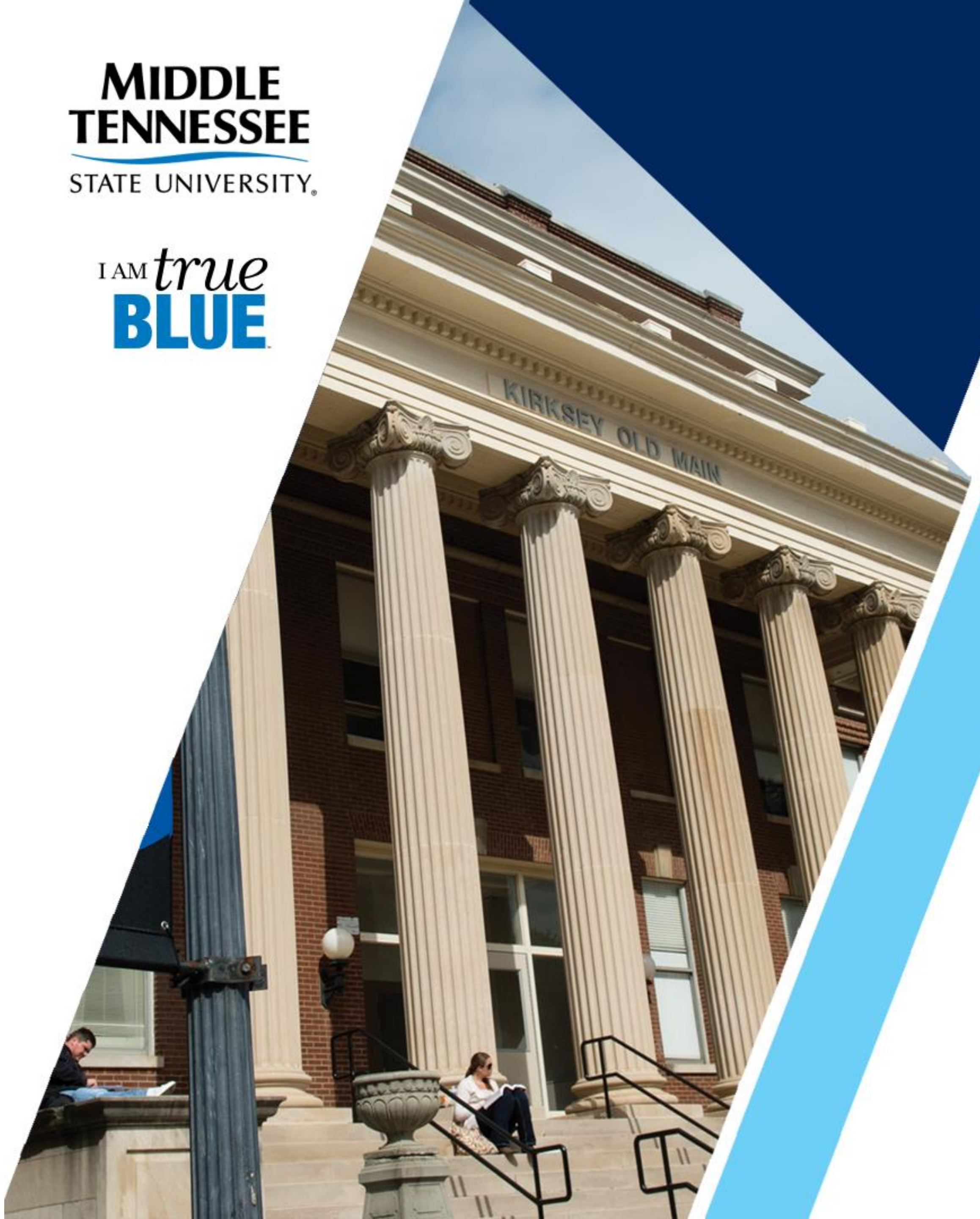

Robust Representations in Deep Learning

Authors: Shu Liu, Qiang Wu

Computational and Data Science, Middle Tennessee State University

Contact email: sl6b@mtmail.mtsu.edu

Shu Liu

- Doctoral student of Computational and Data Science program.

Publications:

- Liu, S., & Wu, Q. (2021, December). Pairwise Learning for Imbalanced Data Classification. In *2021 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 186-189). IEEE.

Robust deep neural networks

We introduce two robust deep neural networks, the robust deep feedforward neural network and the robust long short-term memory neural network.

- Correntropy loss

$$\mathcal{L}(y_i, \hat{y}_i) = \sigma^2 \left(1 - \exp\left(\frac{(y_i - \hat{y}_i)^2}{\sigma^2}\right)\right)$$

- Robust Deep Forward Neural Network

We implement the robust deep forward neural network (RFNN) by minimizing the mean correntropy loss

$$\min_f \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i)$$

MIDDLE TENNESSEE STATE UNIVERSITY

I AM true BLUE

Robust deep neural networks

- Robust Long Short-Term Memory Neural Network

Given a sequence of time series $x_1, x_2, \ldots, x_T$ and corresponding response series $y_1, y_2, \ldots, y_T$, the robust LSTM will minimize the mean correntropy loss

$$\frac{1}{T}\sum_{i=1}^{T}\mathcal{L}(y_t, \hat{y}_t)$$

over the historical period to estimate the network parameters.

Two Stage Algorithms

We extract the features and run the linear regression with either least square (LS) approach or the robust regression (RR) with correntropy loss. This leads to four two-stage algorithms:

- FNN+LS
- FNN+RR
- RFNN+LS
- RFNN+RR

where the FNN+LS uses the features extracted from FNN and least square regression to predict the response variable, FNN+RR uses the features extracted from FNN and robust regression, RFNN+LS uses the features extracted from RFNN and least square regression, and RFNN+LS uses the features extracted from RFNN and robust regression.

Applications

We apply our algorithms to real-world applications and illustrate their effectiveness. There are four data sets are used in two stage RFNN and one data set is used in two stage RLSTM.

TABLE I
MAE ON AIRFOIL AND BOSTON HOUSING DATA

Method	Airfoil	Boston Housing	Agroecosystem
FNN	0.2366 (0.0023)	0.2761 (0.0045)	0.1877 (0.0026)
FNN+LS	0.2235 (0.0016)	0.2719 (0.0040)	0.1720 (0.0007)
FNN+RR	0.2223 (0.0016)	0.2702 (0.0040)	0.1719 (0.0007)
RFNN	0.2279 (0.0018)	0.2706 (0.0035)	0.1779 (0.0014)
RFNN+LS	0.2173 (0.0014)	0.2681 (0.0035)	0.1714 (0.0009)
RFNN+RR	0.2161 (0.0014)	0.2669 (0.0035)	0.1713 (0.0008)

TABLE II
MAE ON AIU AND CSI300 DATA

Method	CSI300
LSTM	0.2221 (0.0007)
LSTM+LS	0.2133 (0.0007)
LSTM+RR	0.2016 (0.0006)
RLSTM	0.2197 (0.0008)
RLSTM+LS	0.2116 (0.0007)
RLSTM+RR	0.2012 (0.0007)

Conclusions and future works

Conclusions

- We proposed to implement robust deep neural networks by using the correntropy loss and four two-stage algorithms.
- Simulation studies on four real data applications show that the robust deep neural networks are more efficient to handle data with outliers or skewed.
- The robust deep neural networks are able to efficiently extract more informative features, indicating the entropy loss plays more roles in robust representation of the data.
- The superiority of two-stage algorithms is a serendipity. The simulations surprisingly show that all two-stage algorithms are consistently better than their one-stage counterparts, regardless the loss function used.

Future Works

- We omitted the study of robust Convolutional Neural Network in this paper. But the idea of two-stage training is promising and it would be interesting to develop two-stage CNN algorithms with appropriate classification loss functions, such as the cross entropy loss.

References

[1] F. Rosenblatt, “The perceptron: a probabilistic model for information storage and organization in the brain.” Psychological review, vol. 65, no. 6, p. 386, 1958.
[2] G. Cybenko, “Approximation by superpositions of a sigmoidal function,” Mathematics of control, signals and systems, vol. 2, no. 4, pp. 303–314, 1989.
[3] K.-I. Funahashi, “On the approximate realization of continuous mappings by neural networks,” Neural networks, vol. 2, no. 3, pp. 183–192, 1989.
[4] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” Neural networks, vol. 2, no. 5, pp. 359–366, 1989.
[5] J. Schmidhuber, “Deep learning in neural networks: An overview,” Neural networks, vol. 61, pp. 85–117, 2015.
[6] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” nature, vol. 521, no. 7553, pp. 436–444, 2015.
[7] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT press Cambridge, 2016, http://www.deeplearningbook.org.
[8] B. Chen, L. Xing, H. Zhao, N. Zheng, and J. C. Principe, “Generalized correntropy for robust adaptive filtering,” IEEE Transactions on Signal Processing, vol. 64, no. 13, pp. 3376–3387, 2016.
[9] I. Goodfellow, P. McDaniel, and N. Papernot, “Making machine learning robust against adversarial inputs,” Communications of the ACM, vol. 61, no. 7, pp. 56–66, 2018.
[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in International Conference on Learning Representations, 2018. [Online]. Available: https://openreview.net/forum?id=rJzIBfZAb

References

[11] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” Communications of the ACM, vol. 24, no. 6, pp. 381–395, 1981.
[12] L. Yin, Q. Wu, and D. Hong, “Statistical methods and software package for medical trend analysis in health rate review process,” J Health Med Inform, vol. 7, no. 219, 2016.
[13] J. E. Dennis Jr and R. E. Welsch, “Techniques for nonlinear least squares and robust regression,” Communications in Statistics-Simulation and Computation, vol. 7, no. 4, pp. 345–359, 1978.
[14] K. P. Kording and D. M. Wolpert, “The loss function of sensorimotor ¨ learning,” Proceedings of the National Academy of Sciences, vol. 101, no. 26, pp. 9839–9842, 2004.
[15] X. Wang, Y. Jiang, M. Huang, and H. Zhang, “Robust variable selection with exponential squared loss,” Journal of the American Statistical Association, vol. 108, no. 502, pp. 632–643, 2013.
[16] F. A. Spiring, “The reflected normal loss function,” Canadian Journal of Statistics, vol. 21, no. 3, pp. 321–330, 1993.
[17] W. Liu, P. P. Pokharel, and J. C. Principe, “Correntropy: properties and applications in non-Gaussian signal processing,” IEEE Transactions on Signal Processing, vol. 55, no. 11, pp. 5286–5298, 2007.
[18] J. C. Principe, Information Theoretic Learning: Renyi’s Entropy and Kernel Perspectives. Springer Science & Business Media, 2010.
[19] Y. Feng, X. Huang, L. Shi, Y. Yang, J. A. Suykens et al., “Learning with the maximum correntropy criterion induced losses for regression.” J. Mach. Learn. Res., vol. 16, no. 30, pp. 993–1034, 2015.

References

[20] Y. Feng, J. Fan, and J. A. Suykens, “A statistical learning approach to modal regression,” Journal of Machine Learning Research, vol. 21, no. 2, pp. 1–35, 2020.
[21] Y. Feng and Y. Ying, “Learning with correntropy-induced losses for regression with mixture of symmetric stable noise,” Applied and Computational Harmonic Analysis, vol. 48, no. 2, pp. 795–810, 2020.
[22] Y. Feng and Q. Wu, “Learning under $(1 + \epsilon)$-moment conditions,” Applied and Computational Harmonic Analysis, vol. 49, no. 2, pp. 495– 520, 2020.
[23] ——, “A framework of learning through empirical gain maximization,” Neural Computation, vol. 33, no. 6, pp. 1656–1697, 2021.
[24] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[25] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: Continual prediction with lstm,” Neural computation, vol. 12, no. 10, pp. 2451–2471, 2000.
[26] D. Dua and C. Graff, “UCI machine learning repository,” 2017. [Online]. Available: http://archive.ics.uci.edu/ml