제한된 데이터의 회전체 결함 진단을 위한 클래스 균형 샘플링기반 소수-샷 학습방법

김하율, 장계봉, 조성배

연세대학교 컴퓨터과학과

{hayoul1999, gyebong.jang, sbcho}@yonsei.ac.kr

A Few-shot Learning Method based on Class-balanced Sampling to Diagnose Rotor Defects of Limited Data

Hayoul Kim, Gyebong Jang, Sung-Bae Cho

Dept. of Computer Science, Yonsei University

요 약

최근 산업용 기기에 딥러닝을 이용한 고장 진단 및 예측 방법은 유망한 결과를 보여주었다. 이러한 학습 방법 대부분은 많은 양의 학습 데이터가 필요하다. 그러나 실제 산업용 기기의 고장 진단을 위한 양질의 데이터 확보가 어렵기 때문에 제한된 데이터를 이용한 소수-샷 학습 기법이 주목받고 있다. 본 논문에서는 기본 학습 데이터와 새로 수집된 데이터 간의 불균형으로 인해 학습 모델의 성능이 떨어지는 문제를 해결하기 위해, 데이터 간 분포의 유사도와 분포를 근거로 선택적인 데이터 훈련을 통해 더욱더 넓은 범주의 일반화를 이루어 낼 수 있다. 또한, 데이터를 고르게 사용함으로써 특정 데이터에 치우치지 않도록 한다. 실험 결과는 기존의 방법대비 약 5~17%의 성능 향상을 보여주었다.

1. 서 론

기계의 고장을 진단하고 예측하는 논제에 대한 많은 연구와 방법론이 제시되어 왔다. 이러한 연구의 어려움은 데이터셋의 부족함과 각 데이터셋들의 불균형성을 이유로 들 수 있다. 불충분한 데이터에서도 효과적인 성능을 확보하기 위해, 소수-샷 학습과 샴 네트워크를 결합한 방법이 제시되었다 [1].

그러나 기존의 방법은 그림 1과 같은 무작위성에 따른 훈련 방식을 사용했기 때문에 매 훈련 시 정확도의 차이가 매우 크게 나타나는 것을 확인할 수 있다. 따라서 이 논문에서는 데이터 분포를 이용한 제약조건을 따르는 방식으로 오차를 줄이고, 범주의 일반화를 통해 모델의 성능을 향상하고자 한다.

2. 관련 연구

딥러닝은 많은 데이터가 필요하다는 한계를 갖고 있다. 또한, 대부분의 현실 문제 상황에서는 충분한 양의 데이터 확보가 어렵다. 이러한 문제를 극복하기 위해, 적은 양의 샘플에서도 원하는 분류 성능을 내는 메타-학습 기법을 이용하여 소수-샷 학습의 포문을 열었다. 그 중 메트릭 학습을 기반으로 하여 같은 범주의 데이터끼리는 깊은 공간에서 더욱더 가깝게, 다른 범주에 대해서는 더 멀게 위치시키는 방법에 대한 연구가 진행되었다 [3]. 그 후에 데이터들의 분포를 이용하여 더욱 효율적으로 훈련을 진행할 수 있는 연구들 또한 진행되었다. 개별 데이터를 하나씩

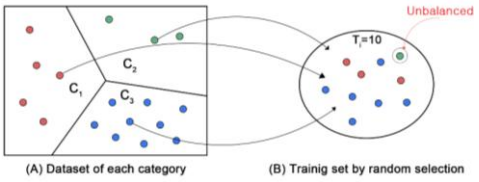


그림 1. 무작위 선택 방식의 한계. C_n은 데이터셋의 각 카테고리, T_i는 무작위 선택된 임의 학습 데이터셋.

훈련하는 것이 아닌 전체 데이터들을 한 번에 훈련하는 방법이 제시되었고 [4], 데이터를 하나씩 훈련했을 경우, 잡음에 대한 취약성과 편향성 제거의 어려움을 예로 들며 통계적 분포의 필요성이 강조되었다 [5].

표 1. 최근 소수-샷 학습 방법의 연구 동향

저자	방법	설명
Zhang, A. [1]	샴 신경 네트워크	데이터 간 거리 학습
Li, W. [3]	메트릭 학습	CNN기반 임베딩 특징 학습 및 이미지 모듈 별 유사도 측정
Li, H. [4]	범주 순회 모듈	추출한 특징 마스크를 축소된 차원에 적용
Zhang, J. [5]	분포 집합	데이터의 분포를 통해 특정-종류 분포로 집합

서식 있음: 가운데

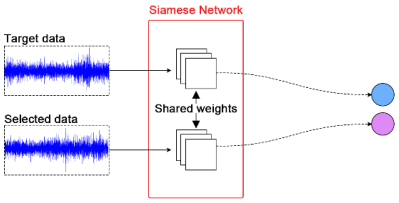


그림 2. 샴 네트워크의 구조

3. 학습모델과 분류방법

3.1 기존의 훈련 방법

기존의 방법[1]은 그림 2와 같이 넓은 첫 층을 가진 CNN을 기반으로 한 샴 네트워크(Siamese Neural Network)를 모델로 가진다. Weight를 공유하는 두 CNN 네트워크에 각각 범주가 같거나 다른 샘플 쌍을 넣어 두 샘플 사이의 거리에 대한 확률을 반환한다. 그 중 샘플을 뽑을 때는 무작위성에 기대어 훈련을 진행한다.

3.2 제안하는 방법

선택되지 않은 기본 클래스의 컬렉션을 F_b, 선택한 기본 클래스를 F_s로 나타내며, 새로운 클래스는 F_n으로 나타낸다. 선택 프로세스는 F_n에서 m개의 요소가 있는 부분 집합 X를 선택하는 것이며 기본 데이터 집합은 F_n과 F_b로 구성된다. 그림 3에서 확인할 수 있듯이, 평균으로 각 카테고리를 구분할 수 있다. 즉, 각 라벨의 평균을 통해 전체 분포를 확인할 수 있다. 그래프의 분포는 다음과 같은 함수를 통해 표현할 수 있다.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (1)$$

여기서 μ 는 평균을 뜻하며, σ^2은 분산, exp 는 자연상수를 뜻한다.

본 논문은 각 라벨의 특성을 나타낼 여러 통계 수치 중에서 유사도(SR) 값으로 평균과 분산을 이용한다. SR은 k-최근접-이웃 알고리즘을 통해 높은 단계의 이미지 특징을 뽑아내 일부 베이스 클래스들과 유사도를 구한 뒤, 다시 전체 베이스 클래스로 나눠 유사도 비율을 구하는 함수다. 훈련에 쓰일 데이터들을 평균을 기준으로 정렬한 후, 훈련 과정에서 같은 범주에 속하는 데이터에 가중치를 평균과의 거리를 기준으로 차등적으로 준다. 다음은 본 논문에서의 제시한 F_n 선택을 위한 SR 함수와 거리 d에 관한 함수이다.

$$SR = \frac{d \pm \sigma}{mean(x)} \quad (2)$$

$$d = abs(mean(arr[x]) - mean(arr[y])) \quad (3)$$

여기서 σ는 분산, x는 데이터 전체, 그리고 y는 선택된 데이터다. 학습 방법은 식 (2), (3) 과 같이 데이터 간의 평균과 분산을 근거로 유사도를 산출하여, 평균에 해당

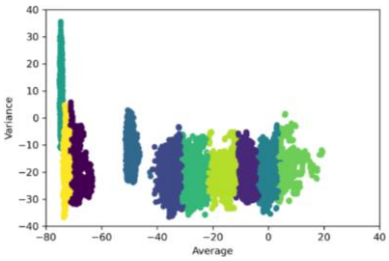


그림 3. 데이터의 분포를 나타낸 그림

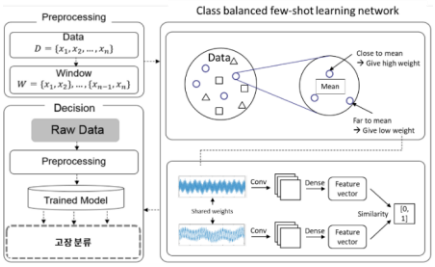


그림 4. 유사도 기반 데이터 선택 훈련 방식

하는 부분에 집중된 데이터를 우선 선택적으로 학습에 사용하며, 점차 그 범위를 확대해 나가도록 설계되었다. 그 후에 훈련 과정에서 데이터의 사용 횟수를 세어 훈련마다 가장 적게 쓰인 데이터를 사용한다. 이 방법으로 앞서 언급한 데이터 불균형과 데이터 부족에 대한 문제를 해결할 수 있다. 그림 4는 제안하는 모델의 구조를 나타낸다.

4. 실험 결과

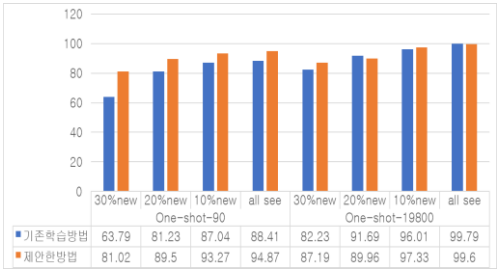
4.1 데이터

소수-샷 학습 알고리즘의 성능을 평가하기 위해 Case Western Reserve University(CWRU)의 베어링 데이터셋을 사용했다. CWRU 데이터셋은 표 2에 나와 있듯이 Fault location을 기준으로 None, Ball, Inner Race, Outer Race로 나뉘어 있고, None을 제외하곤 fault의 diameter를 기준으로 다시 0.007, 0.014, 0.021로 나뉘어 총 10개의 범주를 가진다. 각 라벨은 19,800개의 훈련 데이터와 750개의 테스트 데이터를 가진다.

표 2. 데이터 구성.

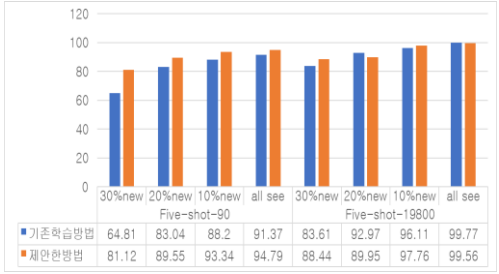
Fault Location	None		Ball			Inner Race			Outer Race	
Diameter (inch)	0	0.007	0.014	0.021	0.007	0.014	0.021	0.007	0.014	0.021
Label	1	2	3	4	5	6	7	8	9	10

표 3. 기존 1-샷 학습의 성능 결과



	30%new	20%new	10%new	all see	30%new	20%new	10%new	all see
	One-shot-90				One-shot-19800			
기존학습방법	63.79	81.23	87.04	88.41	82.23	91.69	96.01	99.79
제안한방법	81.02	89.5	93.27	94.87	87.19	89.96	97.33	99.6

표 4. 기존 5-샷 학습의 성능 결과



	30%new	20%new	10%new	all see	30%new	20%new	10%new	all see
	Five-shot-90				Five-shot-19800			
기존학습방법	64.81	83.04	88.2	91.37	83.61	92.97	96.11	99.77
제안한방법	81.12	89.55	93.34	94.79	88.44	89.95	97.76	99.56

4.2 결과 및 논의

실험은 데이터의 불균형을 해소할 수 있는지 확인하기 위해 10개의 카테고리 중 각 7개, 8개, 9개, 그리고 전부에 대한 훈련을 한 뒤, 10개의 각 카테고리에 해당하는 테스트셋 75개에 대한 정확도를 확인한다. 90개의 데이터로 실험할 때는 각 카테고리마다 6개의 학습 데이터, 3개의 확인 데이터로 훈련을 하고, 19,800개의 데이터로 실험할 때는 각 1,386개의 학습 데이터와 100개의 확인 데이터로 훈련을 한다. 표 3, 표 4는 각각 1-샷 검증과 5-샷 검증으로 평가한 성능을 10번의 결과들을 통해 평균을 낸 자료이다. 표2, 3의 결과에서 확인할 수 있듯이 90개의 데이터를 사용한 경우 전체적으로 정확도가 향상되었다. 특히 90개의 데이터셋에서 70%의 카테고리만 학습한 경우에 성능이 약 5~17% 향상된 것을 확인할 수 있다. 이를 통해 데이터가 적을 때, 불균등할 때 통계를 활용한 학습이 큰 영향을 끼친다는 것을 알 수 있다. 그림 5를 보면, 예측 라벨과 실제 라벨이 정확하게 분류되는 것을 볼 수 있다.

5. 결론

본 논문은 데이터가 적고, 불균등한 상황에서의 기계 장비 고장 예측을 데이터 분포를 통해 정확도를 한층 더 높이고자 하였다. 기존의 소수-샷 학습 모델은 유지하되, 오직 데이터를 선택하는 새로운 기준을

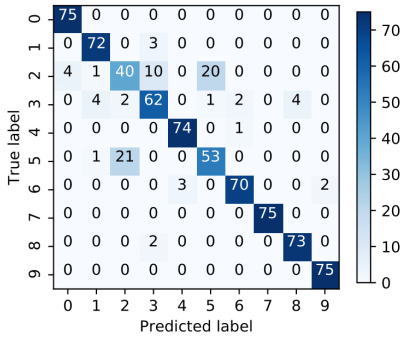


그림 5. 5-샷에서 30% 새로운 사례의 혼동 행렬

만드는 방법을 통해 높은 성능을 얻을 수 있었다. 특히 데이터가 부족한 산업현장 사례에서도 80%가 넘는 결과로 실제 장비에 적용할 수 있는 방법을 도출해냈다. 향후 이 알고리즘을 기존 소수-샷 학습 분야의 많은 문제의 무작위 선택 방식에 적용한다면 보다 향상된 결과를 기대할 수 있을 것이다. 본 논문을 통해 제시된 방법을 실제 장비에 탑재하여, 실시간 고장 진단 성능을 추가 검증할 계획이다.

참고문헌

[1] Zhang, A., Li, S., Cui, Y., Yang, W., Dong, R., and Hu, J., "Limited Data Rolling Bearing Fault Diagnosis With Few-Shot Learning," in *IEEE Access*, vol. 7, pp. 110895-110904, 2019.

[2] Wertheimer, D., & Hariharan, B., "Few-shot learning with localization in realistic settings." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 6558-6567, 2019.

[3] Li, W., Wang, L., Xu, J., Huo, J., Gao, Y., & Luo, J., "Revisiting local descriptor-based image-to-class measure for few-shot learning," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7260-7268, 2019.

[4] Li, H., Eigen, D., Dodge, S., Zeiler, M., & Wang, X., "Finding task-relevant features for few-shot learning by category traversal," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-10, 2019.

[5] Zhang, J., Zhao, C., Ni, B., Xu, M., & Yang, X., "Variational few-shot learning," *In Proceedings of the IEEE International Conference on Computer Vision*, pp. 1685-1694, 2019.