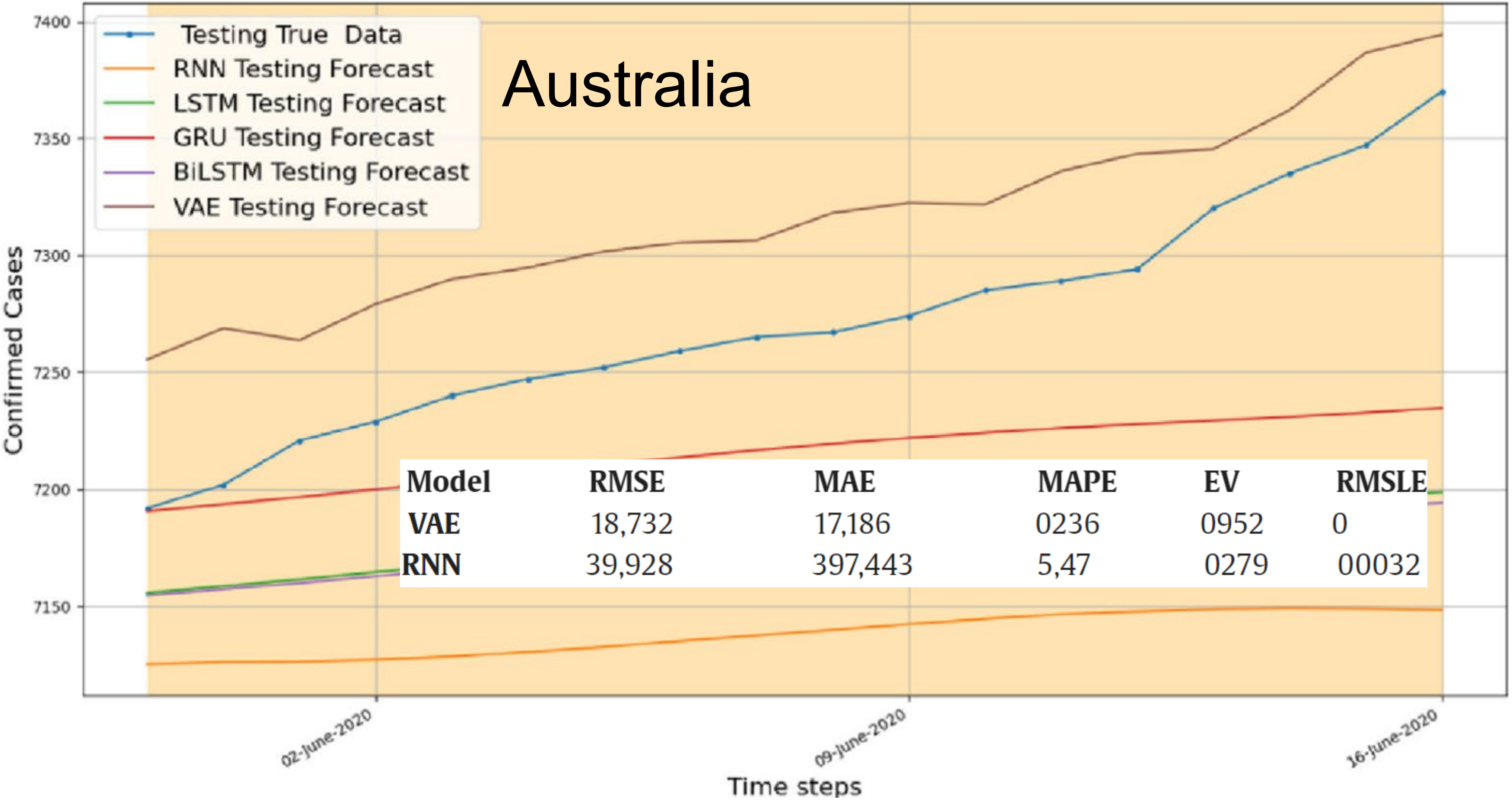
- Forecasted values that are consistently the same distance from the true values will have an EV closer to one.
- A good EV can be visually seen in a forecasted curve consistently matching the slope of the true curve.

“It can be easily seen that the VAE model outperformed the other models by providing good forecasting performance with lower RMSE, MAE, MAPE and RMSLE, and EV values closer to 1.”

- Zeroual et al.



Adapted from Zeroual et al.



Model	RMSE	MAE	MAPE	EV	RMSLE
VAE	18,732	17,186	0236	0952	0
RNN	39,928	397,443	5,47	0279	00032

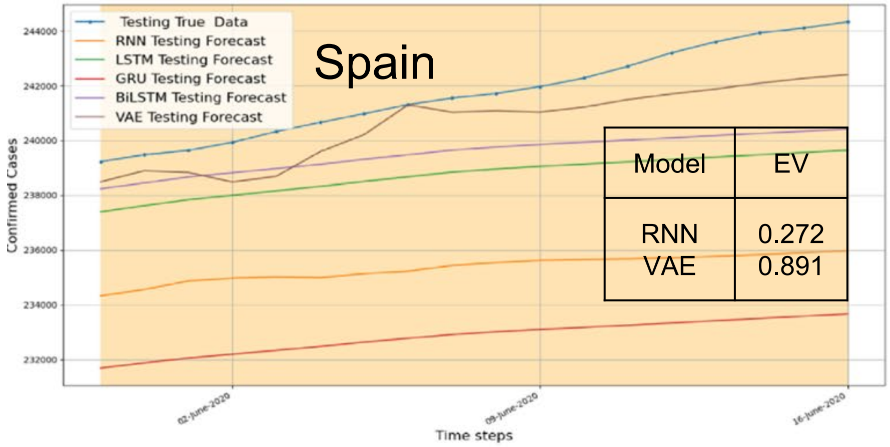
Adapted from Zeroual et al.

Country	Model	RMSE	MAE	MAPE	EV	RMSLE
Italy	RNN VAE	1,070,474 1,386,225	1,062,061 1,385,829	4519 5901	0201 0951	00022 00033
Spain	RNN VAE	1,683,011 5,315,748	167,719 5,288,172	6944 2,19	0272 0891	00052 00005
France	RNN VAE	1,287,786 3,688,083	1,279,681 3,522,353	6827 1,88	0224 0554	00051 00004
China	RNN VAE	1,252,034 11,103	1,250,442 107,873	1485 0128	0095 0843	00002 0
Australia	RNN VAE	39,928 18,732	397,443 17,186	5,47 0236	0279 0952	00032 0
USA	RNN VAE	5,227,287 4,079,244	5,136,497 3,976,682	26,373 2,04	0208 0993	00967 00004

MAPE, EV, and RMSLE missing commas.

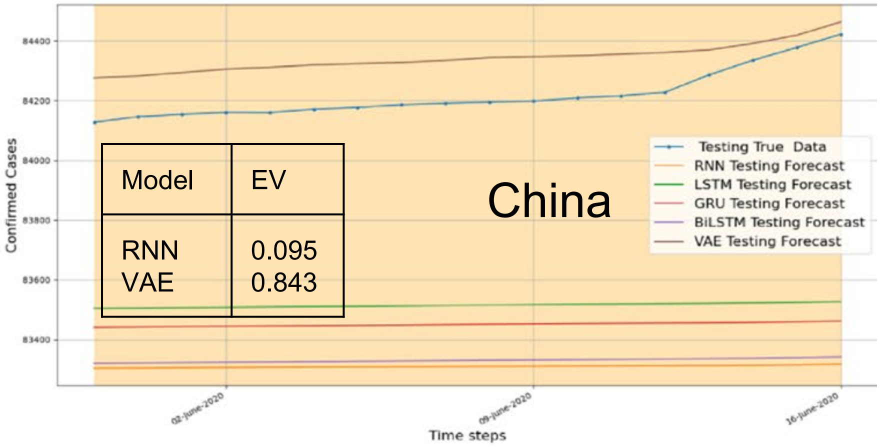
RMSE and MAE scores suggest on average the forecast is off by a million or more

Performance scores for forecasted cases unedited from Zeroual et al (other models excluded).



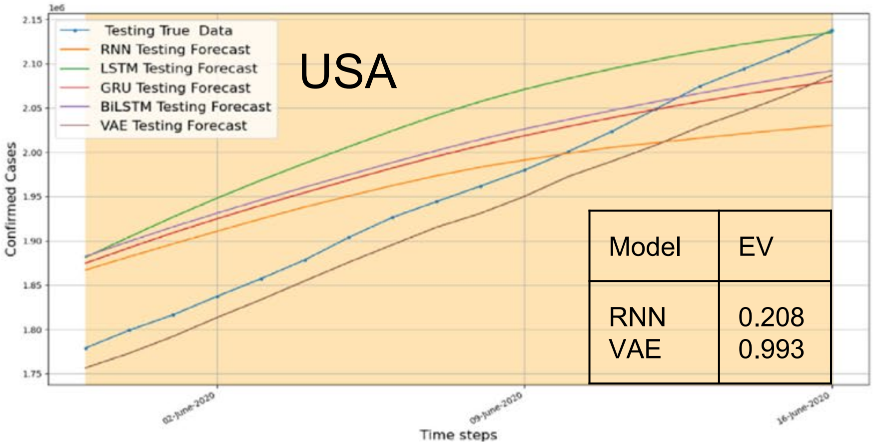
Model	EV
RNN	0.272
VAE	0.891

On average off by thousands of cases.



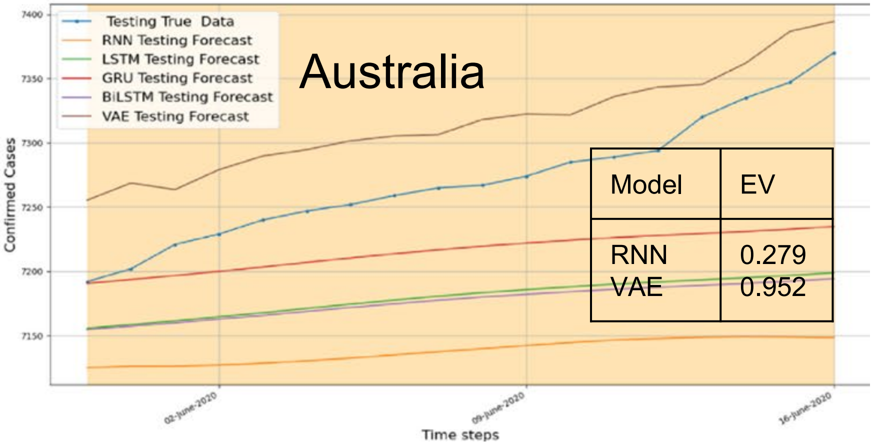
Model	EV
RNN	0.095
VAE	0.843

On average off by hundreds of cases.



Model	EV
RNN	0.208
VAE	0.993

On average off by tens or hundreds of thousands of cases.



Model	EV
RNN	0.279
VAE	0.952

On average off by tens or hundreds of cases.

Adapted from Zeroual et al. Different order from performance metric table.

Conclusion

- An accurate forecast could assist organizing the logistics of fighting COVID-19.
- Neural networks were trained on 131 days of data and tested with a 17 day forecast.
- Unable to talk about specifics of their implementations.
- Performance metrics table, to my knowledge, is problematic. VAE appears to have performed the best based on graphs.

Q&A

References

Josh Starmer. 2020. Neural Networks Pt. 1: Inside the Black Box
https://www.youtube.com/watch?v=CqOfi41LfDw&t=847s

Irhum Shafkat. 2018. Intuitively Understanding Variational Autoencoders.
https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

Josh Starmer. 2022. Recurrent Neural Networks (RNNs), Clearly Explained!!!
https://www.youtube.com/watch?v=AsNTP8Kwu80&t=40s

Abdelhafid Zeroual, Fouzi Harrou, Abdelkader Dairi, and Ying Sun. 2020. Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. Chaos, Solitons & Fractals 140 (2020). https://doi.org/10.1016/j.chaos.2020.110121

Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

y = actual values
ŷ = predicted values

y = [7190, 7230, 7240, 7250]

ŷ=[7260, 7270, 7290, 7300]

Squared differences:

(7260 - 7190)^2 = 4900

(7270 - 7230)^2 = 1600

(7290 - 7240)^2 = 2500

(7300 - 7250)^2 = 2500

Mean of the squared differences:

(4900 + 1600 + 2500 + 2500) / 4 = 2850

Root of the mean squared difference:

RMSE = √(2850) ≈ 53.48

Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

y = actual values
ŷ = predicted values

y = [7190, 7230, 7240, 7250]
ŷ=[7260, 7270, 7290, 7300]

Absolute Differences:
|7260 - 7190| = 70
|7270 - 7230| = 40
|7290 - 7240| = 50
|7300 - 7250| = 50

Mean of absolute differences:
MAE = (70 + 40 + 50 + 50) / 4 = 52.5

Mean Absolute Percentage Error

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

y = actual values
ŷ = predicted values

y = [7190, 7230, 7240, 7250]
ŷ=[7260, 7270, 7290, 7300]

Absolute percentage differences (rounded)
|(7260 - 7190) / 7190| = 0.0097
|(7270 - 7230) / 7230| = 0.0055
|(7290 - 7240) / 7240| = 0.0069
|(7300 - 7250) / 7250| = 0.0069

Mean of absolute percentage differences
(0.0097 + 0.0055 + 0.0069 + 0.0069) / 4 = 0.0075 (rounded)

Converted to a percentage
MAPE = 0.0075 * 100 = 0.75%

Root Mean Squared Logarithmic Error

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(y_i+1)-\log(\hat{y}_i+1))^2}$$

y = actual values
ŷ = predicted values

y = [7190, 7230, 7240, 7250]
ŷ=[7260, 7270, 7290, 7300]

ln(7190) ≈ 8.880
ln(7230) ≈ 8.887
ln(7240) ≈ 8.889
ln(7250) ≈ 8.890

For the predicted values:
ln(7260) ≈ 8.889
ln(7270) ≈ 8.890
ln(7290) ≈ 8.894
ln(7300) ≈ 8.896

Squared differences between the natural logarithms
(8.889 - 8.880)^2 ≈ 0.000081
(8.890 - 8.887)^2 ≈ 0.000009
(8.894 - 8.889)^2 ≈ 0.000025
(8.896 - 8.890)^2 ≈ 0.000036

Mean of the squared differences:
(0.000081 + 0.000009 + 0.000025 + 0.000036) / 4 ≈ 0.00003775

Square root of the mean squared difference
RMSLE ≈ √(0.00003775) ≈ 0.00615 (rounded)

Explained Variance

$$EV = 1 - \frac{\mathrm{Var}(y - \hat{y})}{\mathrm{Var}(y)}$$

y = actual values
ŷ = predicted values

y = [7190, 7230, 7240, 7250]
ŷ=[7260, 7270, 7290, 7300]

7260 - 7190 = 70
7270 - 7230 = 40
7290 - 7240 = 50
7300 - 7250 = 50