사물인터넷융복합논문지 Vol. 12, No. 3, pp. 19-24, 2026 https://doi.org/10.20465/KIOTS.2026.12.3.019

데이터 부족 환경에서 기계학습 성능 향상을 위한 데이터 증강 기반 학습 정책

이현섭*
백석대학교 컴퓨터공학부 교수

A Data Augmentation-Based Learning Policy for Improving Machine Learning Performance in Data-Scarce Environments

Hyun-Seob Lee*
Professor, Division of Computer Engineering, Baekseok University

요 약 최근 인공지능 기술의 발전으로 다양한 분야에서 자동화 시스템 구축이 진행되고 있으나, 고성능 기계학습 모델을 학습시키기 위한 대규모 고품질 데이터를 확보하는 데에는 막대한 비용과 시간이 소요된다. 특히 특정 도메인이나 저자원 환경에서는 수집 가능한 데이터가 제한적이며, 이는 모델의 과적합(Overfitting)과 일반화 성능 저하라는 심각한 문제를 야기한다. 본 연구는 이러한 데이터 부족 문제를 해결하기 위해 데이터 증강(Data Augmentation) 기법을 체계적으로 적용하는 데이터 관리 및 학습 정책을 제안한다. 제안하는 방법은 부족한 원본 데이터를 기반으로 기하학적 변형, 노이즈 삽입 등 다양한 증강 기법을 적용하여 학습 데이터셋의 다양성을 확보하고, 이를 통해 '차원의 저주'를 극복하며 모델의 예측 정확도를 향상시킨다. 본 연구는 제한된 데이터 자원을 가진 환경에서도 데이터 증강 정책을 통해 효율적인 기계학습 구현이 가능함을 입증하며, 이는 고성능 컴퓨팅 인프라나 대규모 데이터셋 접근이 제한된 연구자들에게 실용적인 솔루션을 제공할 것으로 기대된다.

주제어 : 데이터 증강, 기계학습, 데이터 부족, 과적합 방지, 데이터 전처리

Abstract Recent advances in artificial intelligence technology have driven the development of automated systems across various fields. However, acquiring large-scale, high-quality data required for training high-performance machine learning models demands significant costs and time. Particularly in specific domains or resource-constrained environments, the amount of collectable data is limited, leading to serious issues such as model overfitting and reduced generalization performance. This study proposes a data management and learning policy that systematically applies data augmentation techniques to address this data scarcity problem. The proposed method enhances the diversity of the training dataset by applying various augmentation techniques, such as geometric transformations and noise insertion, based on the limited original data. This approach overcomes the ‘curse of dimensionality’ and improves the model's prediction accuracy. This study demonstrates that efficient machine learning implementation is achievable even in environments with limited data resources through data augmentation policies. This is expected to provide a practical solution for researchers with restricted access to high-performance computing infrastructure or large-scale datasets.

Key Words : Data Augmentation, Machine Learning, Data Poverty, Overfitting Prevention, Data Preprocessing

This paper was supported by 2026 Baekseok University Research Fund
*교신저자 : 이현섭(hyunseob@bu.ac.kr)
접수일 2026년 02월 06일 수정일 2026년 05월 03일 심사완료일 2026년 06월 05일

1. 서론

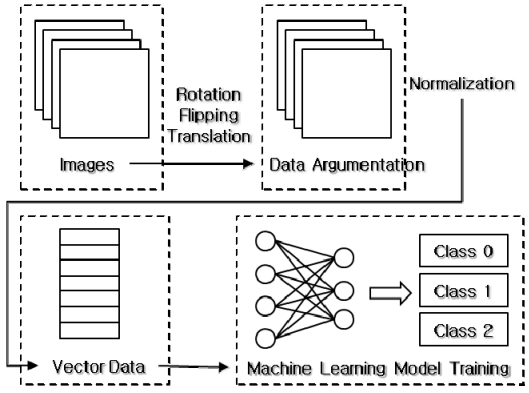
인공지능 기술의 급속한 발전은 현대 산업 전반에 걸쳐 패러다임의 전환을 가져왔다. 제조업에서는 예지보전과 품질 관리 자동화를, 의료 분야에서는 질병 진단 및 예후 예측을, 금융 분야에서는 신용평가 및 사기 탐지를 위한 데이터 기반 의사결정 시스템이 광범위하게 도입되고 있다[1-3]. 이러한 데이터 중심 의사결정은 기업의 경쟁력 강화와 사회적 가치 창출의 핵심 동력으로 자리잡았으며, 다양한 부가가치를 창출할 것으로 전망된다[4]. 그러나 이러한 고성능 기계학습 모델의 학습을 위해서는 대용량의 고품질 데이터가 필수적이다. 고성능 모델 학습을 위해서는 대용량의 데이터가 필수적이며, 이를 구축하고 유지보수하는 데 발생하는 초기 투자 비용은 중소기업이나 개발도상국 연구자들에게 기술 접근의 장벽으로 작용한다. ImageNet과 같은 대규모 데이터셋은 수백만 장의 레이블링된 이미지를 포함하고 있으며[5], 자연어 처리 분야의 대규모 언어모델들은 수십억 개의 토큰으로 학습된다[6]. 문제는 이러한 대용량 데이터를 구축하고 유지보수하는 데 소요되는 막대한 비용과 시간이다. 데이터 수집, 정제, 레이블링, 저장 및 관리를 위한 초기 투자 비용은 수억 원에서 수십억 원에 이르며, 전문 인력의 지속적인 관리가 요구된다[7]. 또한, 데이터가 부족한 환경에서 고차원 데이터를 다룰 경우, 특징 수에 비해 샘플 수가 적어 발생하는 과적합 문제와 이로 인한 성능 저하가 빈번하게 발생한다. 이는 자본력이 제한적인 중소기업이나 개발도상국의 연구자들에게 AI 기술 접근의 결정적 장벽으로 작용하고 있으며, 결과적으로 AI 기술의 불평등한 분배와 디지털 격차(digital divide)를 심화시키는 요인이 되고 있다[8]. 본 연구는 이러한 문제를 해결하기 위해 데이터 증강을 전처리 단계의 핵심 정책으로 활용하여, 부족한 데이터 환경에서도 기계학습 모델의 성능을 극대화할 수 있는 방법론을 제시하고자 한다. 제안하는 방법은 부족한 원본 데이터를 기반으로 기하학적 변형, 노이즈 삽입 등 다양한 증강 기법을 적용하여 학습 데이터셋의 다양성을 확보하고, 이를 통해 '차원의 저주'를 극복하며 모델의 예측 정확도를 향상시킨다. 본 연구는 제한된 데이터 자원을 가진 환경에서도 데이터 증강 정책을 통해 효율적인 기계학습 구현이 가능함을 입증하며, 이는 고성능 컴퓨팅 인프라나 대규모 데이터셋 접근이 제한된 연구자들에게 실용적인 솔루션을 제공할 것으로 기대된다.

2. 배경

현대 기계학습 시스템에서 다루는 이미지, 센서, 유전자 데이터 등은 수만 개 이상의 특징을 포함하는 고차원 데이터인 경우가 많으며, 이는 특징의 수가 늘어날수록 데이터 공간의 부피가 기하급수적으로 팽창하여 데이터 밀도를 희박하게 만드는 차원의 저주(Curse of Dimensionality) 현상을 야기한다. 이러한 고차원 환경에서 통계적으로 유의미한 수준의 데이터 밀도를 유지하기 위해서는 샘플의 수가 기하급수적으로 증가해야 하지만, 실제 산업 현장이나 소규모 연구 환경에서는 대규모 고품질 데이터를 확보하는 데 막대한 비용과 시간이 소요되는 것이 현실이다. 특히 고사양 하드웨어 및 GPU 인프라 구축에 따른 초기 투자 비용과 클라우드 유지보수 비용은 중소기업이나 특정 연구 기관에게 기술 접근의 장벽으로 작용하며, 이로 인해 충분한 학습 데이터를 확보하지 못한 저자원 환경에서의 기계학습은 더욱 큰 어려움에 직면하게 된다. 데이터가 부족한 환경에서 고차원 데이터를 학습할 경우, 기계학습 모델은 데이터의 본질적인 패턴을 학습하기보다 샘플에 포함된 지엽적인 노이즈까지 학습하게 되어 과적합(Overfitting) 문제와 일반화 성능의 심각한 저하를 겪게 된다. 또한, 제한된 메모리와 연산 능력을 가진 임베디드 시스템이나 엣지 컴퓨팅 환경에서는 고차원 데이터 처리에 따른 메모리 부족과 과도한 학습 시간이라는 실무적인 제약이 더해져 기계학습의 실용적인 적용이 제한된다. 따라서 이러한 물리적·경제적 한계를 극복하기 위해 기존의 제한된 데이터를 변형하여 새로운 샘플을 생성함으로써 학습 데이터셋의 양과 다양성을 인위적으로 확장하는 데이터 증강(Data Augmentation) 정책이 필요하다. 데이터 증강은 단순히 중요 정보를 선별하는 특징 선택(Feature Selection)이나 특징 추출(Feature Extraction) 기법과 달리, 원본 데이터 자체의 표본 수를 물리적으로 늘려 데이터 공간의 밀도를 보완함으로써 모델이 다양한 변칙적 환경에서도 견고하게(Robust) 작동할 수 있도록 돕는다. 특히 이미지 데이터의 경우 회전, 반전, 크기 조절 등의 기법을 체계적으로 적용함으로써 데이터 수집에 드는 추가 비용 없이도 데이터 희소성 문제를 해결할 수 있다. 결과적으로 이러한 데이터 증강 정책은 저사양 컴퓨팅 환경에서 발생하는 학습 성능 저하 문제를 해결하는 핵심적인 전처리 전략이 되며, 제한된 하드웨어 자원을 가진 환경에서도 모델의 정확도와 학습 효율성을 동시에 확보할 수 있다.

3. 데이터 증강 기반 학습 정책

3.1 핵심 아이디어



[Fig. 1] Key Idea

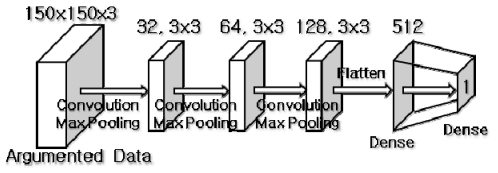
Fig. 1은 제안하는 핵심 아이디어를 보여주고 있다. 그림과 같이 이미지에 회전, 반전, 이동과 같은 효과를 주어 데이터를 증강한다. 그 다음 증강 데이터를 정규화 한다. 이렇게 벡터 형태로 변환된 데이터는 기계 학습을 위해 사용된다. 이러한 과정을 통해 학습된 모델은 회귀, 분류 등 증강되지 않은 일반 데이터에 대한 자동화 처리에 사용된다. 본 논문에서는 실험을 통해 이러한 데이터 증강이 미치는 영향을 분석한다. 결과적으로 분석된 결과는 소규모 연구 환경에서 적은 양의 데이터에 증강 기법을 적용했을 때 기계학습에 미치는 영향을 예측할 수 있을 것으로 기대한다.

3.2 데이터 증강

[Fig. 2] Data Augementation

Fig. 2는 원본 데이터와 증강된 데이터의 예를 보여주고 있다. 그림의 예제에서는 이미지 데이터에 대해 랜덤한 증강을 적용한 결과다. 데이터 증강에 활용된 기법은 이미지를 최대 40도까지 회전시키는 방식으로, 학습 과정에서 일부 데이터에 랜덤하게 적용된다. 이어서 가로 방향으로 최대 20%까지 이동시키는 기법과 세로 방향으로 최대 20%까지 이동시키는 기법을 적용하였다. 또한 이미지에 전단(shear) 변형을 주어 기울어진 형태로 만들거나, 최대 20% 범위 내에서 확대·축소하는 변형도 랜덤하게 선택하여 적용하였다. 마지막으로 좌우 반전 기법을 통해 일부 이미지를 뒤집어 학습에 활용하였다. 이러한 변환 과정에서 발생하는 빈 공간은 가장 가까운 픽셀 값으로 채워져 회전이나 이동으로 생긴 공백을 자연스럽게 메워주었다[9-11].

3.3 기계 학습 모델의 설계



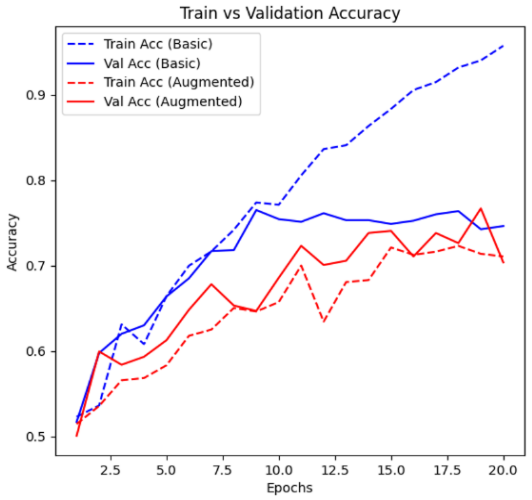
[Fig. 3] Design of Machine Learning Model

Fig. 3은 기계 학습에 사용된 모델의 설계를 보여주고 있다. 먼저 증강된 데이터는 컨볼루전과 최대값 풀링을 통해 32개의 특징맵으로 필터링 된다. 그다음 64개의 특징맵과 128개의 특징맵으로 필터링 된 후 512개의 데이터로 변환된다. 이 데이터는 신경망의 입력 층으로 전달되고, 신경망을 통해 데이터를 분류하기 위한 학습 데이터로 사용된다. 설계한 모델과 같이 본 논문에서는 일반적인 CNN모델[12-14]을 사용했고, 실험을 통해 동일한 모델에서 증강된 데이터와 일반 데이터를 사용했을 때의 차이를 분석한다. 이러한 분석은 증강 데이터가 미치는 영향을 파악할 수 있을 것으로 기대한다.

4. 실험 및 분석

기계학습에서 데이터 증강 정책을 적용했을 때 미치는 영향을 정량적으로 분석하기 위해 공개된 데이터[15]를 이용하여 데이터 증강한 실험을 진행하였다. 데이터는 Kaggle에서 개와 고양이를 분류하는 모델을 학습하기

위한 데이터셋이다. 데이터는 학습용 데이터는 8,000장(개: 4,000, 고양이: 4,000)이고, 검증용 데이터는 10,000장(개: 5,000, 고양이: 5,000)으로 구성되었다. 실험 환경은 AMD Ryzen Threadripper PRO 5975WX 32-Cores, 3600Mhz, 512GB RAM, RTX4090x4를 탑재한 윈도우 OS환경의 워크스테이션에서 진행하였다.

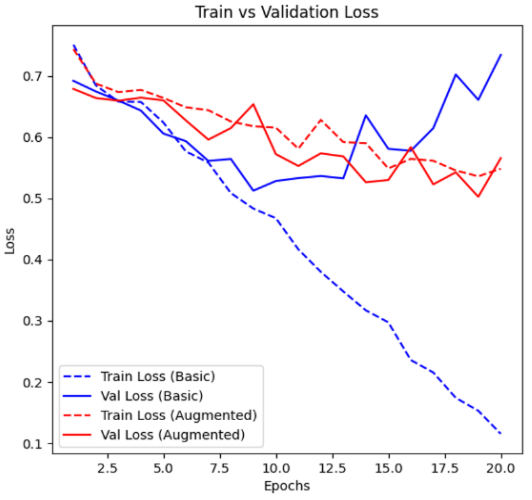


[Fig. 4] Accuracy Result

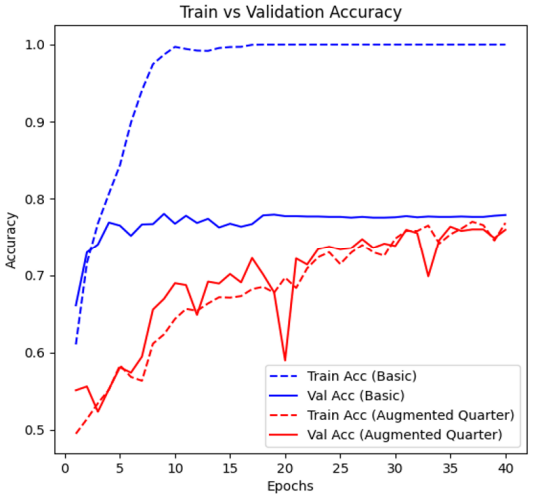
Fig. 4는 일반 데이터와 증강 데이터를 이용한 학습 과정에서 측정된 정확도와 손실 값을 보여주고 있다. 실험에서 x축은 학습이 반복된 횟수를 의미하고 y축은 정확도와 손실값을 의미한다. 증강이 적용되지 않은 기본 데이터는 약 9회 학습 할 때까지 정확도가 약 78% 수준까지 증가하였다. 그러나 이후 학습을 진행할수록, 학습 데이터와 검증 데이터 사이 정확도 값의 차이가 벌어졌고, 20번째 학습에서 학습 데이터는 정확도가 약 95.75%에 도달하였고 검증데이터는 약 74.62%의 결과를 보였다. 이러한 결과는 학습을 할수록 과적합 수준이 증가하는 것을 의미한다. 반면, 증강된 데이터를 이용한 실험에서는 학습에서 최종 약 71.06%이고, 약 70.38%였다. 결과적으로 기본 데이터는 학습과

검증사이 정확도가 큰 차이를 보였으나 검증 과정에서는 5% 미만의 차이로 같은 경향을 보였다.

Fig. 5는 손실값 결과를 보여주고 있다.기본 데이터는 학습을 할수록 저하되어 0.12 까지 저하되었으나 검증 데이터에서는 7번째 학습 이후 점차 증가하여 최종 손실값은 0.73까지 증가하는 것을 확인하였다. 반면 데이터 증강을 적용한 경우 학습과 검증의 손실값이 같은 경향을 보였고, 최종 약 0.55와 0.57%를 보였다. 이 결과는 데이터 증강이 과적합을 방지하는데 효과가 있음을 의미한다.



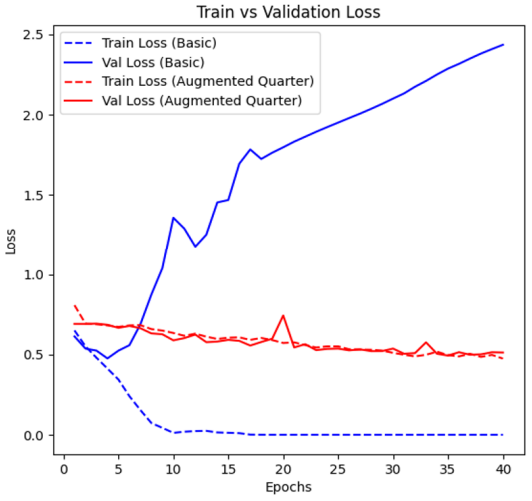
[Fig. 5] Loss Value



[Fig. 6] Accuracy Result with 25% Augementation

소규모 연구 환경에서는 데이터 수집이 어렵고, 대규모 데이터에 대한 최적화된 학습 효과를 도출하기까지 비용 부담을 갖게된다. 따라서 소규모의 데이터를 이용하여 증강하였을 때 기본 데이터와 비교하여 발생되는 성능의 차이에 대한 비교 분석 연구가 필요하다. Fig. 6은 소규모 데이터 환경을 가정하여 전체 학습 데이터의 25%만을 사용하였고, 이 데이터를 증강하였을 때 일반 데이터의 학습 결과와의 차이를 비교하였다. 그림에서 x축은 반복된 학습의 숫자이고, y축은 정확도를 의미한다.

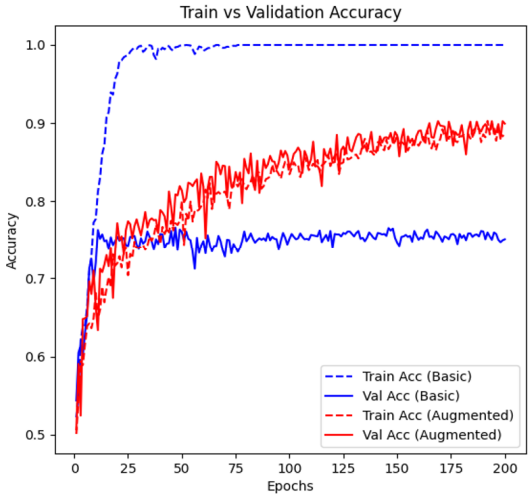
그림과 같이 4,000장의 기본 데이터를 이용한 학습 결과는 학습과정에서 100%의 정확도를 보였으나 5,000장의 검증 데이터를 이용한 검증 과정은 7회 학습 이후 약 77.9%에 도달하였고 이후 유의미한 학습 효과를 보이지 않았다. 즉 일정 횟수 이후 과적합이 발생하는 것을 확인할 수 있다. 반면, 학습 데이터의 25%인 1,000장의 데이터를 증강한 실험에서는 검증과 학습이 동일한 경향을 보였고, 최종적으로, 학습에서 76.87%, 검증에서 75.98%의 성능을 보였다. 이 결과는 소규모의 데이터로 과적합 없이 전체 데이터를 학습한 정확도와 유사한 수준까지 성능을 향상할 수 있음을 의미한다.



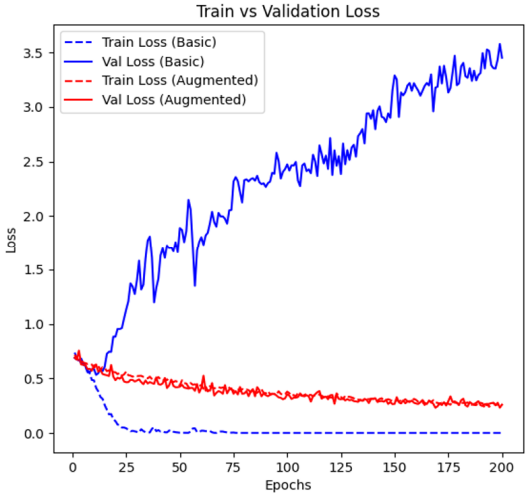
[Fig. 7] Loss Value with 25% Augementation

Fig. 7은 기본 데이터의 25% 데이터만 사용하여 증강하였을 때 학습과정에서 발생하는 손실 값을 보여주고 있다. 그림에서 x축은 학습 횟수를 의미하고 y축은 손실값을 의미한다. 그림의 결과와 같이 증강 데이터에서 학습한 결과는 학습을 할수록 낮아졌고 40회 반복했을 때 학습 데이터와 검증 데이터에서 각각 0.47과 0.51을 보였다. 반면 기본 데이터는 각각 0과 2.44였다.

Fig. 8은 증강 데이터에 대한 학습 횟수를 200회까지 늘렸을 때의 정확도 결과를 보여주고 있다. 그림에서 x축은 학습 횟수를 의미하고, y축은 정확도를 의미한다. 그림과 같이 증강된 데이터에 대한 정확도는 학습과 검증이 같은 경향을 보인다. 또한, 기본 데이터의 경우 일정 횟수의 학습 이후 정확도가 크게 변하지 않고 각각 100%, 75.13%로 과적합 현상을 보인 반면 증강 데이터는 최종 88.65%, 89.94%로 과적합 없이 우수한 성능을 보이는 것을 확인하였다.



[Fig. 8] Accuracy Result with Epochs 200



[Fig. 9] Loss Value with Epochs 200

Fig. 9는 동일한 실험에서 손실값을 보여주고 있다. 손실값 또한 정확도와 동일한 추세를 보여주고 있으며 최종적으로 증강 데이터에서 기계학습을 했을 때 발생한 손실값은 학습 과정과 검증 과정에서 동일하게 약 0.26의 결과를 보였다. 반면, 기본 데이터에서 기계학습의 손실값은 학습 과정에서 0, 검증 과정에서 3.45를 보였다.

이러한 실험을 통해, 한정된 데이터를 이용하여 기계학습을 할 때 데이터 증강을 적용하는 것이 과적합을 줄이고 성능을 개선하는데 효과가 있음을 확인할 수 있었다.

5. 결론

본 논문에서는 데이터가 부족한 환경에서 기계학습의 성능을 높이기 위한 데이터 증강 전처리 정책을 제안하였다. 실험적 배경을 통해 고차원 데이터의 차원의 저주 문제를 확인하였으며, 이를 극복하기 위한 수단으로 데이터 증강의 필요성을 강조하였다. 제안된 정책은 제한된 자원을 가진 임베디드 환경, 엣지 컴퓨팅, 그리고 데이터 수집이 어려운 연구 환경에서 모델의 정확도를 확보하는 데 큰 효과가 있을 것으로 예상된다. 향후에는 증강된 데이터의 양이 늘어남에 따라 발생하는 연산 부담을 줄이면서도 정확도를 추가로 향상시킬 수 있는 최적화 알고리즘에 대해 연구할 예정이다

REFERENCES

[1] I.Hector and R.Panjanathan, "Predictive maintenance in Industry 4.0: a survey of planning models and machine learning techniques," PeerJ Computer Science, Vol.10, e2016, 2024.

[2] N.Thakur, P.Sharma, S.Jain and A.Sharma, "Deep learning approaches for medical image analysis and diagnosis," Cureus, Vol.16, No.5, e59507, 2024.

[3] Z.Feedzai, "2025 AI Trends in Fraud and Financial Crime Prevention Report," Feedzai Press Release, 2025.

[4] W.Zeng, "Image data augmentation techniques based on deep learning: A survey," Mathematical Biosciences and Engineering, Vol.21, No.6, pp.6190-6224, 2024.

[5] C.Shorten and T.M.Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of Big Data, Vol.6, No.1, pp.1-48, 2019.

[6] N.V.Chawla, K.W.Bowyer, L.O.Hall and W.P.Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, Vol.16, pp.321-357, 2002.

[7] H.Zhang, M.Cisse, Y.N.Dauphin and D.Lopez-Paz, "mixup: Beyond empirical risk minimization," International Conference on Learning Representations (ICLR), 2018.

[8] S.Yun, D.Han, S.J.Oh, S.Chun, J.Choe and Y.Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp.6023-6032, 2019.

[9] Y.Pu, Y.Wang, Z.Xia, Y.Han, Y.Wang, W.Gan, Z.Wang, S.Song and G.Huang, "Adaptive Rotated Convolution for Rotated Object Detection," International Conference on Computer Vision (ICCV), pp.6589-6600, 2023

[10] Z.Xie, Z.Zhang, Y.Cao, Y.Lin, Y.Wei, Q.Dai and H.Hu, "On Data Scaling in Masked Image Modeling," IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10365-10374, 2023.

[11] Teerath Kumar; Rob Brennan; Alessandra Mileo; Malika Bendechache, "Image Data Augmentation Approaches: A Comprehensive Survey and Future Directions," Journals & Magazines, Vol.12, pp.187536-187571, 2024.

[12] P.Purwono, A.Ma'arif, W.Rahmaniar, H.I.K.Fathurrahman, A.Z.K.Frisky and Q.M.Haq, "Understanding of Convolutional Neural Network (CNN): A Review," International Journal of Robotics and Control Systems, Vol.2, No.4, 2022.

[13] J.Gupta, S.Pathak and G.Kumar, "Deep Learning (CNN) and Transfer Learning: A Review," Journal of Physics: Conference Series, 2022.

[14] A.Wang, H.Chen, Z.Lin, J.Han and G.Ding, "RepViT: Revisiting Mobile CNN From ViT Perspective," IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.15909-15920, 2024.

[15] https://www.kaggle.com/datasets/tongpython/cat-and-dog/data

이 현 섭(Hyun-Seob Lee) [종신회원]

- 2013년 2월 : 한양대학교 컴퓨터공학과 (공학 박사)
- 2012년 3월 ~ 2021년 2월 : 삼성전자 책임연구원
- 2021년 3월 ~ 현재 : 백석대학교 컴퓨터공학부 조교수

〈관심분야〉
인공지능, 저장시스템, 임베디드 시스템