KIPS Trans. Softw. and Data Eng.
Vol.10, No.9 pp.375~384
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2021.10.9.375

Deep Learning Model Validation Method Based on Image Data Feature Coverage

Chang-Nam Lim† · Ye-Seul Park†† · Jung-Won Lee†††

ABSTRACT

Deep learning techniques have been proven to have high performance in image processing and are applied in various fields. The most widely used methods for validating a deep learning model include a holdout verification method, a k-fold cross verification method, and a bootstrap method. These legacy methods consider the balance of the ratio between classes in the process of dividing the data set, but do not consider the ratio of various features that exist within the same class. If these features are not considered, verification results may be biased toward some features. Therefore, we propose a deep learning model validation method based on data feature coverage for image classification by improving the legacy methods. The proposed technique proposes a data feature coverage that can be measured numerically how much the training data set for training and validation of the deep learning model and the evaluation data set reflects the features of the entire data set. In this method, the data set can be divided by ensuring coverage to include all features of the entire data set, and the evaluation result of the model can be analyzed in units of feature clusters. As a result, by providing feature cluster information for the evaluation result of the trained model, feature information of data that affects the trained model can be provided.

Keywords : Deep Learning, Coverage Testing, Image Feature Extraction, Validation Method, Dataset Splitting Method

영상 데이터 특징 커버리지 기반 딥러닝 모델 검증 기법

임 창 남† · 박 예 슬†† · 이 정 원†††

요 약

딥러닝 기법은 영상 처리 분야에서 높은 성능을 입증 받아 다양한 분야에서 적용되고 있다. 이러한 딥러닝 모델의 검증에 가장 널리 사용되는 방법으로는 홀드아웃 검증 방법, k-겹 교차 검증 방법, 부트스트랩 방법 등이 있다. 이러한 기존의 기법들은 데이터 셋을 분할하는 과정에서 클래스 간의 비율에 대한 균형을 고려하지만, 같은 클래스 내에서도 존재하는 다양한 특징들의 비율은 고려하지 않고 있다. 이러한 특징들을 고려하지 않을 경우, 일부 특징에 편향된 검증 결과를 얻게 될 수 있다. 따라서 본 논문에서는 기존 검증 방법들을 개선하여 영상 분류를 위한 데이터 특징 커버리지 기반의 딥러닝 모델 검증 기법을 제안한다. 제안하는 기법은 딥러닝 모델의 학습과 검증을 위한 훈련 데이터 셋과 평가 데이터 셋이 전체 데이터 셋의 특징을 얼마나 반영하고 있는지 수치로 측정할 수 있는 데이터 특징 커버리지를 제안한다. 이러한 방식은 전체 데이터 셋의 특징을 모두 포함하도록 커버리지를 보장하여 데이터 셋을 분할할 수 있고, 모델의 평가 결과를 생성한 특징 군집 단위로 분석할 수 있다. 검증 결과, 훈련 데이터 셋의 데이터 특징 커버리지가 낮아질 경우, 모델이 특정 특징에 편향되게 학습하여 모델의 성능이 낮아지며, Fashion-MNIST의 경우 정확도가 8.9%까지 차이나는 것을 확인하였다.

키워드 : 딥러닝, 모델 테스팅, 영상 특징 추출, 검증 기법, 데이터 셋 분할 기법

1. 서 론

딥러닝 기법은 영상 처리 분야에서 높은 성능을 입증 받아 의료[1,2], 교통[3], 자동차[4], IoT[5] 등 다양한 분야에서 적용되고 있다. 이러한 분야에서는 다양한 형태의 데이터를 각 분야에 맞추어 학습함으로써 견고한 네트워크 모델을 설계하고 학습한다. 이러한 학습 모델의 성능을 결정하는 요소는 여러 가지가 있지만, 최근에는 학습된 모델에 대한 성능을 개선시키는 것뿐만 아니라 확보된 성능에 대한 체계적인 검증 역시 중요해지고 있다.

전체 데이터를 모두 학습에 사용하는 경우, 가장 좋은 성능을 보이는 분류기를 학습할 수 있으나, 학습에 사용되지 않은 새로운 데이터에 대한 분류기의 성능을 추정할 수 없다. 이러한 학습된 영상 분류기의 성능을 추정하는 것은 사용 단계에서의 정확도를 예측하는 것을 의미하며, 학습한 분류기의 선택 및 결합에 활용될 수 있으므로 중요하다. 따라서 일반적으

※ 본 논문은 산업통상자원부 및 한국산업기술진흥원의 창의산업기술개발기반 구축사업의 일환으로 수행하였음(N0002312, 디지털 헬스케어 소프트웨어 시험평가센터 구축).
※ 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1A2C1007400).
† 준 회 원 : 아주대학교 AI융합네트워크학과 석사
†† 준 회 원 : 아주대학교 AI융합네트워크학과 박사과정
††† 종신회원 : 아주대학교 전자공학과/AI융합네트워크학과 교수
Manuscript Received : June 7, 2021
Accepted : July 4, 2021
* Corresponding Author : Jung-Won Lee(jungwony@ajou.ac.kr)

로 모델의 성능을 추정하기 위해서는 전체 데이터 셋을 분류기의 학습을 위한 훈련 데이터 셋과 평가 데이터 셋으로 분할하며, 평가 데이터 셋을 활용하여 모델의 신뢰성을 검증한다.

딥러닝 모델의 검증을 위해 많이 쓰이는 방법으로는 홀드아웃 검증 방법이나 k-겹 교차 검증 방법, 부트스트랩 방법 등이 있다[6]. 기존의 방법들은 데이터 셋을 서로 겹치지 않는 부분 집합으로 분할하여, 모델의 학습과 평가를 진행한다. 또한, 학습 데이터의 불균형을 막기 위해 데이터 셋을 분할하는 과정에서 층화추출법을 이용해 각 클래스 간의 비율을 유지한다. 하지만 이러한 불균형의 문제는 클래스 간의 비율에만 존재하는 것이 아니다. 같은 클래스 내의 데이터에서도 영상의 구도, 색 분포, 무늬, 형태 등과 같은 다양한 특징들의 차이가 존재할 수 있으며, 이러한 특징들의 균형 또한 고려하여 데이터 셋을 분할해야 한다.

그러나 기존의 검증 방법들은 이러한 특징에 대한 고려 없이 무작위 추출을 이용하여 데이터를 분할하기 때문에, 분할된 평가 데이터 셋이 모델 학습에 사용된 훈련 데이터 셋의 분포에 대해 좋은 커버리지를 갖는지 알 수 없다[7]. 이로 인해 훈련 데이터 셋과 평가 데이터 셋 사이의 데이터 특징 불균형이 심해질 경우 모델이 편향되어 학습되거나 평가될 수 있으며, 분류기의 성능에 악영향을 끼치거나 분류기의 평가 결과에 대한 신뢰성을 떨어뜨릴 수 있다.

이러한 기존의 문제점을 해결하기 위해서 본 논문에서는 데이터 특징 커버리지를 기반으로 한 딥러닝 모델 검증 기법을 제안한다. 제안하는 기법은 학습하고자 하는 전체 데이터에 대하여 특징 벡터를 추출하고, 각각의 클래스별로 추출한 특징을 군집화하여 특징 군집을 생성한다. 이때, 학습하고자 하는 영상 데이터 셋의 특성에 따라 다양한 특징 추출 기법을 적용할 수 있다. 예를 들어, 기존에 널리 사용되던 SIFT[8], HOG[9], Haar[10], Color Moment[11], 등의 전통적인 특징을 사용하거나, 최근 제안되고 있는 합성곱 신경망을 이용한 특징 추출 방법[12-14]으로 추출한 특징을 사용하여 특징 군집을 생성할 수 있다. 이후, 생성한 특징 군집별로 층화추출법을 적용한 홀드아웃 검증 방법이나 k-겹 교차 검증 방법을 통해 데이터 셋을 분할하고, 분할 된 데이터 셋을 이용하여 모델을 학습 및 검증한다. 또한, 학습된 모델을 평가할 때, 특징 군집별로 성능을 측정하여 기존의 검증 방법보다 세밀한 평가를 진행할 수 있다.

본 연구에서 제안하는 방법을 세 종류의 오픈 영상 데이터 셋(Fashion-MNIST[15], CIFAR-10[16])을 이용하여 실험하였다. 구체적으로는 각 데이터 셋에 대해 제안하는 데이터 특징 커버리지를 다르게 하여 분할한 데이터 셋을 이용해 모델을 학습 및 평가하였고, 평가 결과를 비교함으로써 제안하는 데이터 특징 커버리지의 효용성을 검증하였다. 검증 결과, 훈련 데이터 셋의 데이터 특징 커버리지가 낮아질 경우, 모델이 특정 특징에 편향되게 학습하여 모델의 성능이 낮아지며, Fashion-MNIST의 경우 정확도가 8.9%까지 차이나는 것을 확인하였다.

2. 관련 연구

2.1 기존 딥러닝 모델 검증 방법

학습된 분류기의 정확도를 추정하는 것은 사용 단계에서의 정확도를 예측하는 것뿐만 아니라, 분류기의 선택 및 결합에도 매우 중요하다[6]. 이러한 정확도 추정을 위한 방법으로 가장 널리 사용되는 것이 홀드아웃 검증 방법과 k-겹 교차 검증 방법이다. 홀드아웃 검증 방법과 k-겹 교차 검증 방법은 Fig. 1과 같다.

홀드아웃 검증 기법은 데이터 셋을 모델을 학습하는 부분과 학습된 모델을 평가하는 부분으로 나누는 방법이다[6]. 보통, 전체 데이터 셋의 2/3를 훈련 세트로 1/3을 평가 데이터 셋으로 나누며, 6:4, 7:3, 8:2, 9:1 등의 비율로 나누기도 한다. Fig. 1에서 홀드아웃 검증 부분의 예시는 전체 데이터 셋의 2/3를 훈련 데이터 셋으로, 1/3을 평가 데이터 셋으로 나눈 것을 표현하였다. 모델의 학습은 훈련 데이터 셋을 이용하여 이루어지며, 평가 데이터 셋은 사용 단계에서 입력될 데이터에 대한 모델의 성능을 추정하기 위해 사용한다. 또한 홀드아웃 검증 기법은 평가를 위해 일부 데이터를 분리해두기 때문에, 전체 데이터에 대해 학습하지 못해 모델의 성능에 대해 비관적으로 예측하는 경향을 보인다. 또한, 평가 데이터 셋에 사용하기 위해 많은 데이터를 분리하면 학습한 모델에 대한 성능 추정의 비관적인 경향이 증가할 수 있고, 적은 양의 데이터를 분리하면 성능 추정 값의 신뢰구간이 커지게 되는 트레이드오프(Trade-off)가 존재한다. 학습된 모델의 분류 결과는 분류 성공과 실패의 두 가지로 나뉘는 베르누이 시행으로 볼 수 있으며, 평가 데이터 셋을 이용해 분류한 결과들의 합은 베르누이 시행의 합으로 볼 수 있다. 특히, 분류기의 정확도에 대한 신뢰구간과 분산을 계산할 수 있다[6].

k-겹 교차 검증은 전체 데이터를 k개의 동등한 부분으로 나누어, k-1개의 부분으로 학습을 진행하고 나머지 하나의 부분으로 학습된 모델을 평가하는 방법이다[17]. 또한, 이렇게 분할된 k개의 데이터 셋에 모두 한 번씩 평가 데이터 셋으로 사용하기 위해 k번 반복하여 모델을 학습 및 평가한다. Fig. 1에서 k-겹 교차 검증 부분의 예시는 k를 5로 하였을 때, 각 부분을 평가 데이터 셋으로 하는 5개의 분할 데이터 셋이 생성됨을 나타낸다. 이러한 방법은 모델의 학습과 평가

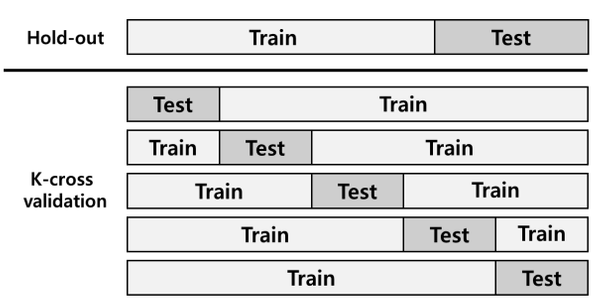

Fig. 1. Example of Verification Method Through Dataset Splitting (Hold-out, K-cross Validation)

에 모든 데이터를 사용할 수 있다는 장점이 있으며, 이러한 장점으로 인해 데이터의 개수가 적은 경우에 많이 사용된다. 또한, k의 크기가 작을 경우 학습에 사용되는 데이터의 개수가 적어지므로 성능의 추정 값의 비관적 편향이 증가되며, k가 커질 경우 연산의 비용이 증가하며, 성능 추정의 분산이 증가하게 된다.

이외에도 부트스트랩, 몬테칼로 교차검증, 부트스트랩 라틴 분할(Bootstrapped Latin Partition)등의 데이터 셋 분할기법이 존재한다. 논문 [18]에 따르면 이러한 기법 및 매개변수에 따른 명확한 우위는 없으며, 사용하는 데이터에 따라 선택하여 사용해야 한다.

2.2 영상 데이터 특징 추출 방법

본 논문에서 제안하는 기법은 특징 군집을 생성하는데, 이때 학습하는 데이터 셋의 특징을 다양한 방식으로 추출하여 군집을 형성할 수 있다. 대표적인 영상 특징으로는 SIFT[8], HOG[9], Haar[10], Ferns[19], LBP[20], MCT[21] 등이 있다. 이러한 전통적인 특징 추출 기법 외에도 여러 특징 추출 기법이 연구되고 있는데, 최근의 몇 연구에서 ImageNet 데이터 셋에 대해 학습된 합성곱 신경망 모델을 이용하여 특징을 추출하는 방법이 제안되었다[12-14]. 제안된 기법은 ImageNet ILSVRC 2013 데이터 셋의 영상 분류를 위해 학습된 분류기의 합성곱 계층을 이용하여 영상의 특징을 추출하는 기법이다. 여러 연구에서 미리 학습된 합성곱 신경망이 다양한 분야에서 영상 특징 추출기로 사용될 수 있으며, 영상 분류 작업에서 뛰어난 성능을 낼 수 있음을 보여주고 있다.

2.3 딥러닝 모델 커버리지 테스팅

최근에 딥러닝 모델의 데이터 특징 커버리지를 체계적으로 다루고자 하는 시도가 있었다. [7]에서는 데이터 셋의 특징을 이용한 딥러닝 모델의 커버리지 테스팅 기법을 제안하였다. 제안된 기법은 다음과 같은 절차로 이루어진다. 먼저, 전체 데이터 셋을 훈련 데이터 셋과 평가 데이터 셋으로 분할한 후, 합성곱 신경망 모델을 학습한다. 그리고 학습된 합성곱 신경망 모델의 합성곱 계층을 이용하여 평가 데이터 셋의 특징을 추출한 후, 평가 데이터 셋에 대해 동등 분할(Equivalence Partitioning), 중심 위치(Centroid Positioning), 경계 조건(Boundary Conditioning) 그리고 페어와이즈 경계 조건(Pair-wise Boundary Conditioning)의 네 가지 품질 척도를 계산한다. 이후, 특징 차원에서 클래스 간 경계 부분의 평가 데이터들을 추가로 생성하고, 생성된 평가 데이터들과 기존의 평가 데이터를 이용하여 학습된 모델을 평가한다. 또한, 생성된 평가 데이터들을 확인하기 위해 역합성곱망에 생성한 평가 데이터를 입력하여 영상을 획득한다. 그러나 제안된 방법은 특징 공간에서 생성된 데이터가 실제로 어떠한 영상으로부터 추출될 수 있는 데이터인지, 해당 클래스에 속하는 데이터인지에 대한 검증을 수행하기 어려우며, 훈련 데이터 셋이 모든 특징 공간에 분포되어 있는지는 평가할 수 없다. 따라서 본 논문에서는 훈련 및 평가 데이터 셋이 전체 데이터 셋의 특징을 고르게 반영하였는지 평가할 수 있는 데이터 특징 커버리지를 제안한다.

3. 특징 커버리지 기반 학습 모델 검증 기법

본 논문에서는 영상 분류를 위한 데이터 특징 커버리지 기반 딥러닝 모델 검증 기법을 제안한다. 제안하는 방식은 특징 군집을 생성하고 데이터 셋을 분할하여 데이터 특징 커버리지를 계산하여 모델의 성능과 비교하는 방식이다.

3.1 영상 데이터 특징 벡터 추출

본 절에서는 영상 데이터의 특징 벡터를 추출하는 기법을 소개하고자 한다. 영상 데이터 특징이란 관찰되는 프로세스의 개별 측정 가능한 속성이다[22]. 가장 흔한 시각적 특징들에는 색상, 질감 그리고 형태 등이 포함된다[23]. 색상은 영상의 가장 중요한 특징 중 하나로, 색상 특징은 색 공간 또는 모델에 따라 정의된다[23]. 색상 히스토그램(Color Histogram)[24], 색상 모멘트(Color Moment)[11], 색상 일관성 벡터(Color Coherence Vector)[25], 색상 코렐로그램(Color Correlogram) [26] 등 여러 특징들이 존재한다. 최근에는 사전 학습된 심층 합성곱 신경망(Deep Convolutional Neural Network)을 이용하여 추출한 특징을 사용한다. 사전 학습된 깊은 합성곱 신경망은 최근 여러 연구에서 범용 이미지 설명자로 사용되어 영상 분류 작업에서 뛰어난 성능을 낼 수 있음을 보여주고 있다[12-14]. 본 연구에서는 이와 같은 사전 학습된 심층 합성곱 신경망 모델을 활용하여 영상 벡터 추출을 수행하였으며, 세부적인 설명은 Fig. 2와 같다.

3.2 k-평균 군집화 기반 특징 군집 생성

본 절에서는 3.1절에서 추출한 특징을 군집화하여 특징 군집을 생성하는 방법을 소개한다. 군집화의 목표는 새로운 범주의 집합을 발견하는 것으로, 새로운 그룹은 그 자체로 의미가 있

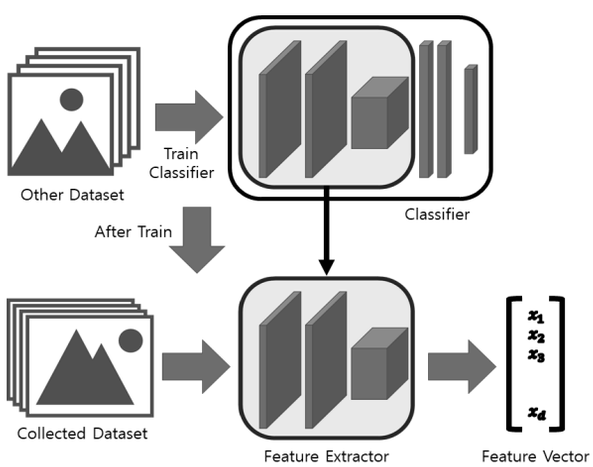

Fig. 2. Feature Extraction Using CNN

다[27]. 군집화는 유사한 데이터가 함께 군집화되고, 서로 다른 데이터가 다른 군집에 속하게 하는 방식으로 데이터를 하위 집합으로 군집화 한다[27]. 이러한 군집화 알고리즘으로 k-평균 군집화, 퍼지(Fuzzy) 군집화, 계층적 군집 분석, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) 등이 있다.

본 연구에서는 다양한 군집화 알고리즘 중에서 k-평균 군집화 알고리즘을 사용하였다. k-평균 군집화 알고리즘은 단순하고, 처리 속도가 빠르기 때문에 실제 응용에서 많이 사용되며 좋은 군집 결과를 생성할 수 있는 매우 효과적인 방법이다[28]. 단, k-평균 알고리즘은 구형의 군집을 생성하고[28], 고차원 데이터에 대한 처리 능력이 떨어진다는 단점이 있다[29]. 이로 인해 추출한 특징의 차원이 높은 경우 군집화 과정의 연산 비용이 커지게 된다. 따라서 본 연구에서는 주성분 분석[30]을 이용하여 추출한 특징 벡터의 차원을 축소하여 군집화를 진행하였다.

주성분 분석은 데이터의 분포에서 분산이 가장 큰 순서대로 방향 벡터를 찾아 주성분을 선정한다. 선정한 주성분의 분산이 클수록, 주성분에 데이터를 투영했을 때 원래 데이터의 분포에 대한 많은 정보를 저장할 수 있다. 이러한 점을 이용해 분산량이 큰 주성분만을 추출하고, 분산량이 작은 주성분은 제거하여 데이터의 차원을 축소할 수 있다. 또한, 차원 축소된 특징 벡터에 대해 k-평균 군집화 알고리즘을 적용하여 특징 군집을 생성한다. k-평균 군집화 알고리즘은 생성하는 군집의 개수 k와 군집화 할 데이터 셋을 입력으로 받으며, 항상 지역 최솟값으로 수렴하고, 수렴하기 전에 일정 횟수 반복하면 반복을 멈추도록 설정할 수 있다[28].

3.3 데이터 특징 커버리지 기반 데이터 셋 분할

3.2절에서 생성된 데이터 특징 군집을 기반으로 본 절에서는 학습 데이터와 검증 데이터를 분할하는 과정을 진행한다. 이 때, 학습에 사용되는 데이터 셋이 전체 데이터 셋의 특징 군집을 얼마나 고르게 포함하는지를 데이터 특징 커버리지로 정의하였으며, 다음과 같은 수식으로 계산된다.

$$Data\,Feature\,Coverage = \frac{1}{N}\sum_{i=1}^{N}\min\left(1, \frac{\left|C_i \bigcap D_d\right|}{\left|C_i\right|}\frac{|D|}{|D_d|}\right)$$

해당 수식에서 N은 전체 클래스에서 생성한 특징 군집의 개수, C_i는 전체 특징 군집 중 i번째 특징 군집에 속한 데이터들의 집합, D는 전체 데이터 셋, 는 D_d분할된 데이터 셋 중 커버리지 측정을 하려는 데이터 셋이다. 제안하는 데이터 특징 커버리지를 통해 분할된 데이터 셋이 전체 특징 군집을 얼마나 포함하는지 나타낼 수 있으며, 데이터 특징 커버리지의 조정을 통해 모든 특징 군집에 대해 학습 혹은 평가가 이루어졌는지 확인할 수 있다.

본 연구에서는 클래스별 특징 커버리지에 따른 세부 분할을 수행하기 위해, 층화추출법을 사용하였다. 층화추출법이란 모집단을 먼저 중복되지 않도록 층으로 나눈 다음 각 층에서 표본을 추출하는 방법이다. Fig. 3은 이러한 층화추출법을 적용한 데이터 셋 분할을 표현한 그림이다.

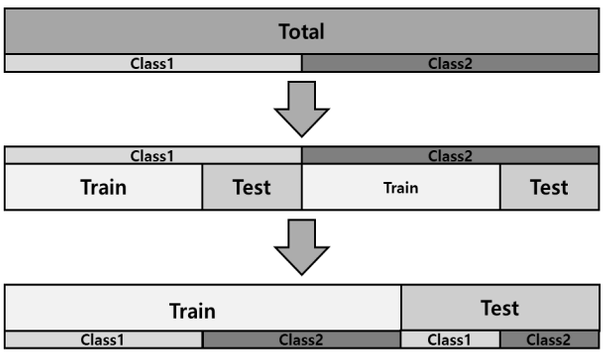

Fig. 3. Example of Dataset Splitting Using Stratified Sampling

Fig. 3에서 2개의 클래스가 존재할 때, 각 클래스의 비율을 유지하기 위해서 각 클래스를 층으로 하여 각각 분할한 후 합치는 것을 표현하였다. 이러한 방법을 통하여 전체 데이터 셋과 클래스 간 비율이 동일한 훈련 데이터 셋 및 평가 데이터 셋을 생성함으로써, 학습 데이터의 불균형을 막을 수 있다. 하지만 이러한 불균형의 문제는 클래스 간의 비율에만 있는 것이 아니다. 같은 클래스 내에서도 영상의 구도, 색상 분포, 질감 등 다양한 특징들의 차이가 존재할 수 있으며, 특징들의 균형 또한 고려하여 데이터 셋을 분할해야 한다.

따라서 본 연구에서는 3.2절에서 생성한 특징 군집에 층화추출법을 적용하여 데이터 셋을 분할하는 방법을 제안한다. 제안하는 방법은 전체 데이터의 특징 추출 및 군집화하고, 생성된 특징 군집에 대한 층화추출을 이용하여 데이터 셋을 분할하는 방법이며, Fig. 4는 한 클래스 내에서 생성한 특징 군집을 층화 추출하는 과정을 나타낸다.

Fig. 4에서 클래스 1에 대해 5개의 특징 군집이 존재하며, 각각의 군집 간 비율을 유지하며 데이터를 분할한 것을 나타낸다. 특징 군집들에 대해 층화추출을 적용하여 데이터 셋을 분할할 경우, 특징 군집 간 비율을 유지하며 데이터 셋을 분할할 수 있다. 이처럼, 분할한 데이터 셋으로 학습 및 검증을 수행하게 되면 모든 특징에 대한 학습과 평가가 가능하다.

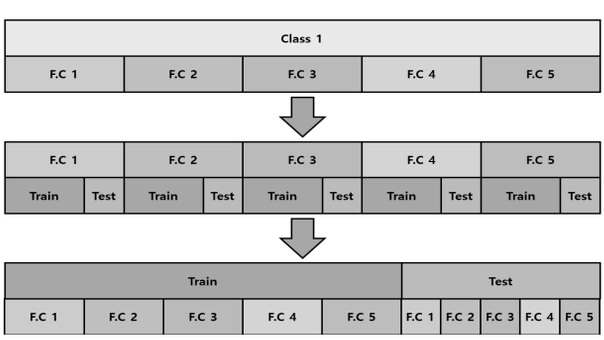

Fig. 4. Example of Feature Cluster Stratified Sampling of Class1 (F.C=Feature Cluster)

또한, 모델을 평가하는 과정에서 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수(F1 score) 같은 평가 지표를 더욱 세분하게 계산하여 모델을 정밀하게 평가할 수 있다. 기존의 검증 방법으로는 전체 모델의 정확도, 정밀도, 재현율, F1 점수와 클래스별 정밀도, 재현율, F1 점수를 계산할 수 있으나, 본 연구에서 제안하는 검증 방법에서는 생성한 특징 군집별 정밀도, 재현율, F1 점수를 계산할 수 있다. 이러한 세분된 평가를 통해 학습된 모델이 각 클래스 내부에서 어떤 특징에 취약한지 분석할 수 있다.

4. 특징 커버리지 기반 모델 학습 성능 분석

4장에서는 데이터 특징 커버리지의 유용성을 증명하기 위해 두 가지 오픈 영상 데이터 셋(Fashion-MNIST[15], CIFAR-10[16]) 학습 과정에 데이터 특징 커버리지를 기반으로 한 성능을 비교 분석한다. 해당 오픈 영상 데이터 셋을 이용하여 실험을 진행한 이유는 서로 다른 도메인에도 제안하는 기법이 적용 가능한지 실험하기 위해 일반화된 데이터를 사용하기 위함이다. 성능 비교 분석을 위한 절차는 아래의 Fig. 5와 같다. 4.1절에서는 ImageNet 데이터 셋에 대해 학습된 DenseNet 121 모델을 이용하여 특징 벡터를 추출한다. 4.2절에서는 주성분 분석을 통해 추출한 특징 벡터의 차원을 축소하고, k-평균 군집화 기법을 이용하여 특징 군집을 생성한다. 4.3절에서는 생성한 특징 군집 정보를 이용하여 모델을 학습하고, 학습된 모델의 성능을 평가한다.

4.1 데이터 셋 특징 추출

본 연구에서는 오픈 영상 데이터 셋인 ImageNet 데이터 셋에 대해 학습된 DenseNet121 모델의 합성곱 계층을 분리하여 특징 추출기로 사용하였다. 사용한 DenseNet121 모델은 최소 입력 크기가 32x32로 제한되어 있어, Fashion-MNIST의 경우 영상의 크기에 의해 입력이 제한되는 문제가 발생한다. 따라서 CIFAR-10의 경우 특징이 충분하게 추출되지 않을 가능성이 있어 영상의 크기를 128x128로 조절하여 특징을 추출하였다. 128x128 크기의 영상을 합성곱 계층에 입력할 경우 1,024차원의 특징 벡터가 추출된다.

4.2 데이터 셋 특징 군집 추출

추출한 특징 벡터는 1,024차원의 높은 차원을 가지며, 이로 인해 특징 군집을 생성하기 전에 주성분 분석 기법을 이용하여 특징 벡터의 차원을 축소하였다. Fig. 6은 각 오픈 영상 데이터 셋의 주성분 분석 결과 각 주성분의 분산 설명량을 나타낸 그림이며, 분산 설명량은 해당 주성분의 정보량이다. 각 그래프 모두 분산 설명량이 1~7차원에서는 매우 급격하게 줄어들다가 8차원 이후 완만하게 줄어드는 것을 볼 수 있다. 분산 설명량이 급격하게 줄어들다가 완만하게 줄어드는 부분을 보통 팔꿈치(elbow)라고 하며, 해당 지점에서 축소하는 차원을 결정하여[31], 두 개의 데이터 셋 모두 10차원으로 특징 벡터를 축소하였다. Fig. 7은 축소한 차원에 따른 분산 설명량의 합을 나타낸 그림이다. Fig. 6과 Fig. 7의 빨간 선은 특징 벡터를 축소한 10차원을 표시한다.

이어서 축소한 특징 벡터에 대해 각각의 클래스별로 k-평균 군집화 알고리즘을 적용하여 특징 군집을 생성하였다. Fig. 8은 각 오픈 영상 데이터 셋에서 각각의 첫 번째 클래스(T-shirt, Airplane, Apple)의 군집 개수 k에 따른 군집 중심점과 데이터 간 거리의 총합을 나타낸 그림이다.

Fig. 8에서 빨간 선은 선택한 군집의 개수인 k가 7일 때를 표시한 것이며, 거리의 총합이 급격하게 줄어들다가 완만하게 줄어드는 부분에서 군집의 개수를 선택하였다[32]. Table 1은 제안하는 방법으로 Fashion-MNIST 데이터 셋에서 생성한 특징 군집의 군집 별 개수를 나타낸다.

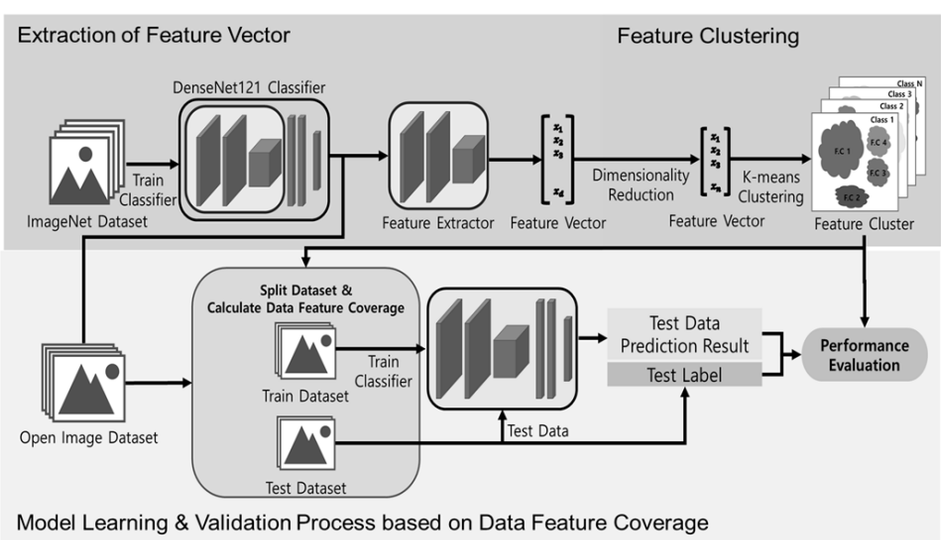

Fig. 5. The Process of Performance Evaluation

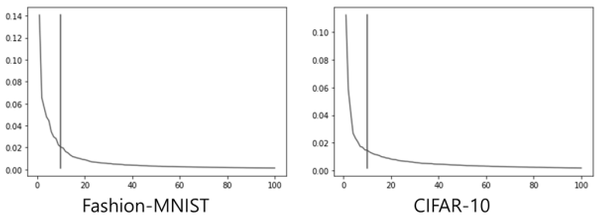

Fig. 6. Principal Component Analysis Result (x: Principal Component y: Variance Explanation Amount)

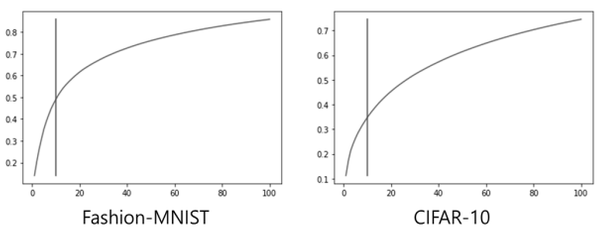

Fig. 7. Analysis of Cumulative Variance Explanation According to the Reduced Dimension

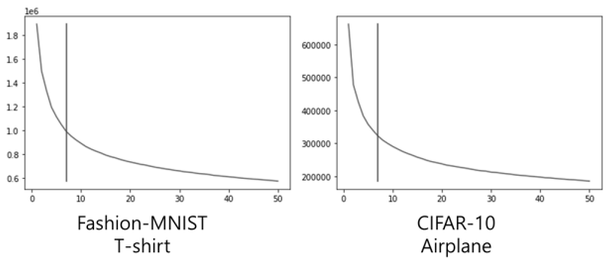

Fig. 8. Analysis of the Distance between Cluster Center Points (x: Number of Clusters k, y: Sum of Distances)

4.3 데이터 특징 커버리지 기반 모델 성능 평가

4.3절에서는 제안하는 기법의 효용성 검증을 위해 오픈 영상 데이터 셋에서 제공하는 분할 데이터 셋을 이용하여 학습 및 평가한 모델과 제안하는 기법을 적용하여 학습 및 평가한 모델을 비교한다. 4.3절에서 진행한 학습에 대한 정보는 다음 Table 2와 같다. 사용된 모델은 모델의 학습을 많이 반복하기 때문에 학습에 걸리는 시간이 짧은 모델을 생성 및 이용

Table 1. Number of Data per Feature Cluster of the Fashion-MNIST Dataset

Dataset	Class	Number of Data in Cluster						
		F.C1	F.C2	F.C3	F.C4	F.C5	F.C6	F.C7
Total Dataset	T-shirt	1339	814	1224	606	950	825	1242
	Trouser	520	982	372	1011	1066	1434	1615
	Pullover	952	1346	1206	1107	833	600	956
	Dress	1027	1278	1094	541	1214	918	928
	Coat	956	1167	830	783	828	1171	1265
	Sandal	2213	1214	1056	537	545	985	447
	Shirt	1118	1234	816	888	1026	1213	705
	Sneaker	776	1170	1117	1029	1305	814	789
	Bag	1293	1235	994	600	633	982	1263
	Ankle Boot	1206	628	639	1494	657	1369	1007
Train Dataset	T-shirt	1139	703	1042	518	815	718	1065
	Trouser	429	852	318	861	903	1240	1397
	Pullover	807	1169	1034	946	711	513	820
	Dress	882	1096	952	465	1033	776	796
	Coat	810	1005	728	667	725	997	1068
	Sandal	1887	1045	916	452	466	840	394
	Shirt	955	1033	705	762	891	1049	605
	Sneaker	666	996	962	882	1122	700	672
	Bag	1103	1076	858	503	541	843	1076
	Ankle Boot	1037	541	544	1300	555	1172	851
Test Dataset	T-shirt	200	111	182	88	135	107	177
	Trouser	91	130	54	150	163	194	218
	Pullover	145	177	172	161	122	87	136
	Dress	145	182	142	76	181	142	132
	Coat	146	162	102	116	103	174	197
	Sandal	329	169	140	85	79	145	53
	Shirt	163	201	111	126	135	164	100
	Sneaker	110	174	155	147	183	114	117
	Bag	190	159	136	97	92	139	187
	Ankle Boot	169	87	95	194	102	197	156

Table 2. Environment of Model Training

Processor	XEON E5-2640V4 * 2 EA	
GPU	NVIDIA GeForce GTX 1080 Ti * 3EA	
RAM	64.0 GB	
Batch Size	32	
Input Image Shape	128 x 128	
Epoch	300	
Optimizer	Stochastic Gradient Descent	
Learning Rate	1~150 epoch	0.1
	151~225 epoch	0.01
	226~300 epoch	0.001
Architecture	SmallNet	

하였으며, SmallNet이라고 명명하였다. 해당 모델은 5개의 합성곱 계층, 2개의 풀링 계층, 2개의 완전연결 계층으로 구성되며, 300 epoch에 약 2시간 정도의 학습 시간이 소요되었다. Fashion-MNIST 및 CIFAR-10 데이터 특징커버리지 분석에 대한 전처리의 경우, 프로세서는 Intel(R) Core(TM) i7-8700K CPU, GPU는 NVIDIA GeForce GTX 1050 Ti의 환경에서 별도로 분석하였으며 시간은 약 20~23분 정도 소요되었다. 본 연구에서는 이와 같은 환경을 기반으로 Fig. 4의 방식을 적용하여 특정 특징 군집을 완전히 배제한 데이터 셋을 생성하였다. 구체적으로는 분류된 층화 추출된 특징 클래스를 훈련 데이터 셋과 평가 데이터 셋에서 아무것도 배제하지 않거나(Number of Excluded Clusters in Train and Test Dataset = 0인 경우), 3개, 5개를 배제한 데이터 셋을 생성하였다. Table 3과 Table 4는 배제된 특징 군집에 따라 두 개의 오픈 영상 데이터 셋에 대해 학습하였을 때 각 모델의 정확도, 정밀도, 재현율, F1 점수, 훈련 데이터 셋의 데이터 특징 커버리지, 평가 데이터 셋의 데이터 특징 커버리지를 계산한 결과이다.

Table 3의 붉은색 음영과 같이 훈련데이터 셋을 하나도 배제하지 않고, 테스트 셋만 0, 3, 5를 배제하였을 때 훈련데이터 커버리지가 떨어지는 것을 확인할 수 있다. 이러한 이유는 평가 데이터 셋에서 하나의 군집을 제외하여도, 학습 과정에서 사용된 훈련 데이터와 평가 데이터의 총 장수는 유지하여야 하므로, 다른 군집들이 평가 데이터에 더 많이 포함되게 되었기 때문이다. 붉은색 음영 구간은 한 군집 내의 데이터 개수는 정해져 있기 때문에 평가 데이터에 더 들어감으로써 훈련 데이터셋에 들어가는 데이터가 적어지게 되고, 이로 인해 커버리지가 조금 떨어지게 된 것을 나타낸다.

Fashion-MNIST 데이터 셋의 평가 결과(정확도, 정밀도, 재현율, F1 점수)에서 제안하는 특징 군집이 학습 모델의 성능에 끼치는 영향을 확인할 수 있다. Table 3의 파란색 음영

Table 3. Evaluation Indicator for Each Number of Excluded Clusters from Fashion-MNIST

Indicator	Number of Excluded Clusters in Train Dataset	Number of Excluded Clusters in Test Dataset		
		0	3	5
Accuracy	0	0.663	0.663	0.646
	3	0.623	0.630	0.630
	5	0.578	0.572	0.574
Precision	0	0.779	0.769	0.759
	3	0.736	0.725	0.731
	5	0.695	0.688	0.678
Recall	0	0.663	0.663	0.646
	3	0.623	0.630	0.630
	5	0.578	0.572	0.574
F1 score	0	0.716	0.712	0.698
	3	0.675	0.675	0.677
	5	0.631	0.625	0.621
Data Feature Coverage of Train Dataset	0	1.000	0.989	0.982
	3	0.957	0.957	0.951
	5	0.928	0.928	0.928
Data Feature Coverage of Test Dataset	0	0.997	0.955	0.927
	3	0.872	0.882	0.854
	5	0.765	0.774	0.777

Table 4. Evaluation Indicator for Each Number of Excluded Clusters from CIFAR-10

Indicator	Number of Excluded Clusters in Train Dataset	Number of Excluded Clusters in Test Dataset		
		0	3	5
Accuracy	0	0.455	0.450	0.450
	3	0.448	0.438	0.449
	5	0.442	0.439	0.442
Precision	0	0.548	0.551	0.545
	3	0.558	0.540	0.549
	5	0.543	0.542	0.534
Recall	0	0.455	0.450	0.450
	3	0.448	0.438	0.449
	5	0.442	0.439	0.442
F1 score	0	0.498	0.495	0.493
	3	0.497	0.484	0.494
	5	0.488	0.485	0.484
Data Feature Coverage of Train Dataset	0	0.999	0.986	0.976
	3	0.957	0.957	0.947
	5	0.928	0.928	0.928
Data Feature Coverage of Test Dataset	0	0.997	0.955	0.927
	3	0.871	0.882	0.854
	5	0.810	0.820	0.831

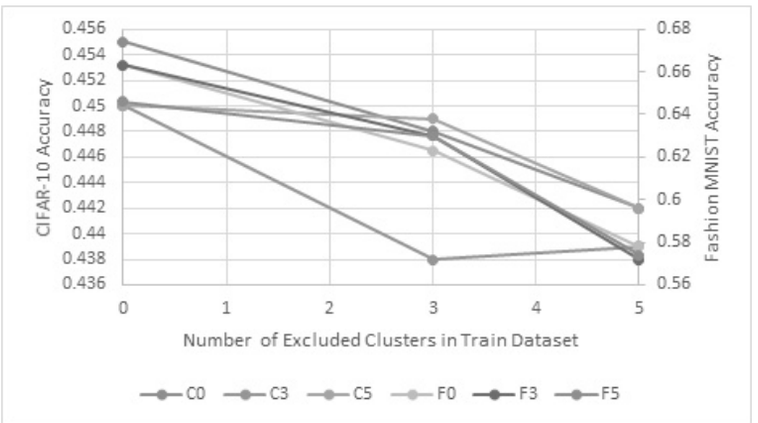

Fig. 9. Accuracy for Each Number of Excluded Clusters from CIFAR-10 and Fashion-MNIST

칸을 예시로 들면, 10개의 클래스 각각에 대해 7개의 군집으로 나뉜 총 70개의 군집 중에서 훈련 셋에서는 67개, 평가 셋에서는 70개, 67개, 65개의 군집을 사용하였을 때의 정확도를 나타낸다. Fashion-MNIST의 각 정확도는 0.623, 0.630, 0.630으로 훈련 셋이 고정된 상태에서 평가 셋이 바뀌어도 성능에는 크게 차이가 없는 것을 확인할 수 있다.

- 훈련 데이터 셋에서 배제되는 군집의 개수가 커짐에 따라 학습된 모델의 여러 성능 평가 지표가 떨어지는 것을 볼 수 있는데, 특징 군집이 훈련 셋에서 배제되어 해당 군집이 대표하는 특징을 모델이 학습할 수 없지만, 평가 데이터 셋에는 해당 군집이 다른 특징 군집보다 다수 포함되어 있기 때문이다.
- 반면, 평가 데이터 셋에서 배제되는 군집의 개수는 크게 영향을 끼치지 않는 것을 확인할 수 있다. 해당 특징 군집이 평가 데이터 셋에서 배제되지만, 해당 군집을 포함하여 다른 군집들도 학습되고 있기 때문으로 판단된다.

Fig. 9는 Table 3과 Table 4에서 제외하는 특징 군집의 개수에 따른 정확도를 표현한 그림이다. Fig. 9의 범례에서 C는 CIFAR-10, F는 Fashion-MNIST 데이터 셋을 의미하며, 범례의 숫자는 훈련 데이터 셋에서 제외된 특징 군집의 개수를 나타낸다. Fig. 9에서 나타나는 것처럼 두 데이터 셋 모두 훈련 데이터 셋에서 제외된 특징 군집의 개수에 따라 정확도가 감소하는 것을 확인할 수 있다. 예를 들어, Fashion-MNIST의 경우, 아무것도 배제하지 않은 경우(정확도 0.663)에 비해 훈련 및 평가 데이터 셋에서 각각 5개의 군집을 배제(정확도 0.574)하게 되면 정확도가 8.9%까지 차이나는 것을 확인하였다. 하지만, 평가 데이터 셋에서 제외된 특징 군집의 개수에 따른 정확도의 규칙성은 관찰하지 못하였다. CIFAR-10의 경우, Fashion-MNIST에 비해 특징 군집에 따른 모델 성능의 변동 폭이 작은데, 이는 최고 성능이 Fashion-MNIST보다 떨어지기 때문으로 판단된다.

5. 결 론

본 연구에서는 데이터 특징 커버리지를 기반으로 하여 딥러닝 영상 분류 모델을 검증하는 기법을 제안한다. 제안하는 기법은 전체 데이터의 특징 벡터를 추출하고, 군집화하여 특징 군집을 생성한다. 또한, 생성된 특징 군집을 층화추출함으로써 모든 특징이 고르게 분포하는 훈련 데이터 셋 및 평가 데이터 셋을 생성할 수 있다. 생성된 데이터 셋은 전체 특징에 대한 훈련 및 평가를 보장할 수 있으며, 특징 군집별 평가를 통해 모델이 어떠한 특징을 가진 데이터에 취약한지 확인할 수 있다. 실험 결과, 훈련 데이터 셋에서 특징 군집이 배제되는 것이 학습된 모델의 정확도에 더 큰 영향을 끼치는 것으로 보이며, Fashion-MNIST의 경우 8.5%에서 9.1%, CIFAR-10의 경우 1.3%에서 1.6%까지 정확도의 차이가 나타나는 것을 확인하였다. 이처럼 데이터 특징 커버리지를 다르게 하여 학습 및 평가 결과를 비교함으로써, 학습에 활용된 특징 군집이 학습 모델의 성능에 끼치는 영향력을 분석할 수 있도록 한다.

References

[1] A. Esteva, et al, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, Vol.542, No.7639, pp.115-118, 2017.

[2] F. Milletari, N. Navab, and S. A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016.

[3] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. Ogale, and D. Ferguson, "Real-time pedestrian detection with deep network cascades," 2015.

[4] M. Bojarski, et al., "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[5] A. Ferdowsi and W. Saad, "Deep learning for signal authentication and security in massive internet-of-things systems," *IEEE Transactions on Communications*, Vol.67, No.2, pp.1371-1387, 2018.

[6] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Ijcai*, Vol.14. No.2, pp.1137-1145, 1995.

[7] S. Mani, A. Sankaran, S. Tamilselvam, and A. Sethi, "Coverage testing of deep learning models using dataset characterization," *arXiv preprint arXiv:1911.07309*, 2019.

[8] T. Lindeberg, "Scale invariant feature transform," pp.10491, 2012.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol.1. IEEE, 2005.

[10] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," *Proceedings. International Conference on Image Processing*, Vol.1. IEEE, 2002.

[11] M. Flickner, et al, "Query by image and video content: The QBIC system," *Computer*, Vol.28, No.9, pp.23-32, 1995.

[12] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[13] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014.

[14] L. Liu, C. Shen, and A. Van Den Hengel, "The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[15] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[16] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," p.7, 2009.

[17] S. Yadav and S. Shukla, "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification," *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, IEEE, 2016.

[18] Y. Xu and R. Goodacre, "On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning," *Journal of Analysis and Testing*, Vol.2, No.3, pp.249-262, 2018.

[19] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua, "Fast keypoint recognition using random ferns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.32, No.3, pp.448-461, 2009.

[20] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, Vol.29, No.1, pp.51-59, 1996.

[21] B. Froba and A. Ernst, "Face detection with the modified census transform," *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings*, IEEE, 2004.

[22] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, Vol.40, No.1, pp.16-28, 2014.

[23] D. ping Tian, "A review on image feature extraction and representation techniques," *International Journal of Multimedia and Ubiquitous Engineering*, Vol.8, No.4, pp.385-396, 2013.

[24] A. K. Jain and A. Vailaya, "Image retrieval using color and shape," *Pattern Recognition*, Vol.29, No.8, pp.1233-1244, 1996.

[25] G. Pass, and R. Zabih, "Histogram refinement for content-based image retrieval," *Proceedings Third IEEE Workshop on Applications of Computer Vision*, WACV'96. IEEE, 1996.

[26] J. Huang, S.R. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Image indexing using color correlograms," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 1997.

[27] L. Rokach and O. Maimon, "Clustering methods," Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA, pp.321-352, 2005.

[28] S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," *2010 Third International Symposium on intelligent Information Technology and Security Informatics*, IEEE, 2010.

[29] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, Vol.16, No.3, pp.645-678, 2005.

[30] R. Bro and A. K. Smilde, "Principal component analysis," *Analytical Methods*, Vol.6, No.9, pp.2812-2831, 2014.

[31] J. Xue, C. Lee, S. G.Wakeham, and R. A. Armstronga, "Using principal components analysis (PCA) with cluster analysis to study the organic geochemistry of sinking particles in the ocean," *Organic Geochemistry*, Vol.42, No.4, pp.356-367, 2011.

[32] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *International Journal*, Vol.1, No.6, pp.90-95, 2013.

임 창 남

https://orcid.org/0000-0002-4525-5176
e-mail : chn0714@naver.com
2015년 아주대학교 전자공학과(학사)
2019년 ~ 2021년 아주대학교 AI융합네트워크학과(석사)
2021년 ~ 현 재 Tmax R&D Center ㈜티맥스오피스연구소 연구원
관심분야 : Medical Images, Machine Learning, Deep Learning, Big-Data Analysis, Embedded Software

박 예 슬

https://orcid.org/0000-0003-2584-7489
e-mail : yeseuly777@gmail.com
2015년 아주대학교 전자공학과(학사)
2017년 아주대학교 전자공학과(석사)
2017년 ~ 현 재 아주대학교 AI융합네트워크학과 박사과정
관심분야 : Data Modeling and Analysis, Machine Learning, Deep Learning, Collaborative Robot, Predictive Maintenance, Intelligent Embedded Software

이 정 원

https://orcid.org/0000-0001-8922-063X
e-mail : jungwony@ajou.ac.kr
1993년 이화여자대학교 전자계산학과(학사)
1995년 이화여자대학교 전자계산학과(석사)
1995년 ~ 1997년 LG종합기술원 주임연구원
2003년 이화여자대학교 컴퓨터학과(박사)
2003년 ~ 2006년 이화여자대학교 컴퓨터학과 BK교수, 전임강사(대우)
2006년 ~ 현 재 아주대학교 전자공학과/AI융합네트워크학과 교수
관심분야 : Context awareness, Big Data Analysis, Predictive Maintenance, Collaborative Robots, Intelligent Embedded Software