

일반논문 (Regular Paper)

방송공학회논문지 제28권 제3호, 2023년 5월 (JBE Vol.28, No.3, May 2023)

<https://doi.org/10.5909/JBE.2023.28.3.302>

ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

자율주행 영상데이터의 신뢰도 향상을 위한 AI모델 기반 데이터 자동 정제

김 가 나^{a)}, 김 학 일^{a)†}

AI Model-Based Automated Data Cleaning for Reliable Autonomous Driving Image Datasets

Kana Kim^{a)} and Hakil Kim^{a)†}

요 약

본 연구는 과학기술정보통신부가 2017년부터 1조원 이상을 투자한 'AI Hub 댐' 사업에서 구축된 인공지능 모델 학습데이터의 품질 관리를 자동화할 수 있는 프레임워크의 개발을 목표로 한다. 자율주행 개발에 사용되는 AI 모델 학습에는 다량의 고품질의 데이터가 필요하며, 가공된 데이터를 검수자가 데이터 자체의 이상을 검수하고 유효함을 증명하는 데는 여전히 어려움이 있으며 오류가 있는 데이터로 학습된 모델은 실제 상황에서 큰 문제를 야기할 수 있다. 본 논문에서는 이상 데이터를 제거하는 신뢰할 수 있는 데이터셋 정제 프레임워크를 통해 모델의 인식 성능을 향상시키는 전략을 소개한다. 제안하는 방법은 인공지능 학습용 데이터 품질관리 가이드라인의 지표를 기반으로 설계되었다. 한국정보화진흥원의 AI Hub을 통해 공개된 자율주행 데이터셋에 대한 실험을 통해 프레임워크의 유효성을 증명하였고, 이상 데이터가 제거된 신뢰할 수 있는 데이터셋으로 재구축될 수 있음을 확인하였다.

Abstract

This paper aims to develop a framework that can fully automate the quality management of training data used in large-scale Artificial Intelligence (AI) models built by the Ministry of Science and ICT (MSIT) in the 'AI Hub Data Dam' project, which has invested more than 1 trillion won since 2017. Autonomous driving technology using AI has achieved excellent performance through many studies, but it requires a large amount of high-quality data to train the model. Moreover, it is still difficult for humans to directly inspect the processed data and prove it is valid, and a model trained with erroneous data can cause fatal problems in real life. This paper presents a dataset reconstruction framework that removes abnormal data from the constructed dataset and introduces strategies to improve the performance of AI models by reconstructing them into a reliable dataset to increase the efficiency of model training. The framework's validity was verified through an experiment on the autonomous driving dataset published through the AI Hub of the National Information Society Agency (NIA). As a result, it was confirmed that it could be rebuilt as a reliable dataset from which abnormal data has been removed.

Keyword : Reliable Dataset, Dataset Cleaning, Dataset Quality, Autonomous Vehicle, Dataset Reconstruction

Copyright © 2023 Korean Institute of Broadcast and Media Engineers. All rights reserved.

“This is an Open-Access article distributed under the terms of the Creative Commons BY-NC-ND (<http://creativecommons.org/licenses/by-nc-nd/3.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited and not altered.”

1. 서론

수년의 걸친 자율주행 기술의 고도화로 현재는 차량 스스로 일부 운전 작업을 수행할 수 있지만, 여전히 운전자의 주의가 필요하다. 완전 자율 주행자동차를 안전하게 도입하기 위해서는 많은 과제가 필요한데, 많은 전문가들이 가장 중요한 요소로 다량의 고품질 데이터를 꼽으며, 가공된 데이터는 AI 모델을 학습하는데 매우 중요하다. 데이터를 수집하고 데이터를 인공지능 모델 학습에 사용할 수 있게 라벨링 하는 가공 과정 뿐만 아니라, 고품질의 인공지능 학습용 데이터셋을 위하여 데이터 구축 후의 품질 관리에 대한 적합한 기준 및 방법이 필요하다. 과학기술정보통신부(MSIT)가 주최하고 한국지능정보사회진흥원(NIA)이 주관하는 ‘인공지능 학습용 데이터 구축사업’을 통해 구축된 데이터셋의 품질을 위한 ‘인공지능 학습용 데이터 품질관리 가이드라인 및 구축 안내서’^[1]에 해당 기준이 명시 되어있다. 데이터셋의 품질을 확보하는 것은, 사용자에게 유용한 가치를 줄 수 있는 수준에 도달하는 것을 의미하며, 가이드라인에는 데이터 품질을 확보하는데 필요한 전문가, 절차, 기준 등이 있다. 사람 및 기계가 라벨링한 데이터에 대한 검증과 편향 및 오류가 있는 데이터에 대한 교정은 필수 적이지만 품질을 관리하고 검사하는 방법은 여전히 품질 검사자가 육안으로 데이터의 오류 등을 확인하거나, 체크리스트 등을 이용하여 수행되며 이는 비용과 시간 소모적이다.

AI 모델의 성능 및 신뢰성을 향상시키기 위해서는 모델이 더 좋은 구조를 갖게끔 수정하거나 더 효율적인 연산을 할 수 있게끔 하는 모델 중심 접근 방식 또한 중요하지만 이러한 모델을 학습시키는데 사용되는 데이터 중심 접근 방식에 집중하는 것 또한 중요하다. 라벨링 데이터(labeled

data) 중 라벨에 오류 및 오차가 있는 데이터, 학습에 방해가 되는 데이터를 걸러내거나 재라벨링(re-labeling)하는 방식으로 모델성능을 향상시키는 방법에는 의료 이미지 의미적 분할(semantic segmentation) 데이터에 대해서 노이즈 라벨(noisy label)을 자동으로 걸러내고 학습에 효율성을 증진시키는 방법^[2], 학습과정에서 노이즈 라벨의 영향을 최소화하여 학습하는 방법^[3-6], 데이터의 재정립을 반복하며 학습되는 모델을 이용하여 노이즈 라벨에 우선순위를 부여하여 재라벨링 하는 방법^[7-10]과 구조화되지 않은 데이터셋에 대해 클래스별 밀도와의 상관관계를 분석하여 데이터셋의 품질을 개선하는 방법^[11] 등이 있다. 또한 데이터 중심 알고리즘을 이용하여 데이터에 변형을 가해 모델이 노이즈에 더욱 강건하게 만드는 방법^[12-14]과 같이 구축된 데이터셋에 대해 데이터를 수정하거나 삭제하는 것이 아닌 데이터셋 자체에 변형을 가해 학습의 효율을 높이는 방법이 있다. 이처럼 데이터셋에 대한 잠재력을 확인하고 모델 성능을 향상시키는 방법보다 본질적으로 데이터셋이 학습에 유효하게 사용될 수 있는지에 대한 연구는 이미지 분류(image classification) 데이터셋에 대해 다양한 합성곱 신경망(Convolutional Neural Networks, CNN)을 사용하여 건축 자재 데이터셋의 품질과 효율성을 검증하는 방법을 제시하여 모델의 인식성능을 향상시키는 연구^[15]가 있다. 하지만 이전의 연구들은 모델 성능 향상을 위해 데이터셋 자체를 자동으로 고치는 것이 아닌, 학습 과정에서 문제가 있는 데이터를 사용하지 않거나, 사람이 개입하여 오류 데이터를 재 라벨링하여 사용한다^[16,17]. 또한 데이터셋의 품질을 검증하는 연구는 이미지 분류에서는 진행되고 있지만, 여러 객체가 장면을 구성하는 객체 인식 데이터에 대해 검증하는 연구는 많지 않다.

본 논문에서는 자율 주행 데이터셋의 이상데이터를 제거하는 데이터셋 정제 프레임워크를 소개하며 이상 데이터 제거에 사용되는 점수를 계산하기 위해 여러 검출 모델을 사용하여 교차 검증을 진행한다. 이상치 제거를 통한 데이터셋 정제 목표치 달성을 위해 해당 프레임워크를 반복하여 수행한다. 이상데이터가 제거된 데이터셋으로 평가 모델을 반복 학습시키며, 인공지능 모델의 성능 향상 및 고품질 데이터셋을 위한 정제를 진행한다. 정제 프레임워크를

a) 인하대학교 전기컴퓨터공학과(Department of Electrical and Computer Engineering, Inha University)

‡ Corresponding Author : 김학일(Hakil Kim)

E-mail: hikim@inha.ac.kr

Tel: +82-32-720-9085

ORCID: <https://orcid.org/0000-0003-4232-3804>

* 이 논문은 2023년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원 (P0017124, 2023년 산업혁신인재성장지원사업)의 지원을 받아 수행된 연구임.

· Manuscript March 22, 2023; Revised April 13, 2023; Accepted April 13, 2023.

통해 MSIT와 NIA의 인공지능 데이터셋 품질 관리 가이드라인에서 언급하는 구축 단계에서의 품질검사 중 전수 검사를 통해 검출될 수 있는 라벨링 오류, 혹은 라벨링 누락에 해당하는 데이터를 확인할 수 있다.

II. 방법

1. 전체 프로세스

본 연구에서 제안하는 데이터셋 정제 프레임워크는 데이터셋에서 부정확, 불완전하거나 손상되거나 중복된 데이터를 감지하고 식별한 다음 삭제하는 과정이다. 부정확하고 불완전한 데이터를 노이즈 라벨 (Noisy Label)로 정의하고 데이터 품질을 확인하는 지표를 이용하여 노이즈 라벨을 자동으로 감지하고자 한다. 낮은 지표는 저 품질 데이터를 의미하며, 사용할 지표는 과학기술정보통신부와 한국정보화진흥원의 인공지능 학습용 데이터 품질 가이드라인^[1]에서의 데이터 정확성 확인 지표 중 의미 정확성 지표의 정의를 활용한다. 프로세스는 크게 모델 학습(Model Training), 점수 앙상블(Score Ensemble), 필터링(Filtering), 정제(Cleaning)의 네가지 솔루션으로 구성된다. 사용한 데이터셋은 AI hub^[18]에 공개된 데이터셋 중 자율 주행 2차원 객체 탐지 데이터를 사용한다.

가이드라인에 따르면 사업관리기관에서는 인공지능 데이터셋의 품질 보장을 위해, 사업 수행 기관이 실시한 품질 관리 활동을 객관적으로 확인하는 제3자 품질 검증 활동을 수행한다. 품질 검사 방법으로 사람의 육안으로 데이터 오류, 손상을 확인하는 육안검사와 정제도구, 체크리스트를 사용하는 도구 검사가 있다. 하지만 이러한 도구를 사용하는 검사 방법조차 검수자가 개입하여 수행하여야 한다. 검수자가 검수할 수 있는 데이터 양은 정해져 있으며, 이는 비용과 시간 소모적이며, 구축된 전체 데이터셋에 대해 검수하기 쉽지 않다. 또한 가공 과정 뿐만 아니라 검수 과정도 편향이 발생할 수 있기 때문에, 검수가 완료된 데이터도 범용성이 떨어질 수 있다. 따라서 이러한 오류 데이터를 감지하여 삭제하는 과정을 여러 인공지능 모델을 이용하여 자동으로 수행하고자 한다. 데이터 정확성 품질 검사 기준으

로 의미 정확성은 95%이상을 권고하며, 이를 정제 프로세스의 목표치로 사용한다.

그림 1은 정제 프레임워크의 알고리즘 흐름도이다. 데이터셋 정제에는 점수(Score)를 정의하여 노이즈 라벨을 감지하고, 점수 임계치(Score Threshold)를 계산하여 데이터셋을 구성하는 각 데이터에 대한 점수를 구하여 임계치와 비교를 통해 노이즈 라벨로 정의하고, 해당 데이터를 삭제한다. 데이터셋 정제 프레임워크는 정제 달성도 (Cleaning Achievement)를 계산하여 노이즈 라벨 삭제 프로세스의 반복 혹은 종료를 결정한다. 점수(Score)는 데이터셋을 구성하는 데이터 중에 오태깅 혹은 미태깅 데이터를 감지하기 위해 본 연구에서 새로이 정의하는 점수이며, 가이드라인에 정의된 데이터 정확성 확인 지표 중 의미 정확성 지표 중 일부를 결합하여 사용한다. 의미 정확성 지표 중 정확도, 정밀도, 재현율 지표를 수치화 하여 결합하여 점수로 정의한다. 데이터의 품질 검수 시 검수자가 육안으로 확인하는 것이 아닌, 라벨링 점수를 통해 자동으로 수행할 수 있게 하기 위해 프레임 내 객체를 인식하는 여러 2차원 객체 검출 모델의 결과를 정답으로 가정하여 라벨링 결과(Ground Truth, GT)를 검사한다. 여러 모델을 사용하는 이유는 모델의 예측 결과에 신뢰도를 높이기 위함이다. 단일 결과가 아닌 여러 의견을 결합함으로써, 모델 결과를 일반화하여 라벨링 데이터를 검사하는 정답 지로서 작용하게 한다.

점수를 도출하여 정제를 진행하기 위한 솔루션 별 수행 방법은 다음과 같다. 첫 번째 솔루션인 모델 학습 단계에서, 여러 개의 기본적인 합성곱 신경망을 노이즈 라벨을 찾을 대상 데이터셋으로 학습하여 사용한다. 학습된 모델을 이용해 각자의 데이터셋을 교차검증을 해야 하므로, 학습에 이용되는 전체 데이터셋을 정제 모델 (Cleaning model)의 개수만큼 데이터 분량을 나눈다. 학습된 모델을 이용하여 각자의 학습 데이터셋을 추론하여 결과를 도출한다. 도출한 결과는 두 번째 솔루션인 점수 앙상블에서 교차결합되어 점수 임계치를 구하는데 사용되며, 세 번째 솔루션에서는 프레임에 대한 필터링 점수가 계산된다. 네 번째 솔루션에서 점수 임계치와 필터링 점수의 비교를 통해 노이즈 라벨이 삭제되며 각 반복 단계에서 정제 달성도가 구해진다.

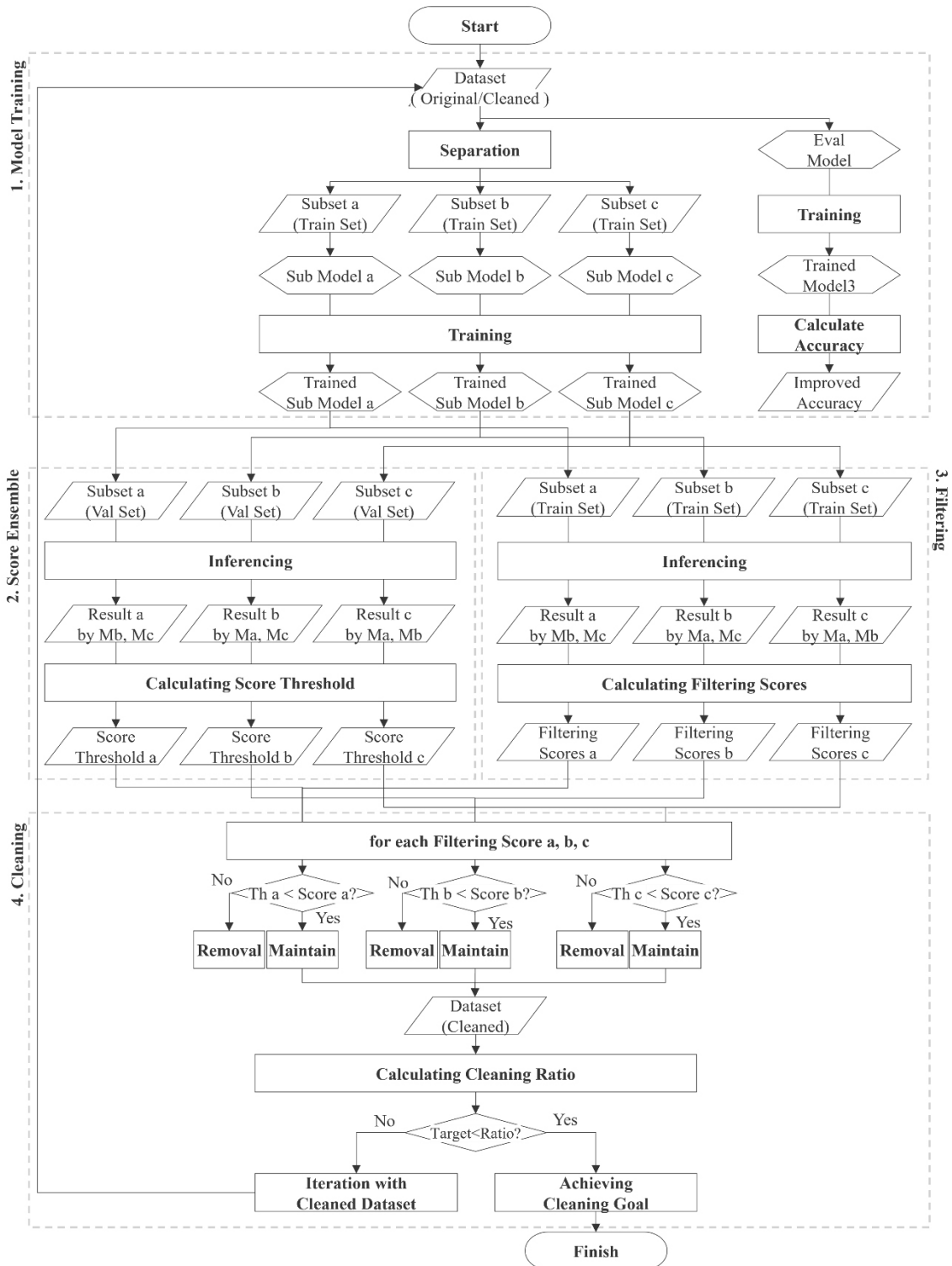


그림 1. 본 논문에서 제안하는 데이터셋 정제 프레임워크의 알고리즘 흐름도
 Fig. 1. Algorithmic flowchart of the dataset cleaning framework proposed in this paper

2. 모델 학습

주어진 데이터의 품질 점수의 계산에 MSIT와 NIA의 ‘인공지능 학습용 데이터 품질관리 가이드라인 및 구축 안내서’^[11]에서 데이터 가공프로세스에서의 품질 관리를 위한 주요 검사 항목을 수치화 적용하였다. 해당 항목을 검수자가 육안으로 확인하는 것이 아닌, 정제 프레임워크를 통해 자동으로 수행할 수 있게 하기 위해, 프레임 내 객체를 인식하여 결과를 제공하여 데이터의 품질 점수를 계산할 수 있는 2차원 객체 검출 AI 모델을 학습하여 사용한다. 데이터 별 점수를 계산하여 검수하는 과정에 모델의 추론 결과를 답지로서 라벨링 데이터의 오류와 누락을 확인한다. 이 때, 단일 모델을 사용하는 것은 신뢰도가 떨어질 수 있다. 따라서 단일 데이터에 대해서 여러 의견을 결합하여 모델의 추론 결과를 일반화하여 사용하기 위하여 교차 검증을 진행한다. 교차 검증을 위해 비슷한 성능을 가지는 다른 모델 3개를 학습하여 사람이 아닌 AI 모델이 데이터에 대해 검증하는 것에 대한 신뢰도를 높이는 것이 목표이다.

여러 개의 인공지능 모델을 학습해야 하므로, 학습의 효율성을 위해 Pytorch 기반의 객체 검출 학습 라이브러리인 Detectron2^[19]를 사용한다. Detectron2의 model zoo에 공개된 비슷한 성능을 도출한 Faster R-CNN 기반 R101-C4, R101-DC5, R101-FPN을 정제 모델 a, b, c (Cleaning model a, b, c)로서 학습시키고, RPN과 Fast R-CNN 기반의 RPN R50-FPN 모델을 평가 모델(Evaluation model)로서 학습시킨다. 학습된 정제 모델은 점수 앙상블(Score Ensemble)에 사용되며, 평가 모델은 재 구축 과정이 반복(iteration)될 때 마다 데이터셋에 대한 성능 평가를 진행할 때 사용된다. 각 정제 모델 a, b, c를 학습에 사용되는 데이터셋은 서브 셋 a, b, c(Subset a, b, c)로 분리하며 학습 데이터셋(training set)과 검증 데이터셋(validation set)의 비율은 8:2로 나누어 사용한다.

3. 점수 앙상블

제안하는 프레임워크에서는, 오류나 누락이 있는 데이터를 자동으로 탐지하기 위해 각 데이터에 대한 지표를 결합하여 점수를 도출한다. 결합된 점수의 일반화를 위해 교차

검증 개념이 도입되는데, 이는 학습된 모델이 자신의 검증 데이터셋을 검사하지 않도록 하여 계산되는 점수에 대한 편향을 방지하기 위함이다. 즉, 데이터를 필터링 하기 전, 필터링에 사용되는 점수 임계치에 대한 기준의 중립을 지켜 신뢰도를 높이기 위해 여러 모델의 데이터에 대한 점수를 결합하는 것이다. 서브셋 a, b, c 로 나누어서 학습된 정제 모델 a, b, c은 각자의 서브셋의 검증 데이터셋에 대해서 추론을 진행하여 점수 임계치를 결정하는데 사용된다.

Average Score: 각 서브셋의 검증 데이터셋에 대한 점수의 평균값에 대한 계산이다. 점수는 각 프레임에 대한 품질에 대한 수치이며, 데이터 품질관리 가이드라인에서 2차원 객체 검출 데이터의 품질을 확인하기 위한 주요 검사 항목을 반영한다. 검사 항목 중 바운딩 박스 정확도를 반영하기 위해 IoU Score를 계산하며, 라벨 정확도를 위해 Confidence Score를 반영한다. 점수(Score)는 식 (1)과 같이 구해지며 각 프레임에 대해서 모델의 예측 정확도인 confidence score, $Conf = Pr(Object)$ 와 프레임 내의 예측된 바운딩 박스와 실제 정답(ground truth)와의 겹치는 정도인 Intersection over Union(IoU)을 곱한 값이며, $IoU = \text{area}(box_{gt} \cap box_p) / \text{area}(box_{gt} \cup box_p)$ 정의된다. 학습된 모델로 추론된 데이터셋을 표현하기 위해 f_{subset} 을 정제 모델로 정의하였으며, θ_{sub} 는 정제 모델 f_{subset} 의 학습에 사용되는 매개 변수, N_{subset} 는 서브 셋을 구성하는 모든 데이터의 수이다. 학습된 f_{subset} 로 서브셋의 검증 데이터셋을 추론한 결과는 $\tilde{d}_{subset} = f_{subset}(\theta_{subset}, d_{subset})$ 로 표현한다. 점수 임계치 계산을 위한 IoU_{th} 는 0.0 초과의 값부터 설정하고, IoU값이 임계 값을 초과하는 예측된 바운딩 박스의 개수를 N_{exceed} 로 정의한다. 점수의 평균을 구한 $AvgScore_{subset}$ 는 검증 데이터셋을 구성하는 프레임안에서 IoU임계값이 넘는 바운딩 박스들로 점수를 계산하고, 각 프레임에 대한 점수를 검증 데이터셋을 구성하는 프레임의 수로 나누어 평균 점수(Average Score)를 계산하며, 식(2)와 같이 표현할 수 있다.

$$Score_d = \sum_{box_p \in \tilde{d}_{subset}} \begin{cases} Conf \times IoU, & \text{if } IoU_{box} > IoU_{th} \\ 0, & \text{else} \end{cases} \quad (1)$$

$$AvgScore_{subset} = \frac{1}{N_{subset}} \sum_1^{N_{subset}} \frac{1}{N_{exceed}} Score_d \quad (2)$$

Score Threshold: 식(3)은 한 서브 정제 모델 자신을 제외한 다른 서브 정제 모델의 평균 점수를 평균 낸 값인 $Score Th_{s_{subset}}$ 으로, 서로 다른 모델 결과와의 결합을 통해 점수 임계치에 대한 신뢰도를 확보할 수 있다.

$$Score Th_{s_{subset}} = \frac{1}{N_{s_{subsets}} - 1} \sum_{s \in \{a, b, \dots, N_{s_{subsets}}\}, s \neq s_{subset}} AvgScore_s \quad (3)$$

4. 필터링

학습된 모델에 사용된 학습 데이터셋을 구성하는 데이터에 대해 구조가 다른 모델들의 예측 결과를 결합하여 교차검증을 통해 신뢰도를 높이는 것을 필터링 솔루션이라 한다. 점수 양상불 솔루션과 계산하는 방식은 같지만, 계산 대상이 다르다. 필터링은 각 정제 모델 학습에 사용된 학습 데이터셋을 구성하는 각 데이터에 대해서 점수를 부여하여 학습된 모델의 추론 결과를 이용하여 학습 과정에서 방해가 됐을 데이터를 자동으로 걸러내기 위한 점수를 구한다. 가이드라인에서는 해당 과정을 점수자가 랜덤 데이터를 추출하여, 정답지를 생성하고 정답과 가공된 데이터를 비교하여 오류 데이터를 찾아내었다면, 필터링 과정에서 도출되는 필터링 점수를 통해 해당 과정을 사람 개입없이 자동으로 수행할 수 있다. 모델의 단일 추론 결과는 결과를 신뢰할 수 없고, 편향이 발생할 수 있기 때문에, 단일 데이터에 대해 여러 모델의 결과를 결합한 필터링 점수를 계산한다. 필터링 점수는 자기 자신을 제외한 다른 서브셋의 학습 데이터셋으로 학습된 모델을 이용하여 추론을 진행하여 각각의 데이터에 대한 지표들을 결합하여 계산한다.

Frame Score: 필터링 점수를 구하기 위한 프레임 점수 (Frame Score)는 점수 양상불 솔루션에서 구한 점수와 같은 방식으로 각 프레임에 대한 confidence score와 IoU가 결합된 $Frame Score_{s_{subset}^{s_{subset}'}}$ 로써 표현되며 식(4)와 같이 구해진다. 여기서 표현되는 $\widetilde{d}_{s_{subseta}}^{s_{subsetb}} = f_{s_{subsetb}}(\theta_{s_{subsetb}}, d_{s_{subseta}})$ 는 서브 정제 모델 b로 추론한 서브셋 a의 학습 데이터셋의 결과이다. 프레임 점수 계산 시에는 가이드라인의 데이터 품질 검사 항목 중 데이터 정확성 중 재현율을 확인

하기 위해 가공 데이터와 예측한 클래스가 다를 때에는 점수에 가중치(weight)를 낮게 적용하는 것이 특징이며, 프레임 점수 계산을 위한 IoU 임계치는 0.3이상부터 매 프레임워크 반복마다 임계치를 늘린다. 재현율은 프레임을 구성하는 객체에 대한 라벨 중 참 값 라벨을 확인하는 수치이다.

$$Frame Score_{s_{subset}^{s_{subset}'}} = \frac{1}{N_{exceed}} \sum_{box_p \in \widetilde{d}_{s_{subset}^{s_{subset}'}}} \begin{cases} Score, & \text{if } IoU_{box_p} > IoU_{th} \\ 0, & \text{else} \end{cases} \quad (4)$$

Filtering Score: 필터링 점수는 해당 프레임 점수의 변별력 및 신뢰도를 높이기 위해 다른 서브셋에 대한 프레임 점수도 결합하여 평균 낸 값을 필터링 점수로 식(5)와 같이 구해진다.

$$Filtering Score_{s_{subset}} = \frac{1}{N_{s_{subset}^{-1}}} \sum_{s \in \{a, b, \dots, N_{s_{subset}}\}, s \neq s_{subset}} Frame Score_{s_{subset}^{s_{subset}'}} \quad (5)$$

5. 정제

세번째 솔루션인 필터링에서 구해진 각 데이터에 대한 필터링 점수와 점수 임계치를 비교하여 이상이 있는 데이터를 자동으로 분류하고 삭제하여 데이터셋을 정제하는 정제 솔루션이다. 필터링 솔루션에서 구한 필터링 점수가 점수 임계치의 이상인지 아닌지의 따라 다음 정제 프레임워크 반복 단계에서 모델 학습시에 사용하지 않을지를 결정한다. 사용하지 않을 데이터로 결정된 데이터는 라벨링 오류 혹은 라벨링 누락 데이터들이 분류된다. 라벨링 누락 데이터는 학습한 모델이 예측한 결과와 가공 데이터(ground truth)를 비교하여 예측 모델이 높은 confidence score를 가지고 학습 데이터셋의 데이터에서 객체를 탐지했음에도 불구하고, 가공 데이터가 라벨링이 되어있지 않은 경우를 분류해 낸다. 라벨링 오류 데이터는 모델의 예측 결과와 가공 데이터와의 품질 점수와 정확한 클래스 인지의 결과를 조합한 값인 필터링 점수가 임계치 보다 현저히 낮을 경우 라벨링 오류 데이터로 분류된다. 즉 모델이 예측한 결과와 가공 데이터의 결과가 차이가 심하게 나는 경우

$$\text{for all } d_{subset_i} \text{ in } N_{subset} \begin{cases} \text{Maintain,} & \text{if } \text{Filtering Score}_{subset_i} \geq \text{Filtering Score Th}_{subset} \\ \text{Remove,} & \text{else} \end{cases} \quad (6)$$

이다. 위의 경우들을 고려하여 위 식(6)에 나타나는 조건에 따라 각각의 서브셋의 학습 데이터셋에 대해서 남길 데이터와 삭제할 데이터를 결정하고 데이터셋을 정제 및 정리한다.

Cleaning Ratio: 정제 비율(Cleaning Ratio)는 매 정제 반복마다 기존 데이터셋과 정제된(Cleaned) 데이터셋의 비율을 계산한 것이며 아래 식(7)과 같이 표현된다. 목표치에 달성할 때까지 정제 프레임워크를 반복한다.

$$\text{Cleaning Ratio} = \frac{N_{cleaned}}{N_{original}} \times 100(\%) \quad (7)$$

III. 결과

제안하는 재구축 프레임워크를 통해 재 구축된 데이터셋의 효과를 실험하기 위해 평가 모델을 따로 학습시켜 사용하였다. 아래 표 1은 프레임 워크 반복을 5번 진행한 후의

평가 모델의 검증 데이터셋에 대한 성능의 결과인 정확도 (Accuracy)와 학습에 사용된 데이터의 수, 정제 비율 (Cleaning Ratio)이다. 성능은 AR¹⁰⁰⁰값으로 계산되며 프레임워크의 반복 시 이상 데이터가 제거되기 때문에, 학습데이터의 수가 줄어든 것을 확인할 수 있다.

표1. Cleaning Framework을 수행한 평가 모델의 결과
Table 1. Results of Evaluation model that performed Cleaning Framework

Iteration Number	Accuracy (AR ¹⁰⁰⁰)	Number of Training set	Cleaning Ratio (%)
1	53.72	207,068	92:8
2	53.95	192,442	92:8
3	53.94	178,464	92:8
4	53.98	164,809	93:7
5	54.01	153,421	-

도출된 결과에 따르면 데이터 수가 줄어들어도 객체 인식 성능은 비슷하거나 조금 향상되었다. 또한 제안하는 프



그림 2. 정제 프레임워크를 통해 걸러진 라벨링 오류 데이터의 예시 1
Fig. 2. Example of miss labeled data filtered through the Cleaning framework

레이아웃은 오류 데이터를 제거하여 신뢰할 수 있는 정제된 데이터셋을 도출하는 것으로, 아래 그림 2부터 그림 6에서 정제 프레임워크를 통해 어떤 데이터가 정제되었는지를 확인할 수 있다. 데이터 품질 관리 가이드라인에서 구축 과

정에서의 품질 검수 후의 계산될 수 있는 오대강율 (정상 라벨링 데이터 중 라벨링 오류 데이터의 비율), 미태강율 (정상 라벨링 데이터 중 라벨링 누락 데이터의 비율) 데이터는 본 연구를 통해 제거대상이 되는 데이터에서도 확인



그림 3. 정제 프레임워크를 통해 걸러진 라벨링 오류 데이터의 예시 2
 Fig. 3. Example of miss labeled data filtered through the Cleaning framework

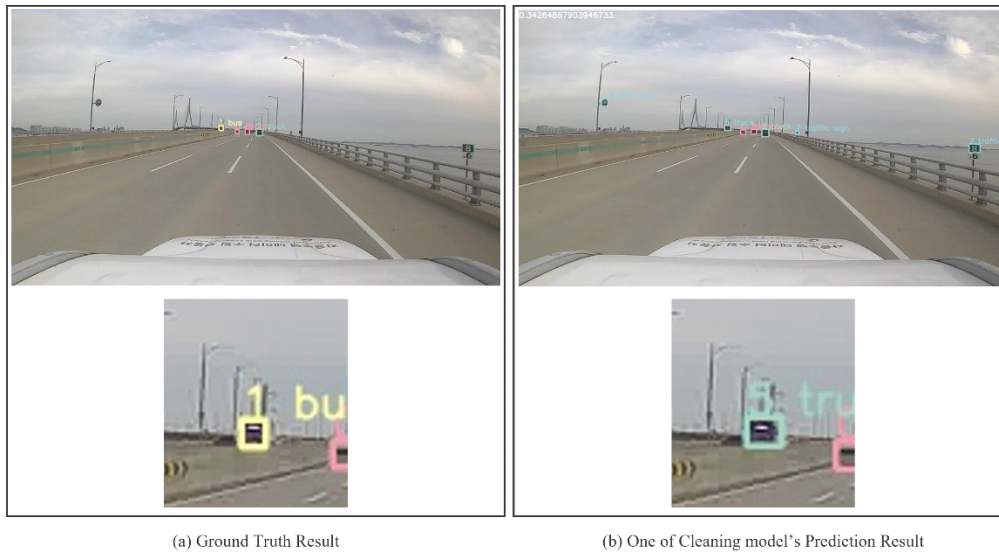


그림 4. 정제 프레임워크를 통해 걸러진 라벨링 오류 데이터의 예시 3
 Fig. 4. Example of miss labeled data filtered through the Cleaning framework

할 수 있다. 삭제되는 데이터 중 그림 2에서 a는 가공되어 구축된 라벨링 데이터 즉, 학습에 사용되는 참 값 데이터이며, b는 정제 모델 중 하나의 학습된 모델이 해당 데이터에 대해 예측한 결과이다. 그림 2-a 와 같이 사람 주석자

(human annotator)의 라벨링 가공 후 저장 단계에서 라벨 지정오류로 인한 오류가 있으며, 그림 2-b처럼 학습된 정제 모델을 이용한 예측 결과와의 비교로 해당 데이터가 오류 데이터로 걸러졌다.

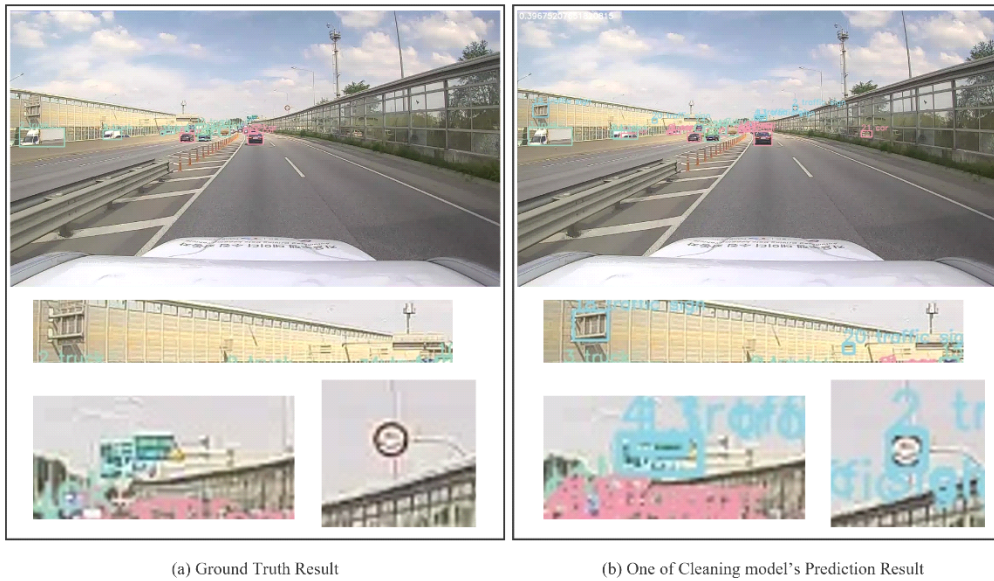


그림 5. 정제 프레임워크를 통해 걸러진 라벨링 누락 데이터의 예시 1
 Fig. 5. Example of data that lack label filtered through the Cleaning framework



그림 6. 정제 프레임워크를 통해 걸러진 라벨링 누락 데이터의 예시 2
 Fig. 6. Example of data that lack label filtered through the Cleaning framework

그림 3 또한 그림 2와 마찬가지로, 프레임을 구성하는 객체에 대한 라벨링 오류로 데이터가 걸러졌으며, 교통 표지판(Traffic Sign) 객체가 버스(Bus) 객체로 라벨링 되어있거나, 차량(Car) 객체가 자전거(Bicycle) 객체로 라벨링 되어 있는 것을 확인할 수 있다. 그림 4는 2,3과 같은 라벨링 오류로 걸러진 데이터이다. 하지만 그림2와 그림 3이 데이터 저장과정에서의 오류였다면 그림 4는 트럭(Truck) 객체를 버스 객체로 착각한 실제 사람 주석자의 인식 오류로 인한 라벨링 오류를 걸러낸 것을 확인할 수 있다.

그림 5와 6은 프레임을 구성하는 객체에 대해 사람 주석자가 라벨링을 하지 않아 걸러진 데이터이다. 이런 누락 라벨이 있는 데이터는 사람이 인식하기에도 어려운 데이터이기 때문에 해당 데이터는 모델을 학습하는 과정에서 노이즈로 작용하게 된다. 따라서 라벨 누락데이터는 학습과정에서 모델에 혼란을 야기하거나 학습 중 loss를 구하는 과정에서 방해가 될 수 있으므로, 데이터 정제과정을 통하여 정제하는 것이 중요하다.

IV. 고찰

본 연구는 Level 5 자율 주행 시스템 개발의 가장 중요한 요소로 기술 개발에 필요한 고품질 데이터를 구축하기 위해, 자율 주행 데이터셋의 오류를 제거하는 데이터셋 정제 프레임워크를 제안하며, 사람이 직접 검증하는 시간적 비용을 줄이고, 편향된 데이터의 교정을 통해 고품질데이터를 재 정립할 수 있다. 본 연구의 결과로 다량의 데이터뿐만 아니라 품질의 데이터도 모델 학습에 중요함을 증명하였고, 이상이 있는 데이터를 자동으로 걸러낼 수 있음을 확인하였다.

초기 정제 모델 학습 시에는 오류가 있는 데이터가 사용될 수 있어 필터링 및 정제에 사용되는 모델의 예측 결과를 사용하여 계산되는 점수를 완전히 신뢰할 수 없다. 하지만 여러 모델의 결과를 결합하여 사용하기 때문에 이러한 단점을 보완할 수 있다. 또한 점수 계산에는 confidence score와 IoU의 값이 결합되는데, 이때 IoU 즉 참값(ground truth)와 예측 바운딩 박스(bounding bo)의 겹치는 정도가 작으면 점수가 낮아지는데 이는 작은 객체에 좀 더 민감하게 작용

되기 때문에 이를 보완할 수 있는 방법이 필요해 보인다.

다른 연구[2]에서는 데이터셋의 클래스 밀도를 정의하여 클래스별 정확도와와의 상관관계를 분석하여 데이터셋의 품질을 개선하였지만, 이는 데이터셋 내의 모든 프레임에 대한 각각의 분석 결과를 제공하지는 않는다. 또 다른 연구[3]에서는 여러 모델을 사용한 교차 검증을 통해 이상 데이터를 제거하여 데이터셋의 품질을 향상시켰지만, 데이터셋 제거에 사용한 기준은 confidence score 하나로만 판단하였으며, 범용성이 좋은 데이터셋이 아닌 이미지 분류 데이터셋을 이용한 실험이며 정확한 평가 지표를 제시하지 않았다.

본 연구를 통해 데이터셋 정제에 단일 모델을 사용하지 않고 여러 모델의 교차 검증을 통해 데이터셋 정제의 기준에 대한 신뢰성을 확보하였으며 학습에 방해가 되는 이상 데이터를 자동으로 분류 후 정제된 데이터셋으로 학습된 모델의 인식 성능이 좋아지는 것을 확인하였다.

V. 결론

본 연구에서 제안한 데이터셋 정제 프레임워크를 통해 국내 및 해외에 공개된 데이터셋에 대해 오류가 있거나 학습에 해가 되는 데이터를 자동으로 걸러내어 학습의 효율성을 높여 인공지능 모델의 대한 오류를 줄일 수 있다. 과학기술정보통신부와 한국지능정보 사회진흥원의 ‘인공지능 학습용 데이터 품질관리 가이드라인’에 따라 품질 평가 항목으로 사용되는 정확도, 정밀도 등의 수치를 모델 학습결과로 구하여 사용하기 때문에 자율 주행 데이터셋 뿐만 아닌 모든 객체 탐지 데이터셋에 대해 해당 프레임워크를 적용시킬 수 있는 범용성이 있다.

참 고 문 헌 (References)

- [1] Ministry of Science and ICT, National Information Society Agency, "Data Quality Management Guidelines and Construction Guidelines for AI Learning v3.0," Vol.1 Quality Management Guidelines. doi : <https://aihub.or.kr/aihubnews/qlityguidance>, 2023.
- [2] H. Zhu, J. Shi, J. Wu, "Pick and Learn: Automatic Quality Evaluation for Noisy-labeled Image Segmentation," International Conference on

- Medical Image Computing and Computer-Assisted Intervention, pp. 576-584, 2019.
doi: http://dx.doi.org/10.1007/978-3-030-32226-7_64
- [3] Y. Kim, J.M. Kim, Z. Akata, J. Lee, "Large Loss Matters in Weakly Supervised Multi-Label Classification," Computer Vision and Pattern Recognition Conference, pp. 14156 - 14165, 2022.
doi: <https://doi.org/10.1109/CVPR52688.2022.01376>
- [4] J. Li, R. Socher, and S. C. Hoi, "DivideMix: Learning with noisy labels as semi-supervised learning," International Conference on Learning Representations, pp. 1 - 14, 2020.
doi: <https://doi.org/10.1109/IJCNN55064.2022.9892927>
- [5] A. Ghosh, N. Manwani, and P. Sastry, "On the robustness of decision tree learning under label noise," The Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 685 - 697, 2017.
doi: https://doi.org/10.1007/978-3-319-57454-7_53
- [6] V. Mnih and G. E. Hinton, "Learning to label aerial images from noisy data," International Conference on Machine Learning, pp. 567 - 574, 2012.
- [7] M. Bernhardt, D.C. Castro, R. Tanno, A. Schwaighofer, K.C. Tezcan, M. Monteiro, S. Bannur, M.P. Lungren, A. Nori, B. Glocker and J. Alvarez-Valle, "Active label cleaning for improved dataset quality under resource constraints", Nature communications, 13(1), pp.1-11, 2022.
doi: <https://doi.org/10.1038/s41467-022-28818-3>
- [8] O Sener, S Savarese, "Active learning for convolutional neural networks: A core-set approach", International Conference on Learning Representations, 2018.
doi: <https://doi.org/10.48550/arXiv.1708.00489>
- [9] G. Contardo, L. Denoyer and T. Artieres, "A meta-learning approach to one-step active-learning" in arXiv:1706.08334, 2017.
doi: <https://doi.org/10.48550/arXiv.1706.08334>
- [10] P. Bachman, A. Sordoni and A. Trischler, "Learning algorithms for active learning," in arXiv:1708.00088, 2017.
doi: <https://doi.org/10.48550/arXiv.1708.00088>
- [11] A. Byerly and T. Kalganova, "Class Density and Dataset Quality in High-Dimensional, Unstructured Data," arXiv preprint arXiv:2202.03856, 2022.
doi: <https://doi.org/10.48550/arXiv.2202.03856>
- [12] Y. Zhong, L. Wu, X. Liu and J. Jiang, "Exploiting the Potential of Datasets: A Data-Centric Approach for Model Robustness," arXiv preprint arXiv:2203.05323, 2022.
doi: <https://doi.org/10.48550/arXiv.2203.05323>
- [13] D. Gamberger, N. Lavrac, and S. Dzeroski, "Noise detection and elimination in data preprocessing: Experiments in medical domains," Applied Artificial Intelligence, vol. 14, no. 2, pp. 205 - 223, 2000.
doi: <http://dx.doi.org/10.1080/088395100117124>
- [14] T. Liu, D. Tao, "Classification with noisy labels by importance reweighting," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 3, pp. 447 - 461, 2015.
doi: <http://dx.doi.org/10.1109/TPAMI.2015.2456899>
- [15] Y. Sun, Z. Gu, "Using computer vision to recognize construction material: A Trustworthy Dataset Perspective," Resources, Conservation and Recycling, 183, p.106362, 2022.
doi: <http://dx.doi.org/10.1016/j.resconrec.2022.106362>
- [16] H. Song, M. Kim, D. Park, J. Lee, "Learning from noisy labels with deep neural networks: A survey," arXiv preprint arXiv:2007.08199, 2021.
doi: <http://dx.doi.org/10.1109/TNNLS.2022.3152527>
- [17] B. Settles, "Active learning literature survey," *Computer Science Technical Report* 1648, University of Wisconsin-Madison, January 2010.
- [18] Korea National Information Society Agency, AI Hub, <https://aihub.or.kr>
- [19] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo and R. Girshick, "Detectron2", 2019. <https://github.com/facebookresearch/detectron2>

— 저 자 소 개 —



김 가 나

- 2021년 8월 : 인하대학교 정보통신공학과 학사
- 2021년 9월 ~ 현재 : 인하대학교 전기컴퓨터공학과 석사과정
- ORCID : <https://orcid.org/0009-0001-8866-7162>
- 주 관심분야 : 인공지능, 자율주행, 컴퓨터비전

저 자 소 개



김 학 일

- 1983년 2월 : 서울대학교 제어계측공학과 학사
- 1985년 8월 : Purdue 대학교 전기컴퓨터공학 석사
- 1990년 8월 : Purdue 대학교 전기컴퓨터공학 박사
- 1990년 9월 ~ 현재 : 인하대학교 스마트모빌리티공학과 교수
- ORCID : <https://orcid.org/0000-0003-4232-3804>
- 주 관심분야 : 패턴인식, 컴퓨터비전, 자율주행