하드디스크의 잔존 수명 예측에 1D CNN-LSTM 을 이용한 모델 적용 연구

서양진*
*이포즌
yjseo@epozen.com

A Study on Applying a Model Using 1D CNN-LSTM to the RUL Prediction of HDD

Yangjin Seo*
*EPOZEN Co., Ltd.

요 약

제품이나 부품의 잔존 수명을 정확하게 예측할 수 있다면 고장이나 중단으로 인한 손실을 방지하는 것이 가능해질 것이다. 제품의 잔존 수명은 시계열 데이터 분석을 통해 예측될 수 있으며, 최근에는 딥러닝을 이용한 잔존 수명 예측 연구가 활발하게 진행되고 있다. 본 연구에서 우리는 컴퓨터 기반 시스템의 주요 고장 요소가 되고 있는 하드디스크의 잔존 수명을 예측하는 문제에 1D CNN-LSTM 을 이용한 모델을 적용하고, RMSE 와 R-Square 값을 이용해 적용한 모델의 성능을 평가하였다.

1. 서론

어떤 제품이나 부품의 잔존 수명(Remaining Useful Lifetime, 이하 RUL)은 이를 측정 또는 예상하는 시점부터 해당 제품이나 부품이 고장 나거나 교체되기까지 정상적으로 가동되는 시간이다. RUL 을 정확하게 예측할 수 있다면 시스템의 예상치 못한 고장이나 중단이 가져오는 손해를 사전에 막는 것이 가능해진다. 사물 인터넷과 빅데이터로 대변되는 4 차 산업혁명 시대를 맞아 보다 효과적이고 효율적인 고장 예지 및 건전성 관리를 위한 다양한 시도가 진행 중이다 [1][2]. 제품 잔존 수명은 순차적인 특징을 가지고 있기에 시계열 데이터 분석을 통해 이를 예측하는 연구가 수행되어 왔는데 [3][4], 과거에는 전문적인 도메인 지식과 기법의 적용 없이 시계열 데이터를 다루는 것이 힘들었지만 현재는 딥러닝 기술의 발전을 통해 이런 제약이 해소되고 있을 뿐만 아니라 더 효율적인 분석이 가능하게 되었다 [5-10].

컴퓨터를 기반으로 동작하는 다양한 시스템에 있어 하드디스크의 고장은 시스템 중단을 발생시키는 주요 원인이 되어 왔다 [11][12]. 이러한 문제의 해결을 위해 업계에서는 하드디스크의 현재 상태를 확인할 수 있는 S.M.A.R.T.(Self-Monitoring, Analysis and Reporting Technology) [13] 데이터를 정의하고 이를 활용한 하드디스크 관리를 수행해 왔으나 제조사나 제품마다 상이한 데이터를 발생시킬 뿐만 아니라 동일한 제품이라도 운영 환경에 따라 다른 고장 유형이나 다른 사용 수명을 가질 수 있기에 이 데이터를 바탕으로 한 잔존 수명 예측이 쉽지 않다 [14][15].

이에 본 연구는 하드디스크의 잔존 수명 예측에 1D CNN-LSTM 을 이용한 모델을 적용하고 그 성능을 평가하였다. 1D CNN 은 시계열 S.M.A.R.T. 데이터의 지역적인 특성을 효과적이고 효율적으로 추출할 수 있으며, LSTM 은 시계열 데이터에 기반한 잔존 수명 예측이 가능하도록 한다.

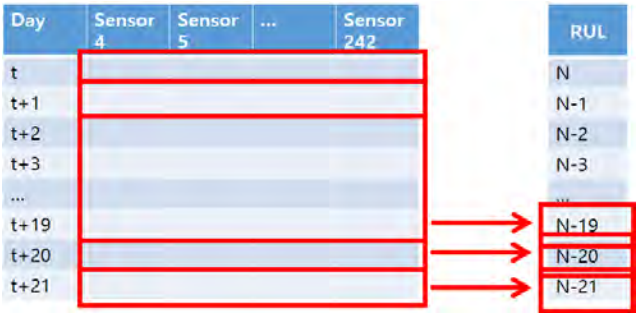
본 논문의 구성은 다음과 같다. 먼저 2 장에서는 본 연구에 사용된 데이터에 대해 설명하고 3 장에서는 본 연구에 사용된 RUL 모델과 1D CNN-LSTM 모델을 설명한다. 4 장에서는 실험을 통해 결정된 1D CNN-LSTM 모델의 상세와 성능 평가 결과를 제시하며, 마지막으로 5 장에서 결론을 맺는다.

2. 데이터 정의

본 연구에 사용된 데이터는 미국 데이터 스토리지 업체인 백블레이즈의 데이터 센터에서 얻은 것으로 하드디스크의 실제 운영 데이터로 구성되어 있다. 해당 데이터는 date, serial_number, model, capacity_bites, failure, smart sensor 들의 항목들로 구성되어 있으며,

serial_number 는 각 하드디스크 별로 붙여진 고유 번호, model 은 하드디스크 모델 번호, capacity_bites 는 하드디스크 용량, failure 는 하드디스크의 고장 여부 값을 가진다. 마지막 항목인 smart sensor 들은 1 번부터 255 번까지 있는데 그 중 일부 공개되지 않은 값이 존재한다.

학습과 시험에는 2016 년 2 분기부터 2019 년 1 분기까지 3 년 동안 순차적으로 기록된 데이터를 사용하였다. 딥러닝 적용 시 원시 데이터를 그대로 활용하는 것이 가능한 경우도 있으나, 탐색적 데이터 분석과 전처리를 통해 데이터를 정제하는 것이 학습 및 모델의 성능에 도움을 주기에 본 연구에서는 다음과 같은 탐색적 데이터 분석과 전처리를 수행하였다. 먼저 결측치가 70% 이상인 열들은 시계열 분석의 정확도를 감소시킬 수 있다고 판단하여 삭제하였으며, 0 으로만 채워져 있는 열들도 제거하였다. 여러 하드디스크 모델 중 데이터를 많이 보유하고 있는지, 고장난 데이터가 많은지, 긴 평균 생존 기간을 가지고 있는 지의 여부를 조건으로 ST4000DM000 을 선택하였다. 7 일 이상 날짜가 비어있거나 사용한지 30 일 안에 고장 난 데이터는 필터링하여 제거하였으며, 시간의 흐름에 따라 데이터가 연속되도록 선형 보간법을 사용하였다. 열 간의 상관관계를 구해 상관도가 0.9 가 넘어가는 열은 1 개만 남기고 삭제하였으며, RUL 목표 값을 특징으로 추가해 f-regression 함수를 이용하여 목표 값에 영향을 미치지 않는 열들은 삭제하였다. 탐색적 데이터 분석과 전처리 과정을 거친 후 최종적으로 남은 센서 열은 4, 5, 7, 9, 12, 187, 188, 190, 193, 197, 240, 241, 242 번이다.



(그림 1) 데이터 증강.

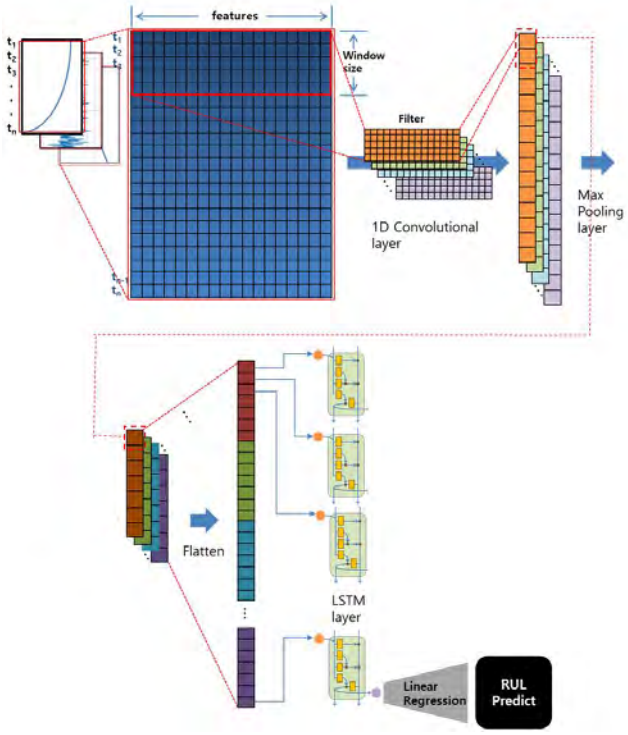
전처리가 끝난 데이터는 모든 하드디스크 데이터가 연결되어있는 형태이기 때문에 하드디스크들 간에 데이터가 연속하지 않는다는 것과 하드디스크 데이터의 생존 길이가 모두 다르다는 두가지 사실을 고려하며 (그림 1)과 같이 데이터 증강을 수행하였다. (그림 1)에서 데이터는 하드디스크 별로 20 일 단위로 나뉘며 하드디스크 별로 나뉜 데이터를 한 칸씩 이동하며 20 일 간의 순차적 데이터가 마지막 날의 RUL 값을 예측하도록 정의하였다.

3. 모델 정의

데이터 기반의 예측을 위해서는 각각의 입력 데이터에 대한 출력 RUL 값을 결정할 수 있어야 하는데, 물리적 모델 없이 각 단계의 시스템 상태를 정확하게 결정하는 것은 사실상 불가능하다. RUL 의 직관적인 모델로서 전체 생존 기간에서 해당 데이터 시점의 차이를 구해 잔존 수명을 할당하는 방식을 사용할 수 있겠으나, 시스템 성능저하가 일반적으로 어느 정도 시간이 경과된 후에 시작된다는 점을 감안할 때에 아래 식과 같은 구간별 함수(piecewise function)를 사용하는 것이 보다 현실에 가깝다고 판단하였다 [16].

$$RUL = \begin{cases} pw, & \text{if } 0 \le t \le pw \\ -t + tmr, & \text{if } t > pw \end{cases}$$

식에서 t 는 각 데이터의 시점으로 일(day) 단위이며, tmr 은 가장 오래 생존한 제품의 기간, pw 는 시스템의 성능 저하가 발생하기 시작하는 시점이다.



(그림 2) 1D CNN-LSTM 모델

(그림 2)는 본 연구에 사용된 1D CNN-LSTM 모델의 전체 구조이다. 먼저 1D CNN 모델은 합성곱층(Convolutional Layer)과 최대 풀링층(Max Pooling Layer)으로 구성되어 있다. 입력은 여러 센서 값을 포함하는 다변량(Multivariate) 데이터이다. 합성곱층에서는

창을 아래 방향으로 이동하며 합성곱 연산을 수행하며 이 과정에서 데이터의 지역적 특징들이 추출되어 특징맵(Feature Map)을 생성한다. 이어지는 최대 풀링층은 더 의미 있는 특징들을 추출하며 데이터의 크기를 줄인다. 그렇게 합성곱층과 최대 풀링층을 각각 4 번, 1 번 거친 결과 생성되는 행렬 데이터는 플래튼(Flatten) 계층을 거쳐 시계열 데이터를 기반으로 예측을 수행하는 대표적인 딥러닝 모델인 LSTM 계층에 입력된다. LSTM 계층에서는 입력된 데이터에 대해 반복 연산을 수행하며 이전 데이터들의 피드백을 받아 데이터의 시점들 간의 관계를 학습한다.

4. 실험 및 평가

합성곱층과 최대 풀링층의 개수와 필터의 개수, 커널 사이즈, 활성화 함수, LSTM 계층의 층의 개수, 유닛 개수, 드랍아웃 층의 개수와 마스킹 비율 및 배치(batch) 크기와 학습 비율을 변화시켜가며 실험을 수행하였으며, 최종적으로 결정된 모델의 상세는 <표 1> 과 같다. LSTM 은 두 층을 사용하였으며 LSTM 계층 사이와 LSTM 계층과 선형 회귀 함수 사이에 0.2 비율의 드랍아웃 계층을 설정하였다.

<표 1> 모델 상세

순서	계층 타입	Output Shape
1	Conv 1D	(None, 17, 64)
2	Conv 1D	(None, 14, 128)
3	MaxPooling 1D	(None, 7, 128)
4	Conv 1D	(None, 4, 128)
5	Conv 1D	(None, 4, 64)
6	Time Distributed Faltten	(None, 4, 64)
7	LSTM	(None, 4, 100)
8	DropOut	(None, 4, 100)
9	LSTM	(None, 50)
10	DropOut	(None, 50)
11	Dense	(None, 1)
12	Activation	(None, 1)

최종적으로 에폭(epoch)은 228 회 진행했으며, 127 번째 에폭에서 최적의 결과가 나왔다. 훈련 데이터 셋에 대한 RMSE 값은 0.0018, R-Square 값은 0.9620 이 나왔으며, 테스트 데이터에 대해서는 0.00084 의 RMSE 값과 0.9815 의 R-Square 값이 나왔다.

5. 결론

본 연구에서 우리는 하드디스크의 잔존 수명 예측을 위한 모델을 제안하고 해당 모델의 성능을 평가하였다. RUL 의 예측에 1D CNN-LSTM 모델을 사용하였는데, 1D CNN 은 데이터의 지역적 특징을 추출하는 역할을, LSTM 은 앞서 추출된 특징들의 장기적인 패턴을 인식하는 역할을 담당한다. RUL 목적 함수로는 제품 수명 패턴과 유사한 물리적 모델의 RUL 목적 함수를 적용하였다. 제안한 1D CNN-LSTM 모델의 성능을 RMSE 와 R-Square 를 통해 평가한 결과 장기간 쌓인 하드디스크 S.M.A.R.T 데이터로부터 의미 있는 특징들을 선별하고 이를 바탕으로 하드디스크의 잔존 수명을 예측하는 일에 우수한 성과를 보이는 것을 확인할 수 있었다.

참고문헌

[1] 이수학, 윤병동, “Industry 4.0 과 고장예지 및 건전성 관리 기술 (PHM) 의 방향”, 한국소음진동공학회지, 25 권, 1 호, 22-28, 2015.

[2] D. Kwon, M. R. Hodkiewicz, J. Fan, T. Shibutani and M. G. Pecht, “IoT-based Prognostics and Systems Health Management for Industrial Applications”, IEEE Access, 4, pp. 3659-3670, 2016.

[3] H. T. Pham, B. S. Yang, and T.T. Nguyen, “Machine Performance Degradation Assessment and Remaining Useful Life Prediction Using Proportional Hazard Model and Support Vector Machine”, Mechanical Systems and Signal Processing, 32, pp.320-330, 2012.

[4] T. H. Loutas, D. Roulias, and G. Georgoulas, “Remaining Useful Life Estimation in Rolling Bearings Utilizing Data-driven Probabilistic E-support Vectors Regression”, IEEE Transactions on Reliability, Vol. 62, no. 4, pp.821-832, 2013.

[5] L. Zhang, J. Lin, B. Liu, Z. Zhang, X. Yan and M. Wei, “A Review on Deep Learning Applications in Prognostics and Health Management”, IEEE Access, 7, pp. 162415-162438, 2019.

[6] O. Fink, Q. Wang, M. Svensén, P. Dersin, W.J. Lee and M. Ducoffe, “Potential, Challenges and Future Directions for Deep Learning in Prognostics and Health Management Applications”, Engineering Applications of Artificial Intelligence, 92, p. 103678-103684, 2020.

[7] N. Gugulothu, V. Tv, P. Malhotra, L. Vig, P. Agarwal and G. Shroff, “Predicting Remaining Useful Life Using Time Series Embeddings Based on Recurrent Neural Networks”, arXiv preprint arXiv:1709.01073, 2017.

[8] A. Z. Hinchi and M. Tkiouat, “Rolling Element Bearing Remaining Useful Life Estimation Based on a Convolutional Long-short-term Memory Network. Procedia Computer Science, 127, pp. 123-132, 2018.

[9] J. Niu, C. Liu, L. Zhang and Y. Liao, “Remaining Useful Life Prediction of Machining Tools by 1D-CNN LSTM Network” In 2019 IEEE Symposium Series on Computational Intelligence (SSCI), China, December 2019, pp. 1056-1063.

[10] J. R. Jiang, J. E. Lee and Y. M. Zeng, “Time Series Multiple Channel Convolutional Neural Network with Attention-Based Long Short-Term Memory for

Predicting Bearing Remaining Useful Life", Sensors, Vol. 20, no. 1, pp. 166-184, 2020.

[11] K. V. Vishwanath and N. Nagappan, "Characterizing Cloud Computing Hardware Reliability", In Proceedings of the 1st ACM Symposium on Cloud Computing, USA, June 2010, pp. 193-204.

[12] Y. Xu, K. Sui, R. Yao, H. Zhang, Q. Lin, Y. Dang, P. Li, K. Jiang, W. Zhang, J. G. Lou and M. Chintalapati, "Improving Service Availability of Cloud Systems by Predicting Disk Error", In 2018 USENIX Annual Technical Conference, USA, July 2018, pp. 481-494.

[13] Wikipedia, "S.M.A.R.T.", Retrieved Sep. 29, 2020, from https://en.wikipedia.org/wiki/S.M.A.R.T.

[14] S. Basak, S. Sengupta and A. Dubey, "Mechanisms for Integrated Feature Normalization and Remaining Useful Life Estimation Using Lstms Applied to Hard-disks", In 2019 IEEE International Conference on Smart Computing (SMARTCOMP), USA, June 2019, pp. 208-216.

[15] S. Lu, B. Luo, T. Patel, Y. Yao, D. Tiwari and W. Shi, "Making Disk Failure Predictions SMARTer!", In 18th USENIX Conference on File and Storage Technologies, USA, February 2020, pp. 151-167.

[16] P. Anantharaman, M. Qiao and D. Jadav, "Large Scale Predictive Analytics for Hard Disk Remaining Useful Life Estimation", In 2018 IEEE International Congress on Big Data, USA, July 2018, pp. 251-254.