

최소 손실 기반 동적 re-Batch에서 이미지 데이터 증강 효과

이석원*, 강태원**

The Effect of Image Data Augmentation in Minimum Loss-based Dynamic Re-Batch

Seokwon Lee*, Taewon Kang**

요약

기존 데이터 선택·증강 결합 기법의 연산 복잡성을 극복하기 위해, 본 논문은 단일 손실값 지표 기반의 동적 re-Batch와 이미지 데이터 증강을 결합한 새로운 학습 파이프라인을 제안한다. 손실 값이 낮은 샘플을 학습 데이터셋에 일정한 간격으로 분배하여 학습 안정성을 높이고, 증강을 통해 데이터 다양성을 확보하는 두 가지 전략의 이점을 동시에 취한다. CIFAR-10 데이터셋에서 전체 학습 데이터의 75%만 사용하고, 그중 8%만 증강해도 증강 없이 100% 전체 데이터를 사용한 기존 학습보다 높은 정확도를 달성함을 실험적으로 입증하였다. 이러한 결과는 re-Batch의 학습 안정화 효과와 증강의 데이터 다양성 확보가 시너지를 발휘하여 학습 데이터가 부족한 조건에서도 우수한 성능을 달성할 수 있음을 보여준다.

Abstract

To overcome the computational complexity of existing data selection and augmentation methods, this paper proposes a new training pipeline that combines a single-loss-metric-based dynamic re-Batch technique with image data augmentation. By distributing samples with low loss values at regular intervals across the training dataset to enhance training stability, securing data diversity through augmentation, it simultaneously takes advantage of the two strategies. Experiments on the CIFAR-10 dataset demonstrate that using only 75% of the total training data, with just 8% of it augmented, achieves higher accuracy than conventional training using 100% of the data without augmentation. These results show that the stabilizing effect of re-Batch and the data diversity provided by augmentation work synergistically to achieve excellent performance even in conditions under scarce training data.

Keywords

limited data, image data augmentation, dynamic data selection, dynamic re-Batch, minimum loss-based

* 강원대학교 원주캠퍼스 산업대학원 석사과정
- ORCID: <https://orcid.org/0009-0005-8755-4325>
** 강원대학교 원주캠퍼스 컴퓨터공학과 교수(교신저자)
- ORCID: <https://orcid.org/0000-0003-4343-5517>

• Received: Jun. 15, 2026, Revised: Jul. 23, 2026, Accepted: Jul. 26, 2026
• Corresponding Author: Taewon Kang
Dept. of Computer Science & Engineering, Kangwon National University, Korea
Tel.: +82-33-760-8666, Email: twkang@kangwon.ac.kr

I. 서 론

인공 신경망 기반의 딥러닝 기술은 이미지 인식, 자연어 처리 등 다양한 분야에서 혁신적인 성과를 거두고 있지만, 모델의 성능은 학습 데이터의 양과 질에 크게 의존한다.

이를 극복하기 위한 기법들 중 데이터 증강은 기존 샘플에 회전, 크롭, 색상 변형 등 다양한 변환을 가하여 인위적으로 데이터의 다양성을 높임으로써 모델의 일반화 능력을 향상한다[1].

단순히 데이터 다양성을 높이는 것을 넘어, 유의미한 샘플을 선별하거나 배치 구성을 최적화하는 전략들이 있다. 그중 동적 re-Batch 전략[2]은 도메인에 특화된 복잡한 휴리스틱이나 가중치 부여 알고리즘 대신, 샘플의 손실값이라는 단일하고 범용적인 지표를 사용하여 배치를 동적으로 재구성한다. 특히 최소 손실 기반의 동적 re-Batch가 모델의 수렴 안정성과 성능 개선에 효과적임이 입증된 바 있다.

본 연구는 학습 데이터가 부족한 조건에서 학습 성능을 개선하는 것을 목표로 한다. 이를 위해 최소 손실 기반의 동적 re-Batch 전략과 이미지 데이터 증강 기법을 결합한 새로운 학습 파이프라인을 제안한다. 먼저, 손실값이 낮은 샘플들을 일부 선정하여 학습 데이터세트에 일정한 간격으로 분배한다. 그리고 이미지 데이터 증강을 통해 학습 데이터의 다양성을 확보한다. 이 학습 파이프라인이 CIFAR-10 데이터세트에서 효과적임을 실험적으로 검증하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 살펴보고, 3장에서는 제안하는 통합 학습 파이프라인을 설명한다. 이어 4장에서 실험 환경을 소개하고 실험 결과 분석한 뒤, 5장에서 결론 및 향후 과제를 제시한다.

II. 관련 연구

2.1 이미지 데이터 증강

이미지 데이터 증강의 전통적인 기법인 수평 뒤집기, 크롭[3], 회전[4] 등의 기하학적 변환에서 시작하여, 최근에는 두 개 이상의 샘플을 혼합하는

Mixup[5], 특정 영역을 타 이미지로 대체하는 CutMix[6], 그리고 클래스 불균형 해소를 위해 소수 클래스에 가중치를 두는 Remix[7]와 같은 다중 인스턴스 혼합 기법으로 진화하였다.

또한, 최적의 증강 정책을 자동으로 탐색하는 AutoAugment[8], 그리고 단일 무작위 증강만으로도 우수한 성능을 보이는 TrivialAugment[9]등이 제안되며 학습 성능을 높여왔다.

2.2 동적 데이터 선택과 증강의 결합

데이터 효율성을 극대화하기 위해 전체 데이터 중 유의미한 샘플만을 선별하여 학습하는 데이터 선택(Data selection) 연구가 활발히 진행되고 있다.[10]

초기 연구는 쉬운 샘플부터 학습하는 커리큘럼 학습이나[11] 어려운 샘플에 집중하는 하드 예시 마이닝 등이 있으며[12], 최근에는 데이터 선택과 증강을 하나의 프레임워크로 통합하려는 시도도 등장했다.

Yang, Ye, Shen 그리고 Zhou[13]가 제안한 기법은 이 분야의 대표적 선행 연구인데, 샘플의 밀도와 멀티모달 시맨틱 일관성을 결합하여 증강에 적합한 샘플을 선별한다. 특징 공간에서 밀도가 낮은 샘플을 우선 선택하여 증강을 적용하고, CLIP(Contrastive Language-Image Pre-training)과 같은 사전 학습된 멀티모달 모델을 통해 노이즈 샘플을 필터링한다.

샘플의 신뢰성을 평가하기 위해 CLIP이라는 사전 학습된 멀티모달 모델이 필수적이며, 온라인 학습 중 밀도 계산을 위한 근사 최근접 이웃 탐색이 동반되어야 하므로 시스템 구현이 복잡하고 연산 오버헤드가 크다. 또한, 이미지-텍스트 쌍의 일관성에 의존하므로 멀티모달 정보가 없는 도메인에는 적용하기 어렵다.

2.3 기존 연구와 차별점

본 연구에서 활용하는 동적 re-Batch 전략은 선행 연구보다 단순하며 범용적인 방법론이다. CLIP이나 복잡한 밀도 계산 대신, 샘플별 손실값만을 지표로 사용하여 학습 난이도를 정의한다. 이는 추가적인 모델이나 사전 분석 없이도 모든 데이터세트에 즉

시 보편적으로 적용 가능한 방법론이다.

또한 Yang, Ye, Shen 그리고 Zhou[13]가 제안한 기존 기법은 학습이 부족한 데이터 샘플을 선택하기 위해 특징 공간에서 밀도가 낮은 샘플을 찾아낸다. 반면에, 본 연구는 손실값이 작은 샘플들을 이용한다.

III. 제안 방법론

3.1 동적 re-Batch

학습 안정성을 확보하기 위해 손실값에 기반하여 배치를 재구성하는 동적 re-Batch 기법을 적용한다.

매 에포크 종료 후, 전체 학습 데이터에 대해 증강을 적용하지 않은 상태에서 손실값을 측정한다. 이 중 손실이 가장 낮은 샘플들을 최적 샘플로 정의하고, 정렬하여 관리한다. 선별된 최적 샘플들을 학습 데이터 사이에 일정한 간격으로 분배하여 배치를 재구성한다.

전체 학습 데이터 크기를 N 이라고 하고, re-Batch 비율을 r_{re} 라고 할 때, 매 에포크마다 선정되는 최적 샘플 수 K 와 재배치 간격 I 는 다음과 같다.

$$\begin{aligned} K &= \lfloor N \times r_{re} \rfloor \\ I &= \lfloor N/K \rfloor \end{aligned} \quad (1)$$

직관적인 이해를 돕기 위해 그림 1에 re-Batch 과정을 간략히 나타내었다.

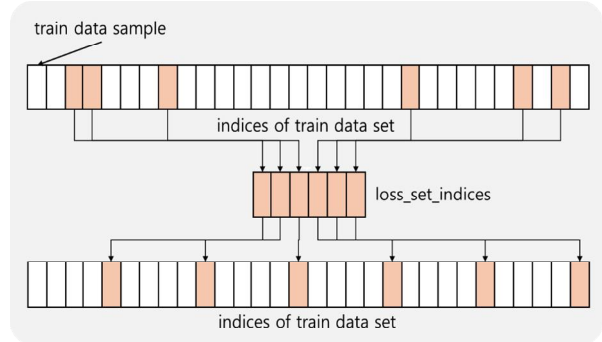


그림 1. 동적 re-Batch의 샘플 재분배 과정 예시 이미지[2]

Fig. 1. Example image of the process of repositioning data samples for re-Batch strategy[2]

3.2 동적 re-Batch 이용 증강

본 알고리즘은 매 에포크마다 3단계의 루프를 반복 수행하며, 제안하는 동적 re-Batch 기반 증강 기법의 전체적인 의사코드는 알고리즘 1과 같다.

알고리즘 1. 동적 re-Batch 이용 증강 알고리즘

Algorithm 1. algorithm of augmentation using dynamic re-Batch

Input	Trainind dataset D of size N , Model Θ , Total epochs E , re-Batch ratio r_{re} , Augmentation ratio r_{aug}
Output	Trained model Θ
1:	Calculate re-Batch samples count K and augmentation samples count M
2:	initialize sample indices $\mathcal{I}_{train}$
3:	for epoch=1 to E
4:	// step1 : Training
5:	Set model to Train mode
6:	Randomly select a subset of indices $\mathcal{I}_{aug}$ of size M from $\mathcal{I}_{train}$
7:	for batch in dataloader(D , $\mathcal{I}_{train}$):
8:	for each sample in batch:
9:	if sample index i is in $\mathcal{I}_{aug}$:
10:	sample = ApplyAugmentation(sample)
11:	Forward pass, compute loss, backward pass and update Θ
12:	//step2 : Dynamic re-Batching for next epoch
13:	Set model to Eval mode
14:	initialize an empty list $Loss_List$
15:	for each sample in D :
16:	Calculate $Loss$
17:	Append (sample index, $Loss$) to $Loss_List$
18:	//step3 : Update re-Batch target
19:	Sort $Loss_List$ in ascending odrder based on $Loss$
20:	$\mathcal{I}_{best}$ = Extract indices of top- K sample from $Loss_List$
21:	$I = \lfloor N/K \rfloor$
22:	$\mathcal{I}_{train} = MergeWithinterval(\mathcal{I}_{best}, interval = I)$

1 단계, 설정된 증강 비율에 따라 샘플들에 증강을 적용하여 모델을 학습시킨다. 증강 비율을 r_{aug} 라고 할 때 증강이 적용되는 총 샘플 수 M 은 다음과 같다.

$$M = \lfloor N \times r_{aug} \rfloor \quad (2)$$

2 단계, 에포크 종료 후 다시 증강 없는 전체 데이터 손실을 재계산하여 다음 에포크에서 사용할 최적 샘플 리스트를 최신화한다. 3 단계, 갱신된 인덱스 순서에 따라 다음 에포크에 사용할 배치를 재구성한다. 이와 같은 통합 전략은 손실값이 작은 일부 샘플들을 학습 데이터세트에 일정한 간격으로 분배함으로써 얻는 수렴 안정성과 데이터 증강을 통해 얻는 학습 데이터의 다양성 확보를 동시에 달성한다. 이는 동적 re-Batch를 위한 최소한의 과정을 제외하면 추가적인 연산 오버헤드가 거의 발생하지 않는다.

IV. 실험 결과 및 분석

4.1 실험 설정

모든 입력 이미지는는 크롭, 수평 뒤집기, TrivialAugment를 증강 알고리즘으로 채택하였다.

실험의 재현성을 확보하기 위해 모든 무작위 시드를 42로 고정하였으며, 기존 연구들과 대부분의 설정을 동일하게 하여 일관성을 유지했다[13][14]. 본 연구의 인공지능명 모델로 ResNet18을 채택하였다. Optimizer는 Nesterov 모멘텀이 적용된 SGD를 사용하였다. 모멘텀 계수는 0.9, 가중치 감소는 $5E-4$ 로 설정하였다. 학습률 스케줄러는 OneCycleLR 방식을 적용하였다. 기본 학습률은 0.001에서 시작하여 학습 진행에 따라 최대 학습률 0.2까지 변화하도록 설정하였다. 기존 연구들은 에포크를 200으로 설정했지만, 본 연구에서는 학습 시간을 단축하기 위해 100 에포크 동안 학습을 수행하였으며, 배치 크기는 128로 설정하였다. 손실 함수로는 CrossEntropyLoss를 사용하였다. 모든 실험 결과는 각 10회씩 실행하여 평균을 낸 값이다.

실험에는 CIFAR-10[15] 표준 데이터세트를 사용

하였다. 학습 데이터 부족 상황을 가정하기 위해 원본 학습 데이터 50,000장을 75%의 학습 데이터 (37,500장)로 분할하였으며, 별도의 10,000장 테스트 데이터를 통해 최종 모델 성능을 평가하였다.

4.2 성능 평가

전체 학습 데이터(100%)를 사용하여 증강하지 않았을 때의 정확도는 89.55%이며 표본 표준 편차는 0.269이다. 동적 re-Batch 비율 0.005, 0.03, 0.08, 0.15, 0.3, 0.5일 때의 실험 결과는 표 1에 나타내었으며, 그에 해당하는 표본 표준 편차는 표 2에 나타내었다. 직관적인 이해를 위해 표 1의 동적 re-Batch 비율별 도표를 그림 2, 그림 3, 그림 4에 각각 나타내었다. 비교를 위해 동적 re-Batch를 적용하지 않고 무작위 증강만 사용한 결과는 No re-Batch에 표기했다. 실험 결과 동적 re-Batch 비율 0.005일 때 가장 성능이 높으며, 비율이 높아질수록 성능이 감소한다. 동적 re-Batch와 증강을 통한 다양성 확보 사이의 적절한 균형점을 찾는 것이 성능 개선의 핵심임을 시사한다.

표 1. 실험 결과 정확도(%)

Table 1. Experiment result accuracy(%)

Augment select ratio	No re-Batch	re-Batch 0.005	re-Batch 0.03	re-Batch 0.08
0	87.48	86.63	86.15	84.59
0.03	88.05	88.44	87.88	86.47
0.08	88.81	89.77	89.65	88.26
0.15	89.53	91.18	91.01	89.81
0.3	91.35	92.79	92.61	91.54
0.5	93.04	93.69	93.55	92.96

표 2. 실험 결과 표본 표준 편차

Table 2. Experiment result sample standard deviation

Augment select ratio	No re-Batch	re-Batch 0.005	re-Batch 0.03	re-Batch 0.08
0	0.225	0.217	0.244	0.337
0.03	0.169	0.288	0.216	0.347
0.08	0.093	0.175	0.322	0.414
0.15	0.262	0.1	0.17	0.27
0.3	0.231	0.192	0.194	0.317
0.5	0.171	0.154	0.141	0.222

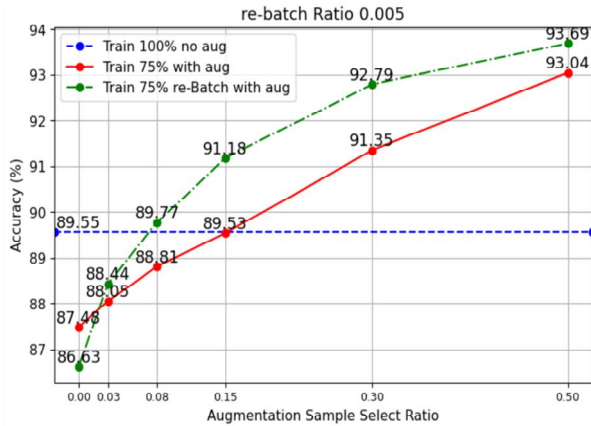


그림 2. re-Batch 비율 0.005 실험 결과
Fig. 2. re-Batch ratio 0.005 experiment result

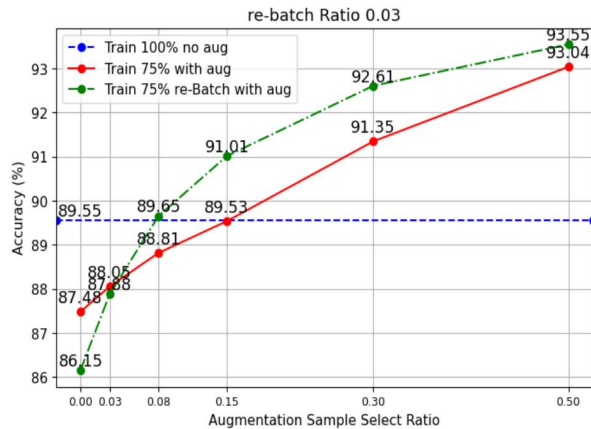


그림 3. re-Batch 비율 0.03 실험 결과
Fig. 3. re-Batch ratio 0.03 experiment result

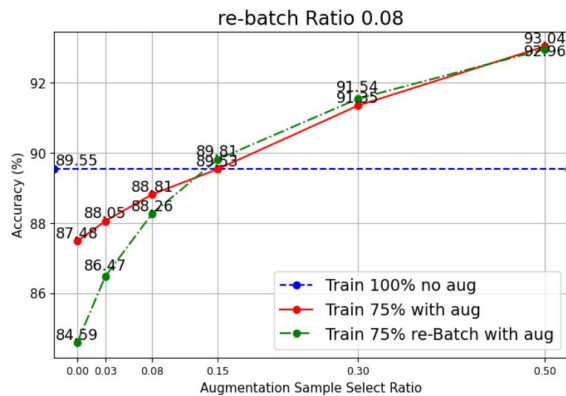


그림 4. re-Batch 비율 0.08 실험 결과
Fig. 4. re-Batch ratio 0.08 experiment result

Yang, Ye, Shen 그리고 Zhou가 제안한 기존 기법 역시 데이터 선택과 증강을 다루고 있으나 사전 학습된 멀티모달 모델(CLIP)의 미세조정 등 많은 컴퓨팅 자원을 요구한다. 본 연구가 제안하는 파이프

라인의 핵심 목적은 외부 모델이나 복잡한 밀도 계산 없이 오직 단일 손실값 지표만 사용하는데 있다. 따라서 선행 연구들과 직접적인 정량 비교는 실험의 범위를 벗어나는 것으로 판단하여 제외하였다.

제안한 기법은 동적 re-Batch 비율 0.005의 거의 모든 증강 선택 비율 구간에서 동일한 데이터 양(75%)을 사용하여 단순히 무작위 증강한 결과(No re-Batch)보다 일관되게 높은 정확도를 기록했다. 특히 증강 선택 비율 0에서 0.3 사이의 구간에서 단순 무작위 증강 대비 성능 향상 폭이 가장 두드러지게 나타났다. 구체적으로 전체 데이터의 75%만 사용하더라도 동적 re-Batch 비율을 0.005으로 설정하고 데이터의 0.08만 증강했을 때, 89.77%의 정확도를 기록하였으며, 증강 없이 전체 데이터(100%)를 모두 사용한 기존 학습의 성능 89.55%를 상회하는 성과를 거두었다. 결과가 우연적 오차인지 확인하기 위해, 10회 반복 실험 결과에 대해 t-검정을 수행하였다. 검정 결과, 기준 대비 $p\text{-value} < 0.05$ ($p=0.048$) 수준에서 통계적으로 유의미한 성능 향상을 달성했음을 확인하였다. 이는 비교적 적은 비율의 데이터 증강만으로도 최적의 re-Batch 스케줄링이 결합된다면, 학습 데이터가 부족한 조건에서도 자원을 효율적으로 사용하며 동시에 모델의 일반화 성능을 전체 데이터 학습 이상으로 끌어올릴 수 있음을 증명한다.

V. 결론 및 향후 과제

본 연구에서는 학습 데이터가 부족한 조건에서 학습 성능을 개선하기 위해 최소 손실 기반의 동적 re-Batch 기법과 이미지 데이터 증강을 결합한 새로운 학습 파이프라인을 제안하고, CIFAR-10 데이터 세트를 통해 그 성능을 실험적으로 검증하였다. 실증 분석 결과, 제안 기법은 데이터의 75%만 사용하고 그중 일부만을 증강하는 조건 속에서도 100% 전체 데이터를 사용한 기존 학습 방식을 상회하는 우수한 성능을 보여주었다. 이는 학습 초기부터 손실값이 작은 샘플들을 학습 데이터세트에 일정한 간격으로 분배하여 수렴 안정성을 확보하는 동적 re-Batch의 메커니즘이 무작위 증강으로 인한 노이즈가 수렴 안정성을 훼손하지 않고 모델의 일반화 성능 향상으로 직결되도록 최적의 학습 환경을 간

접적으로 조성하는 효과를 보인 것으로 해석된다. 특히 사전 학습된 멀티모달 모델의 도입 없이 단일 손실값 지표만을 활용하여 추가 연산 비용을 거의 발생시키지 않는다는 점에서 실용성이 매우 높다.

향후 연구에서는 75% 뿐만아니라 다양한 데이터 부족 상황에서 실험할 계획이다. 동적 re-Batch 비율이 증가하면서 성능이 낮아지는 현상도 향후 연구에서 정확한 원인 파악이 필요하다. 소량의 데이터로 고도의 정확도를 요구하는 의료 영상 분석, 자율주행 객체 인식, 위성 이미지 처리 등 데이터 부족 및 라벨링 비용 문제가 심각한 실제 산업에서 제안 방법론의 범용성과 실증적 가치를 더 검증할 계획이다.

References

- [1] Z. Wang, et al., "A Comprehensive Survey on Data Augmentation", IEEE Transactions on Knowledge and Data Engineering, Vol. 38, No. 1, pp. 47-66, Jan. 2026. <https://doi.org/10.1109/TKDE.2025.3622600>.
- [2] S. Lee and Kang, T, "Dynamic re-Batch for Neural Network Training", The Journal of Korean Institute of Information Technology, Vol. 23, No. 12, pp. 79-88, Dec. 2025. <https://doi.org/10.14801/jkiit.2025.23.12.79>.
- [3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25", NIPS 2012, Lake Tahoe, Nevada, USA, pp. 1097-1105, Dec. 2012.
- [4] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis", Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings, Edinburgh, Scotland, UK, pp. 958-963, Aug. 2023. <https://doi.org/10.1109/ICDAR.2003.1227801>.
- [5] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization", International Conference on Learning Representations, Vancouver, BC, Canada, pp. 1-13, Apr.-May 2018. <https://doi.org/10.48550/arXiv.1710.09412>.
- [6] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features", 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 6022-6031, Oct.-Nov. 2019. <https://doi.org/10.1109/ICCV.2019.00612>.
- [7] H.-P. Chou, S.-C. Chang, J.-Y. Pan, W. Wei, and D.-C. Juan, "Remix: Rebalanced Mixup", Computer Vision - ECCV 2020 Workshops (ECCV 2020), Vol. 12540, pp. 95-110, Aug. 2020. https://doi.org/10.1007/978-3-030-65414-6_9.
- [8] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning Augmentation Strategies From Data", 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, pp. 113-123, Jun. 2019. <https://doi.org/10.1109/CVPR.2019.00020>.
- [9] S. G. Müller and F. Hutter, "TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation", Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, pp. 774-782, Oct. 2021. <https://doi.org/10.48550/arXiv.2103.10158>.
- [10] B. B. Moser, A. S. Shanbhag, S. Frolov, F. Raue, J. Folz, and A. Dengel, "A Coreset Selection of Coreset Selection Literature: Introduction and Recent Advances", arXiv preprint, Jan.2026.<https://doi.org/10.48550/arXiv.2505.17799>
- [11] P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe, "Curriculum Learning: A Survey", International Journal of Computer Vision, Vol. 130, No. 19, pp. 1526-1565, Jun. 2022. <https://doi.org/10.1007/s11263-022-01611-x>.
- [12] A. Shrivastava, A. Gupta, and R. Girshick, "Training Region-based Object Detectors with Online Hard Example Mining", Proc. IEEE Conf.

on Computer Vision Pattern Recognition(CVPR),
Las Vegas, USA, p. 761-769, Jun. 2016.
<https://doi.org/10.48550/arXiv.1604.03540>.

[13] S. Yang, P. Ye, F. Shen, and D. Zhou, "When
Dynamic Data Selection Meets Data Augmentation:
Achieving Enhanced Training Acceleration", arXiv
preprint, arXiv:2505.03809, May 2025.
<https://doi.org/10.48550/arXiv.2505.03809>.

[14] Z. Qin, et al., "InfoBatch: Lossless Training Speed
Up by Unbiased Dynamic Data Pruning", The
Twelfth International Conference on Learning
Representations, Vienna, Austria, pp. 1424-1445,
May 2024. <https://doi.org/10.48550/arXiv.2303.04947>.

[15] A. Krizhevsky, "Learning multiple layers of
features from tiny images", Technical Report,
University of Toronto, pp. 1-60, Apr. 2009.

저자소개

이 석 원 (Seokwon Lee)



2020년 2월 : 강릉원주대학교
컴퓨터공학과(학사)
2020년 2월 ~ 2021년 7월 :
(주)케이에스씨엔씨 전임
2021년 7월 ~ 2025년 4월 :
(주)디엑스솔루션즈 팀장
2025년 3월 ~ 현재 : 강원대학교

원주캠퍼스 컴퓨터공학과 산업대학원 석사과정
관심분야 : 복잡계, 인공지능, 인공지능

강 태 원 (Taewon Kang)



1985년 2월 : 연세대학교, 서울
수학과(이학사)
1988년 2월 : 고려대학교, 서울
전산과학과(이학사)
1991년 2월 : 고려대학교, 서울
수학과 (이학석사)
1996년 8월 : 고려대학교, 서울

컴퓨터학과(이학박사)
1997년 3월 ~ 현재 : 강원대학교 원주캠퍼스
컴퓨터공학과 교수
관심분야 : 복잡계, 인공지능, 인공지능, 소프트웨어