Knowledge distillation for anomaly detection in multivariate time series data*

Seonyoung Kim, Myoungho Kim
School of Computing, Korea Advanced Institute of Science and Technology
{sykim, mhkim}@dbserver.kaist.ac.kr

Abstract

Knowledge distillation (KD) is a technique that trains a compact model while maintaining comparable performance to a larger, pre-trained model. In this paper, we propose a knowledge distillation method for anomaly detection in multivariate time-series data, employing a mean squared error (MSE)–based loss function. Experiments on real-world datasets demonstrate that the distilled model achieves comparable anomaly detection performance with the teacher model, while requiring approximately 60% fewer parameters.

1. Introduction

With the recent proliferation of diverse sensors in various industrial sectors, such as smart factories and IoT environments, a vast amount of multivariate time series data is being generated in real-time. Anomaly detection is a critical technology that analyzes this multivariate time series data to identify abnormal situations, such as equipment failures or hazardous conditions, which do not conform to the system's normal operational patterns. This capability allows for the early detection of potential risks within a system. While early work relied on data-mining techniques, recent studies have increasingly focused on leveraging deep learning models, such as Long Short-Term Memory (LSTM) networks, which have garnered significant attention for their superior performance.

However, a challenge arises when employing deep learning models for anomaly detection. To enhance performance, these models often feature a deeper architecture with a greater number of layers. This structural complexity leads to a significant increase in inference time—the time required for the model to produce an output for a given input. This prolonged inference time poses a practical problem for systems that need to analyze multivariate time series data in real-time. Specifically, for systems with a high data generation frequency, the extended response time of the anomaly detection model makes it difficult to provide real-time results, thereby limiting the applicability of pre-trained models in such environments.

To address the aforementioned challenges, this paper proposes the application of knowledge distillation to anomaly detection models. Knowledge distillation is a technique that trains a compact "student" model by leveraging a pre-trained, larger deep learning model, referred to as the "teacher" model. The objective is for the student model to achieve detection performance comparable to the teacher model while having significantly fewer computational operations on its weights. For our anomaly detection framework, we utilize an LSTM-based autoencoder model for multivariate time series data [1]. Diverging from conventional knowledge distillation methods in image classification that typically employ a cross-entropy loss function, this paper proposes a novel approach that utilizes a Mean Squared Error (MSE) function to train the student model.

The remainder of this paper is organized as follows. Chapter 2 provides an overview of the anomaly detection method using the LSTM autoencoder model [1] and the principles of knowledge distillation. Chapter 3 introduces our proposed knowledge distillation technique for the LSTM autoencoder model, which leverages the Mean Squared Error function. In Chapter 4, we present experimental results demonstrating that our proposed method effectively reduces the number of parameters, a key indicator of model computational cost, while maintaining comparable performance. Finally, Chapter 5 concludes the paper and discusses future research directions.

2. Background

2.1 Anomaly Detection using an LSTM Autoencoder Model

The LSTM autoencoder, as a sequence-to-sequence model, was proposed for sequence learning applications such as chatbots and machine translation [2]. In [1], an LSTM autoencoder model is utilized to detect anomalies in a multivariate time series $X = \{x_1, x_2, \ldots, x_L\}$ of length L. The LSTM autoencoder is composed of an encoder, which maps an input sequence to a fixed-length vector representation, and a decoder, which reconstructs the input sequence from that vector. The LSTM autoencoder is trained to minimize the difference between the input sequence and the reconstructed sequence.

*) Acknowledgments: This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2020R1A2C1004032) and by the Bio-Synergy Research Project (2013M3A9C4078137) funded by the MSIT and NRF.

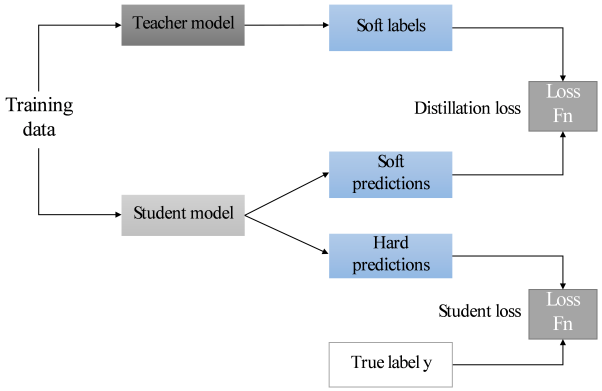

Figure 1: Method for training student models via knowledge distillation.

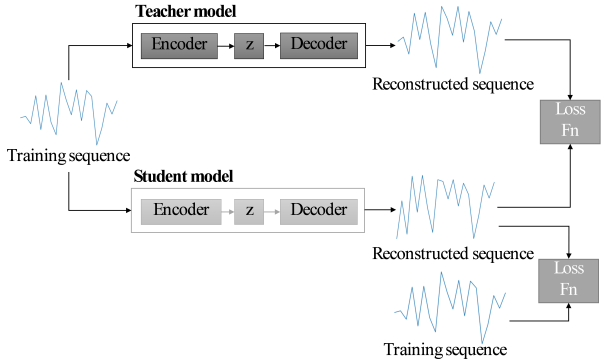

Figure 2: Proposed knowledge distillation method for LSTM autoencoders.

After training is completed, an anomaly score is calculated for an input sequence to detect anomalies. The formula for the anomaly score for an individual point within the sequence is as follows:

$$a_t = (e_t - \mu)^T \Sigma^{-1} (e_t - \mu) \tag{1}$$

Here, $e_t = \|x_t - x'_t\|$ signifies the reconstruction error between the input point x_t and the reconstructed point x'_t. The terms μ and Σ represent the mean vector and the covariance matrix, respectively, of a normal distribution $\mathcal{N}(\mu, \Sigma)$ fitted to the reconstruction errors e_t from the validation data. For a given threshold τ, if the number of points x_t for which the anomaly score a_t is greater than τ exceeds a certain count, the sequence X is classified as an anomaly.

2.2 Knowledge Distillation

Knowledge distillation is a technique for transferring the knowledge encapsulated within a large model to a smaller model. In this framework, the large model is referred to as the teacher model, and the small model as the student model. By learning from the teacher's knowledge, the student model's loss function converges more rapidly. Consequently, the student model, despite having significantly fewer parameters than the teacher, can achieve comparable performance and enables faster computation. In [3], it defines the output of the teacher model as knowledge and demonstrates that it can be effectively transferred to the student model through the distillation process.

Figure 1 illustrates the method of training a student model via knowledge distillation. First, to extract the knowledge, the teacher model is trained using the training data. The primary objective of pre-training the teacher model is to enable it to provide its knowledge-rich outputs to the student. Subsequently, the student model is trained on the same training data, utilizing both the outputs obtained from the teacher model and the ground-truth labels from the original data. Furthermore, [3] showed that applying a hyper-parameter called $TemperatureT$ to the final outputs allows for the transfer of more informative knowledge to the student model.

3. Proposed Method

In this paper, we propose a method for applying the knowledge distillation technique to an anomaly detection model that utilizes an LSTM autoencoder, as illustrated in Figure 2. For the teacher model, we employ the LSTM autoencoder from [1], and for the student model, we use a smaller LSTM autoencoder with a reduced number of parameters, achieved by decreasing the number of hidden units in its LSTM layers.

The multivariate time series data is represented as $X = \{x_1, x_2, \ldots, x_L\}$ of length L, where $x_t \in \mathbb{R}^m$ is an m-dimensional vector. The same training sequence is fed to both the pre-trained teacher model and the untrained student model. The student model learns using the sequence reconstructed by the teacher model and the original input sequence. The loss function L_S proposed in this paper for knowledge distillation is as follows:

Here, L_{MSE} denotes the Mean Squared Error loss function. X_S and X_T represent the sequences reconstructed by the student model and the teacher model, respectively. T and α are hyperparameters. In our experiments, we used a value for T in the range of 1 to 2, and for α in the range of 0.3 to 0.4.

Table 1: Performance Comparison.

	Precision	Recall	F1 score	Parameter
Teacher model	0.27	0.71	0.40	1,124,201
Student model	0.35	0.67	0.39	482,601

4. Experiment

In our experiments, we evaluated the anomaly detection performance of the teacher and student models using the Server Machine Dataset (SMD) [4], a multivariate time series dataset. The SMD dataset consists of data collected from 38 sensors at 1-second intervals over a five-week period. The training and testing sets contain 708,405 and 708,420 samples, respectively, with the proportion of anomalous data in the test set being 4.16%. To train the LSTM autoencoder, input sequences X were generated using a sliding window with a length of L=100 and a step size of 10. For detection, anomalous data were classified as Positive, and normal data as Negative.

The anomaly detection performance was measured by evaluating the model's ability to classify each input sequence X as anomalous or normal. The performance metrics used were Precision, Recall, and the F1-score, defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Figure 3 shows the loss value L_S for both the teacher and student models as a function of the training epochs. It is observed that the loss value for the student model, much like the teacher model, consistently decreases and converges over time.

Table 1 presents the final experimental results. The student model proposed in this paper has only about 40% of the parameters compared to the teacher model. Nevertheless, its Precision score increased by 0.08, while its Recall and F1-scores decreased by only 0.04 and 0.01, respectively, thus demonstrating anomaly detection performance comparable to that of the teacher model. This indicates that the student model can achieve similar detection performance to the teacher model with significantly fewer computational operations.

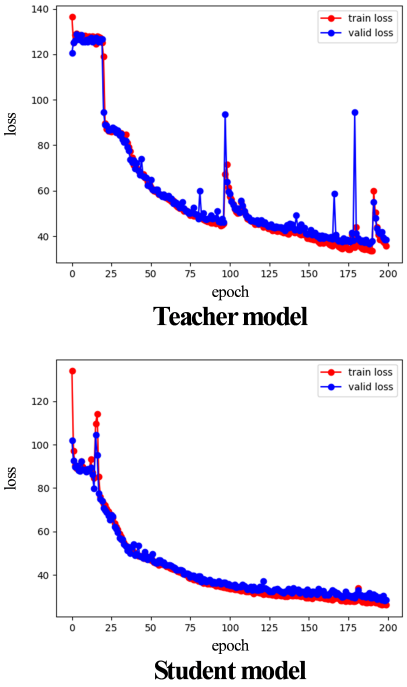

Figure 3: Learning curve.

5. Conclusion and Future Work

In this paper, we proposed a method for applying the knowledge distillation technique to the LSTM autoencoder, a deep learning model for anomaly detection. The proposed method utilizes the Mean Squared Error from both the teacher and student models in its loss function for knowledge distillation. Experiments on a real-world multivariate time series dataset demonstrated that through this knowledge distillation technique, it is possible to train a student model that achieves comparable anomaly detection performance while reducing the number of model parameters.

For future work, we plan to investigate methods for applying knowledge distillation to various other anomaly detection models beyond the LSTM autoencoder.

References

[1] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "Lstm-based encoder-decoder for multi-sensor anomaly detection," *arXiv preprint arXiv:1607.00148*, 2016.

[2] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[3] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[4] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2828–2837, 2019.