석 사 학 위 논 문

이미지 분류 모델 테스트를 위한 효과적인 데이터 선정 방법

최 영 원

부산대학교

정보융합공학과

2022년 2월

이미지 분류 모델 테스트를 위한 효과적인 데이터 선정 방법

최영원

2022년
2월

이미지 분류 모델 테스트를 위한 효과적인 데이터 선정 방법

이 논문을 공학석사 학위논문으로 제출함

최영원

부산대학교 대학원

정보융합공학과

지 도 교 수 채흥석

최영원의 공학석사 학위논문을 인준함

2021년 12월 17일

위원장	조 환 규	인
위 원	홍 봉 희	인
위 원	채 흥 석	인

차 례

<제목 차례>

표 차례

<표 차례>

그림 차례

<그림 차례>

이미지 분류 모델 테스트를 위한 효과적인 데이터 선정 방법

최영원

부산대학교 대학원 정보융합공학과

요약

심층 신경망의 성능이 향상됨에 따라 여러 분야에서 심층 신경망을 활용한다. 그러나 이미지 분류 모델의 경우 적대적 공격 기법에 취약함이 발견되었고, 자율 주행을 비롯한 안전 필수 시스템에서 활용되는 이미지 분류 모델이 결함을 발생하는 경우 인적/금전적 손실이 발생할 수 있다. 이에 이미지 분류 모델의 테스트를 위하여 테스트 데이터 생성 방법과 방어 기법 등이 연구되었다.

기존 이미지 분류 모델 테스트에서는 적대적 공격 기법과 데이터 증강 기법 등 데이터 변형 기법들을 활용하여 테스트 데이터를 생성하였다. 그러나 무분별한 데이터 변형으로 인해 이미지 분류 모델 테스트에 적합하지 않은 데이터가 생성될 수 있다. 그러므로 변형된 테스트 데이터 중 이미지 분류 모델 테스트에 효과적인 테스트 데이터를 선정할 기준이 필요하다.

본 연구에서는 피처 맵 기반으로 이미지 분류 모델 테스트에 효과적인 데이터를 선정하는 방법을 제안하였다. 이미지 분류 모델의 추론 결과는 입력 데이터에 대한 모델의 뉴런 활성화 값에 영향을 받으며, 뉴런 활성화 값은 모델의 레이어별로 피처 맵 형태로 출력할 수 있다. 학습 데이터의 피처 맵과 테스트 데이터의 피처 맵 간의 뉴런 활성화 값 차이를 측정 방법으로 피처 맵 거리를 정의하였다. 테스트 데이터의 피처 맵 거리를 기반으로 FMD 기준치를 선정하였다. FMD 기준치보다 더 큰 피처 맵 거리를 가지는 테스트 데이터들을 FMD 테스트 데이터로 선정한다.

사례 연구로 STL10 데이터 셋과 ResNet-20 모델에 대하여 실험을 진행하였다. STL10의 10개 레이블 중 6개 레이블에서 피처 맵 거리가 클수록 정확도가 감소하여 음의 상관관계를 가짐을 확인하였다. 원본 테스트 데이터 중 오분류 테스트 데이터의 평균 피처 맵 거리를 FMD 테스트 데이터 기준 피처 맵 거리로 선정하였다. 데이터 증강 기법이 적용된 테스트 데이터 셋 20개에 대하여 FMD 테스트 데이터를 선정하였고 기존 테스트 데이터 셋과 FMD 데이터 셋의 테스트 효과성을 비교하여 FMD 테스트 데이터의 테스트 효과성이 평균적으로 더 크게 측정됨을 확인하였다.

I. 서론

심층 신경망의 성능은 2012년 AlexNet을 시작으로 GoogLeNet, ResNet 등 심층 신경망의 구조가 발전함에 따라 향상되었다[1-3]. 심층 신경망의 성능 향상으로 심층 신경망은 자율 주행, 의료 진단 등 여러 분야에서 활용되고 있다.

심층 신경망 중에서 이미지 분류 모델의 취약성이 발견됨에 따라, 심층 신경망의 품질에 대한 테스트가 중요시 되고 있다. 이미지 분류 모델은 적대적 데이터에 대해 정확도가 낮게 측정되며 취약성을 확인되었다[4]. 이미지 분류 모델을 적용한 자율주행 자동차, 안면 인식 보안 시스템 등 안전 필수 시스템에서 이미지 분류 모델의 오동작이 발생할 경우 막대한 인적/금전적 손실이 발생할 수 있다. 대표적인 사례로 구글의 자율주행 택시 충돌사고와 테슬라의 자율주행 시스템 차량 충돌사고가 있으며, 테슬라의 사례에서는 자율주행 시스템으로 운행하던 차량이 트레일러의 흰색을 하늘로 판단하여 트레일러와 충돌하는 사고가 발생하였다[5].

이미지 분류 모델의 테스트는 테스트 케이스를 통해 이미지 분류 모델의 품질을 평가하는 과정이다. 이미지 분류 모델의 품질은 테스트 케이스에 대한 심층 신경망의 정확도로 평가할 수 있다. 테스트 데이터에 대한 이미지 분류 모델의 정확도가 학습 데이터에 대한 정확도보다 큰 차이로 낮게 측정되는 경우, 해당 모델은 결함이 있다고 평가할 수 있다. 테스트 데이터에 대한 모델의 분류 성능이 낮게 측정되는 원인으로는 모델이 학습 데이터에 대해 오버피팅(Overfitting)되었거나 특정 레이블에 편향된 데이터 셋을 모델 학습에 사용하는 것 등이 이다.

이미지 분류 모델의 테스트에서 결함은 심층 신경망이 테스트 데이터의 레이블을 올바르게 분류하지 못하는 것이다. 결함이 없는 이미지 분류 모델은 테스트 데이터를 실제 레이블로 올바르게 분류할 수 있어야 한다. 따라서 이미지 분류 모델 테스트에 효과적인 데이터는 심층 신경망이 실제 레이블로 올바르게 분류하지 못하는 데이터이다.

기존 이미지 분류 모델 테스트에서는 모델의 결함을 찾기 위한 변형된 테스트 데이터들을 생성하였다. 원본 테스트 데이터는 올바르게 분류하는데 변형된 테스트 데이터를 올바르게 분류하지 못한다면 모델이 결함을 가지고 있다고 판단하였다. 테스트 데이터 변형 방법으로 적대적 공격(Adversarial Attack) 기법과 데이터 증강(Data Augmentation) 기법을 활용하였다[5-8]. 적대적 공격 기법은 사람이 인지하지 못하는 미세한 정도로 데이터 값을 변형시킨다. 데이터 증강 기법은 이미지의 색상 또는 아핀(Affine) 공간을 변형시킨다.

이미지 분류 모델의 테스트에 효과적인 데이터 생성에는 테스트 데이터를 선별하는 기준이 필요하다. 무분별한 적대적 공격 기법과 데이터 증강 기법 적용은 이미지 분류 모델 테스트에 효과적이지 않은 데이터가 생성될 수 있다. 이미지의 변형 강도가 미약하여 모델이 분류하는 데이터가 생성되거나, 이미지 변형 강도가 과도하여 사람도 레이블을 구별하기 어려운 데이터가 생성될 수 있다.

본 연구에서는 이미지 분류 모델에서 레이블을 추론하는 과정에서 출력되는 데이터를 기반으로 테스트 데이터를 선정하고자 한다. 이미지 분류 모델의 성능은 뉴런 활성화 값(Neuron Activation Value)에 영향을 받는다. 심층 신경망은 학습 데이터에 따라 뉴런별 웨이트(Weight) 값을 가지게 되며, 뉴런의 활성화 함수(Activation Function)는 입력 데이터의 값과 웨이트 값으로 뉴런 활성화 값을 계산한다[5]. 이미지 분류 모델의 레이어 순서대로 뉴런 활성화 값이 계산되고, 심층 신경망의 출력 레이어(Output Layer)에서 이전 레이어의 뉴런 활성화 값을 기반으로 입력 데이터의 레이블을 추론한다.

이미지 분류 모델의 테스트에 적합한 데이터 선정 기준으로 학습 데이터에 대한 모델의 뉴런 활성화 값과 테스트 데이터에 대한 모델의 뉴런 활성화 값의 차이를 사용하고자 한다. 이미지 분류 모델의 레이어별 뉴런 활성화 값들은 레이어별로 피처 맵 형태로 출력할 수 있다. 마지막 히든 레이어에서 측정된 학습 데이터의 피처 맵과 테스트 데이터의 피처 맵을 비교한 결과 모델이 올바르게 레이블을 분류한 데이터들의 피처 맵은 학습 데이터의 피처 맵과 유사함을 보였지만, 모델이 다른 레이블로 잘못 분류한 데이터들은 학습 데이터의 피처 맵과 뉴런 활성화 값의 크기와 분포에서 차이가 있음을 확인하였다. 따라서 학습 데이터와의 뉴런 활성화 값 차이가 큰 테스트 데이터들이 이미지 분류 모델 테스트에 효과적인 데이터일 가능성이 있다.

본 연구에서는 학습된 레이블의 학습 데이터의 뉴런 활성화 값과 테스트 데이터의 뉴런 활성화 값 간의 차이를 측정하여 심층 신경망 테스트에 효과적인 데이터를 선정하는 방법을 제안한다. 학습 데이터의 피처 맵과 테스트 데이터의 피처 맵 간의 뉴런 활성화 값 차이를 측정 방법으로 피처 맵 거리(Feature Map Distance)를 정의하고, 이미지 분류 모델 테스트에 효과적인 데이터를 선정하는 기준이 되는 피처 맵 거리를 FMD 기준치라고 정의한다. 다른 레이블로 잘못 분류되는 테스트 데이터의 피처 맵에서 학습 데이터의 피처 맵보다 더 큰 뉴런 활성화 값을 가지는 뉴런들이 더 많다는 차이를 확인하여, FMD 기준치보다 더 큰 피처 맵 거리를 가지는 데이터들을 FMD 테스트 데이터 셋으로 정의한다.

본 연구에서는 STL10 데이터 셋과 ResNet-20 모델을 대상으로 사례 연구를 수행하

였다. STL10 데이터 셋의 테스트 데이터 피처 맵 거리를 측정하고, 레이블별 피처 맵 거리와 올바르게 분류된 데이터 비율의 상관관계를 측정하였다. STL10의 10개 레이블 중 6개 레이블에서 음의 상관관계를 가짐을 확인하여 피처 맵 기반의 테스트 데이터 선정의 효과성을 확인하였다.

피처 맵 거리가 커질수록 정확도가 감소함에 따라 STL10의 잘못 분류된 테스트 데이터의 평균 피처 맵 거리로 FMD 기준치를 선정하였고, 여러 데이터 증강 기법을 적용한 테스트 데이터 셋 20개에 대하여 FMD 기준치보다 더 큰 피처 맵 거리를 가지는 데이터들을 FMD 테스트 데이터로 선정하였다. 테스트 데이터 셋 20개에 대하여 FMD 기준치로 데이터를 선정하기 이전의 테스트 데이터 셋의 테스트 효과성과 FMD 테스트 데이터의 테스트 효과성을 비교하였으며, FMD 테스트 데이터 셋의 테스트 효과성이 더 크게 측정됨을 확인하였다. 레이블별로 테스트 데이터와 FMD 테스트 데이터의 테스트 효과성을 비교한 결과 10개 레이블 중 6개 레이블에서 테스트 효과성이 상승하였다.

본 논문의 2장에서는 연구 배경을 설명한다. 3장에서는 피처 맵 기반 데이터 선정 방법을 정의하고 피처 맵 거리를 활용한 테스트 데이터 선정 과정을 설명한다. 4장에서는 본 연구에서 수행한 사례 연구의 실험 환경과 실험 결과를 설명한다. 5장에서는 기존 이미지 분류 모델의 테스트를 위한 데이터 생성 방법들을 소개한다. 6장에서는 본 논문의 사례 연구에 대한 결론과 향후 연구를 설명한다.

II. 연구 배경

본 장에서는 이미지 분류 모델 테스트를 위한 효과적인 데이터 선정 방법을 제안하게 된 배경 지식들을 설명한다. 본 연구에서는 모델에서 출력되는 값을 기반으로 테스트 데이터 중에서 이미지 분류 모델의 테스트에 효과적인 테스트 데이터를 선정하고자 하였다. 모델의 레이어별로 출력되는 피처 맵을 기반으로 효과적인 테스트 데이터를 선정한다. 이에 이미지 분류 모델의 구조와 뉴런 활성화 값, 피처 맵을 설명한다.

1. 이미지 분류 모델 구조

심층 신경망은 뉴런으로 구성된 사람의 뇌에서 영감을 받아 뉴런으로 구성되어 있다[5]. 그림 1은 심층 신경망 중 이미지 분류 모델의 구조와 입력 데이터에 대한 결과 추론 과정을 표현한다. 모델은 여러 레이어로 구성되고, 각 레이어는 여러 뉴런으로 구성된다. 레이어는 계층에 따라 Input Layer, Hidden Layer, Output Layer로 분류된다. Input Layer는 심층 신경망의 첫 번째 레이어로, 입력 데이터를 입력받아 Hidden Layer에 입력 데이터 값을 전송한다. Hidden Layer는 Input Layer와 Output Layer 사이에 있는 레이어들이고, 이미지 분류 모델의 Hidden Layer는 Input Layer로부터 전달받은 입력 데이터에서 피처를 추출하여 Output Layer에 전송한다. Output Layer는 Hidden Layer로부터 피처를 바탕으로 최종적으로 모델이 입력 데이터에 대한 레이블을 추론한 결과를 출력한다.

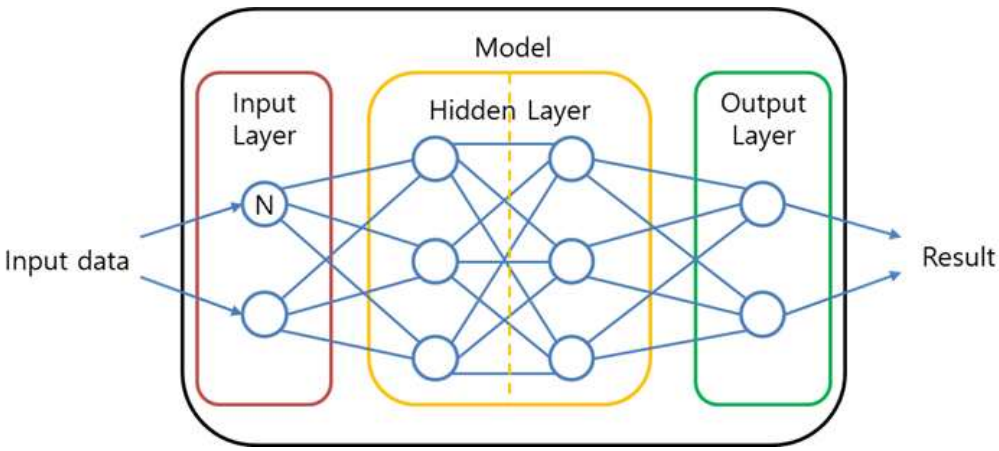


그림 1 심층 신경망의 추론 과정

2. 뉴런 활성화 값

뉴런 활성화 값(Neuron Activation Value)은 입력 데이터에 대하여 뉴런의 활성화 함수가 출력하는 값이다[5]. 그림 2는 뉴런 활성화 값 계산 과정을 표현하고, 식 (1)과 (2), (3)은 뉴런 활성화 값 계산 과정을 정의한다. 레이어 l의 뉴런 집합 N_l 중에서 n번째 뉴런 n_l은 활성화 함수 A에 입력된 입력 데이터 i에 대한 이전 레이어 $l-1$의 뉴런 활성화 값들 a_{l-1}^{i}과 이전 레이어의 뉴런들과의 웨이트 값 집합 $w_{l-1}^{n_l}$에 따라 뉴런 활성화 값 $a_{n_l}^{i}$을 출력한다. 웨이트 값 집합 $w_{l-1}^{n_l}$은 뉴런 n_l과 이전 레이어 $l-1$의 뉴런 집합 N_{l-1}과의 보정값 집합이다.

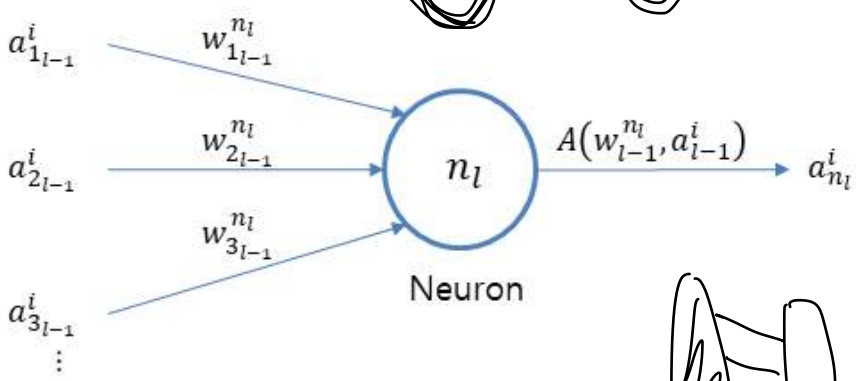


그림 2 뉴런 활성화 값 계산 과정

$$a_l^i = \left\{ a_{n_l}^i | n_l \in N_l \right\} \tag{1}$$

$$w_{l-1}^{n_l} = \left\{ w_{m_{l-1}}^{n_l} | n_l \in N_l, m_{l-1} \in N_{l-1} \right\} \tag{2}$$

$$a_{n_l}^i = A(w_{l-1}^{n_l}, a_{l-1}^i) = \sigma\left(\sum_{m=1}^{|N_{l-1}|} w_{m_{l-1}}^{n_l} a_{m_{l-1}}^i \right) \tag{3}$$

뉴런 활성화 값은 학습 데이터와 입력 데이터에 영향을 받는다. 활성화 함수에 입력되는 웨이트 값은 심층 신경망의 학습 과정에서 학습 데이터를 올바르게 분류하기 위해 최적화된다. 심층 신경망이 학습을 완료한 이후에는 웨이트 값이 변경되지 않는다. 그러므로 테스트 데이터에 대한 심층 신경망의 정확도를 측정할 때 웨이트 값이 사용된다.

따라서 심층 신경망의 학습 데이터에 대한 정확도와 테스트 데이터에 대한 정확도는 학습 데이터와 테스트 데이터의 차이에서 발생한다.

3. 피처 맵

피처 맵은 입력 데이터에 대한 심층 신경망 내부 레이어의 뉴런 활성화 값 집합이다. 그림 3은 피처 맵 출력 과정을 표현하고, 식 (4)와 (5)는 피처 맵 계산 과정을 정의한다. 입력 데이터 i가 이미지 분류 모델 M에 입력되면 레이어별로 피처 맵이 출력되며, 레이어 l에서는 R행과 C열의 2차원 공간의 피처 맵 FM_l^i를 출력한다.

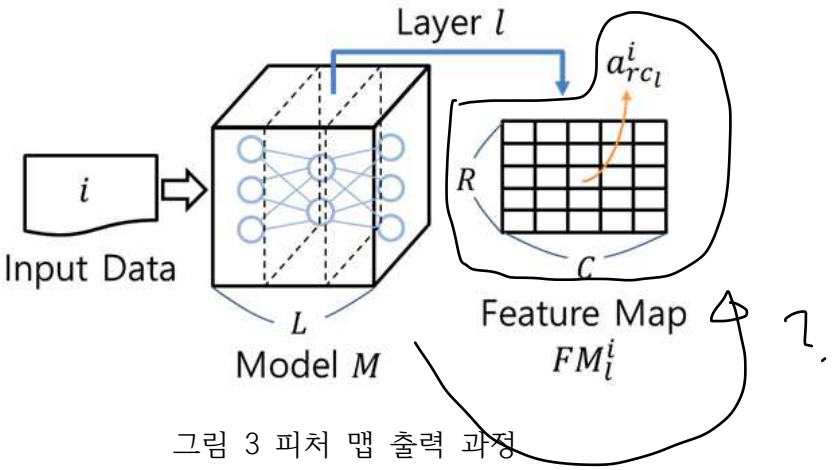


그림 3 피처 맵 출력 과정

$$M = \{l | 1 \le l \le L\} \tag{4}$$

$$FM_l^i = \left\{a^i_{rc_l} | 1 \le l \le L, 1 \le r \le R, 1 \le c \le C\right\} \tag{5}$$

이미지 분류 모델에서 피처 맵을 출력한 결과 올바르게 분류된 정분류 테스트 데이터(Classified Test Data)의 피처 맵과 다른 레이블로 잘못 분류된 오분류 테스트 데이터(Misclassified Test Data)의 피처 맵은 올바르게 분류된 정분류 학습 데이터(Classified Training Data)의 피처 맵과 차이를 보인다. 그림 4와 그림 5는 각각 STL10 데이터 셋의 Airplane 레이블과 Car 레이블의 정분류 학습 데이터의 평균 피처 맵과 정분류 테스트 데이터 셋의 평균 피처 맵, 오분류 테스트 데이터 셋의 평균 피처 맵이다. 정분류 테스트 데이터 셋의 평균 피처 맵과 정분류 학습 데이터의 피처 맵 간에는 뉴런 활성화

값 분포가 유사함을 보이지만, 오분류 테스트 데이터 셋의 피처 맵은 정분류 학습 데이터의 피처 맵과 큰 뉴런 활성화 값을 가지는 뉴런의 분포에서 차이가 있음을 보인다. 이를 기반으로 본 연구에서는 학습 데이터의 피처 맵과 테스트 데이터의 피처 맵 간의 차이를 기준으로 이미지 분류 모델의 테스트에 효과적인 데이터를 선정하고자 하였다.

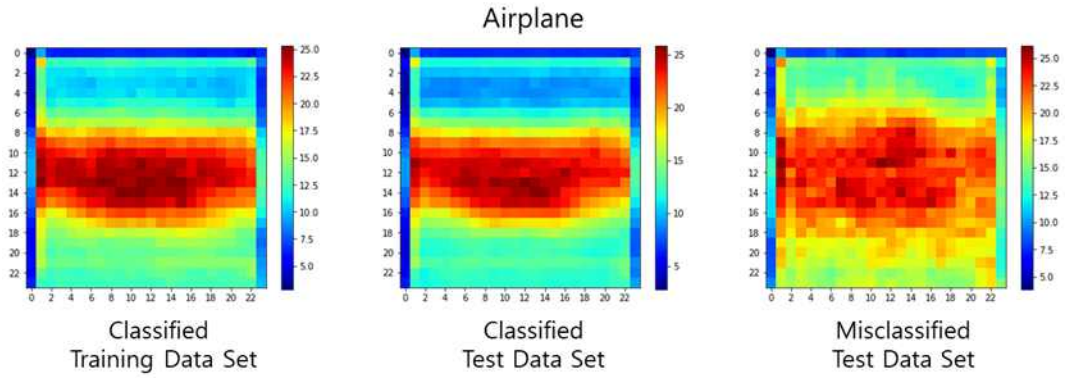


그림 4 STL10 Airplane 레이블 데이터 셋별 평균 피처 맵

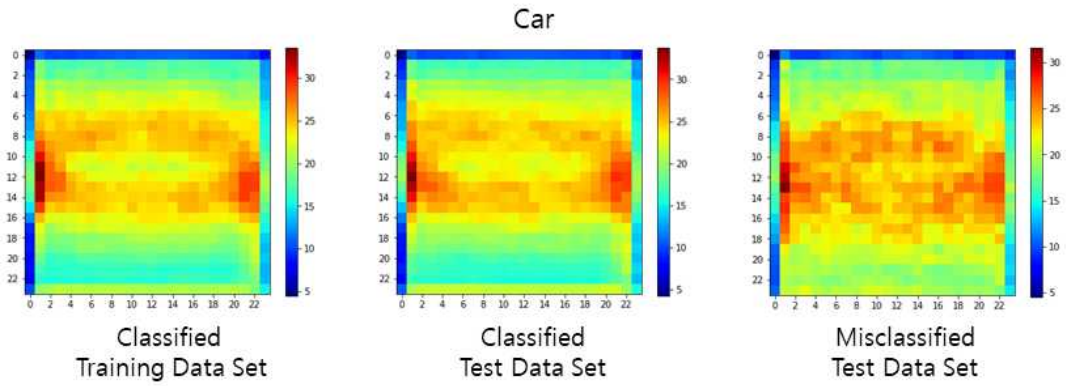


그림 5 STL10 Car 레이블 데이터 셋별 평균 피처 맵

III. 피처 맵 거리 기반 테스트 데이터 선정 방법

본 연구에서는 심층 신경망 테스트에 효과적인 데이터를 선정하기 위한 피처 맵 거리 기반 데이터 선정 과정을 설명한다. 그림 6은 피처 맵 거리를 기반으로 이미지 분류 모델 테스트에 효과적인 데이터를 선정하는 과정을 나타내며, 크게 테스트 데이터 기준 선정과 테스트 데이터 선정 두 단계를 거쳐 수행한다. 세부적으로 테스트 데이터 기준 선정은 피처 맵 거리 측정 단계와 FMD 기준치 선정 두 단계를 순차적으로 수행한다. 테스트 데이터 선정은 테스트 데이터 셋 생성과 피처 맵 거리 비교 두 단계로 구성된다.

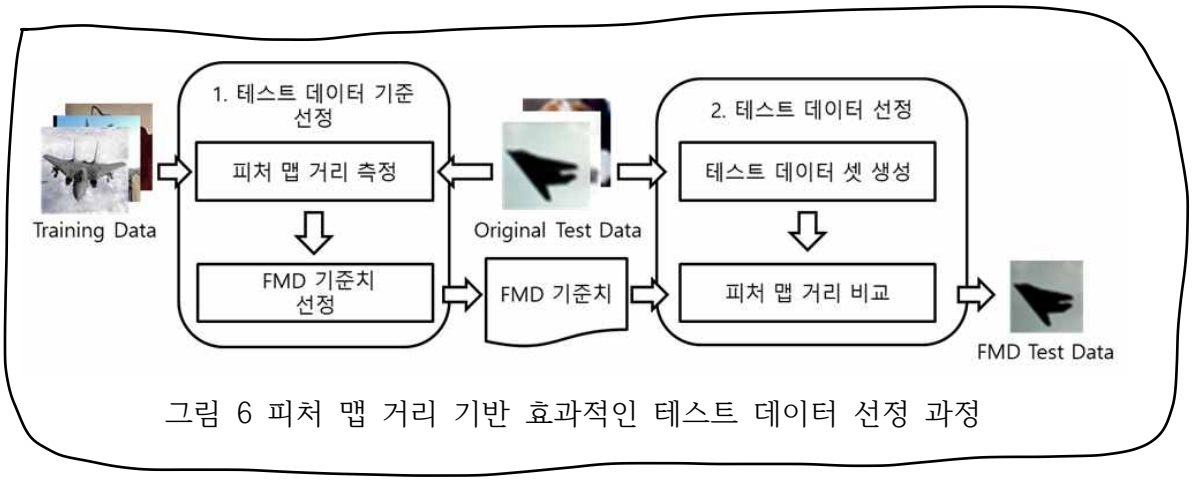


그림 6 피처 맵 거리 기반 효과적인 테스트 데이터 선정 과정

1. 테스트 데이터 기준 선정

피처 맵 거리 기반 데이터 선정 방법에서는 테스트 데이터 선정을 위한 기준으로 FMD(Feature Map Distance) 기준치를 사용한다. 그림 7은 피처 맵 거리 기반 테스트 데이터 선정을 위한 기준이 되는 피처 맵 거리를 선정 과정을 표현한다. 학습 데이터와 테스트 데이터에 대한 모델의 피처 맵을 출력하고 학습 데이터의 레이블별로 베이스 피처 맵(Base Feature Map)을 생성한다. 레이블별로 베이스 피처 맵과 테스트 데이터의 피처 맵 간의 피처 맵 거리(Feature Map Distance)를 측정한다. 레이블이 올바르게 분류되는 테스트 데이터의 피처 맵 거리와 다른 레이블로 잘못 분류되는 테스트 데이터의 피처 맵 거리로부터 효과적인 테스트 데이터 선정 기준이 될 FMD 기준치를 선정한다.

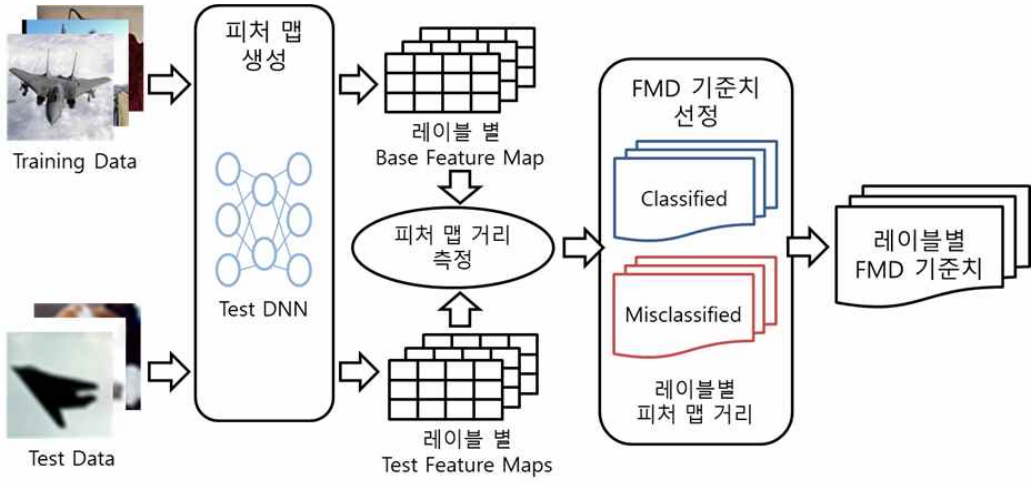


그림 7 테스트 데이터 기준 선정 과정

가. 베이스 피처 맵 생성

베이스 피처 맵(Base Feature Map)은 피처 맵 거리 측정의 기준이 되는 피처 맵으로, 학습 데이터 피처 맵의 평균이다. 그림 8은 베이스 피처 맵 생성 과정을 표현하며, 식 (6)과 (7)은 베이스 피처 맵 계산 과정을 정의한다. 레이어 l에서 출력된 학습 데이터 피처 맵 집합 FM_l^{train}은 학습 데이터 수 T만큼 피처 맵을 가지며, 학습 데이터별 피처 맵에서 동일한 좌표의 뉴런 활성화 값들의 평균을 계산하여 베이스 피처 맵 BFM_l을 생성한다.

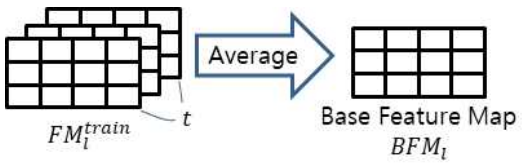


그림 8 베이스 피처 맵 생성 과정

$$FM_l^{train} = \left\{ FM_l^{train_t} | 1 \le t \le T \right\} \quad (6)$$

$$BFM_l = \left\{ \frac{\sum_{t=1}^{T} a_{rc_l}^{train_t}}{T} | a_{rc_l}^{train_t} \in FM_l^{train_t} \right\}, \quad (7)$$

$$1 \le l \le L, 1 \le r \le R, 1 \le c \le C$$

베이스 피처 맵 생성 시에는 레이블별로 생성해야 한다. 그림 9는 STL10 데이터 셋 [11]의 Airplane과 Car, Ship 레이블의 베이스 피처 맵 예시이며, 레이블별로 큰 뉴런 활성화 값의 분포가 다른 것을 확인할 수 있다. 이는 입력 데이터의 레이블에 따라서 심층 신경망 내부에 활성화되는 뉴런이 다름을 보여준다.

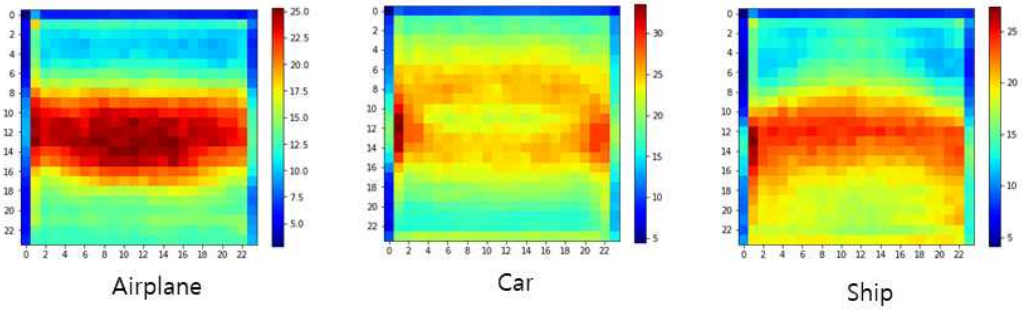


그림 9 STL10 레이블별 베이스 피처 맵 예시

나. 테스트 데이터 피처 맵 거리 측정

피처 맵 거리(Feature Map Distance)는 동일한 레이블의 특정 테스트 데이터의 피처 맵과 베이스 피처 맵 간의 거리이다. 그림 10은 피처 맵 거리 측정 과정을 표현한다. 레이어 l에서의 테스트 데이터 n에 대한 피처 맵 거리 $D_l^{test_n}$은 테스트 데이터 n의 피처 맵 $FM_l^{test_n}$과 베이스 피처 맵 BFM_l의 동일한 행과 열의 뉴런 활성화 값의 차이 $d_{rc_l}^{test_n}$로 계산된다.

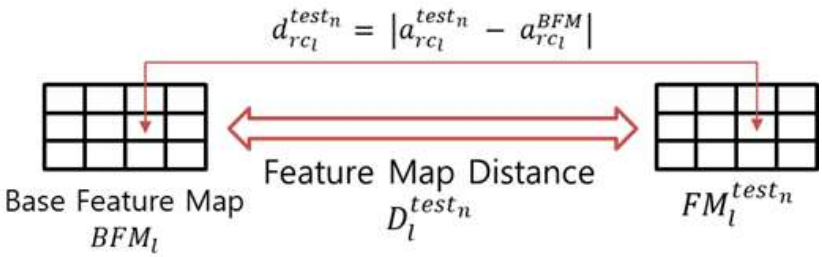


그림 10 피처 맵 거리 계산 과정

두 피처 맵 간의 거리는 L_p Norm 거리로 측정하며, 식 (8)과 (9), (10)은 피처 맵 거

리 계산 과정을 정의한다. 본 연구에서는 L_1 Norm 거리와 L_2 norm 거리로 피처 맵 간의 거리를 측정한다. L_1 Norm 거리로 측정한 테스트 데이터 n의 피처 맵 거리를 L_1 피처 맵 거리 $L_1D_l^{test_n}$라고 정의하고 피처 맵의 행렬 좌표별 뉴런 활성화 값 차이의 절대값의 합으로 계산한다. L_2 Norm 거리로 측정한 테스트 데이터 n의 피처 맵 거리는 L_2 피처 맵 거리 $L_2D_l^{test_n}$라고 정의하고 행렬 좌표별 뉴런 활성화 값 차이의 제곱들의 합을 제곱근한 값으로 계산된다.

L_1, L_2 : 맵해든 거리

$$d_{rc_l}^{test_n} = \left| a_{rc_l}^{test_n} - a_{rc_n}^{BFM} \right|, 1 \le l \le L, 1 \le r \le R, 1 \le c \le C \tag{8}$$

$$L_1D_l^{test_n} = \sum_{r=1}^{R}\sum_{c=1}^{C} d_{rc_l}^{test_n} \tag{9}$$

$$L_2D_l^{test_n} = \sqrt{\sum_{r=1}^{R}\sum_{c=1}^{C} (d_{rc_l}^{test_n})^2} \tag{10}$$

다. FMD 기준치 선정

FMD 기준치는 테스트 데이터 중 심층 신경망의 테스트에 효과적인 데이터를 선정하는 기준이 되는 피처 맵 거리이다. 그림 11은 원본 테스트 데이터의 피처 맵 거리의 예상 분포를 Box plot으로 표현한다. 본 연구에서는 오분류 테스트 데이터의 피처 맵 거리가 정분류 테스트 데이터의 피처 맵 거리보다 클 것으로 예상하므로, FMD 기준치를 오분류 테스트 데이터의 평균 피처 맵 거리로 선정한다.

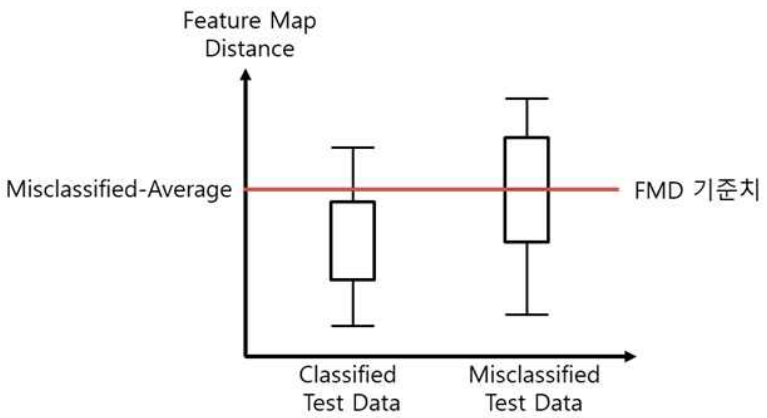


그림 11 피처 맵 거리 예상 분포

2. 테스트 데이터 선정

그림 12는 테스트 데이터 선정 과정을 표현한다. 테스트 데이터 선정 과정은 테스트 데이터 셋 생성과 피처 맵 거리 비교 두 단계로 구성된다. 테스트 데이터 셋 생성 단계에서는 다양한 테스트 데이터 생성을 위해 여러 데이터 증강 기법들을 적용하여 변형된 테스트 데이터로 구성된 테스트 데이터 셋을 생성한다. 피처 맵 거리 비교 단계에서는 변형된 테스트 데이터 셋들을 각 레이블별 FMD 기준치보다 큰 피처 맵 거리를 가지는 데이터들을 FMD 테스트 데이터 셋(FMD Test Data Set)으로 선정한다.

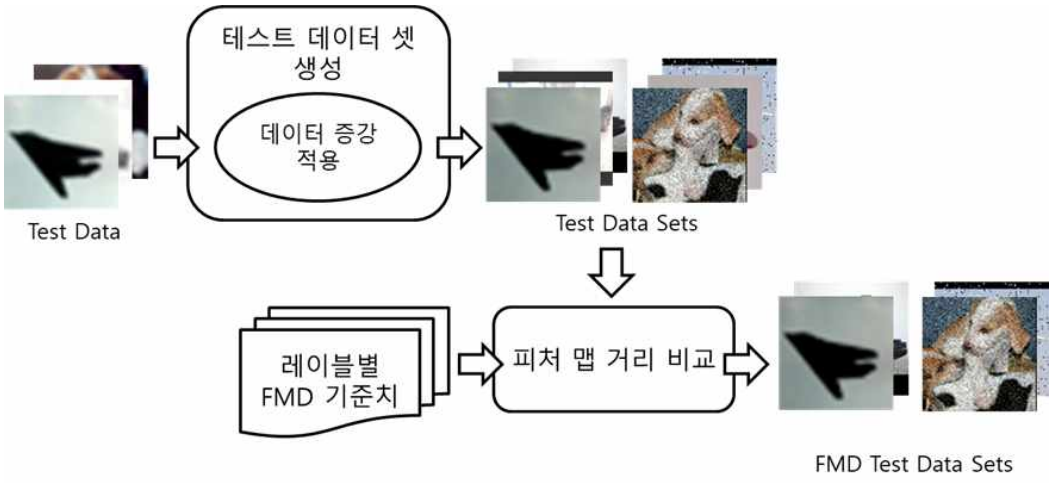
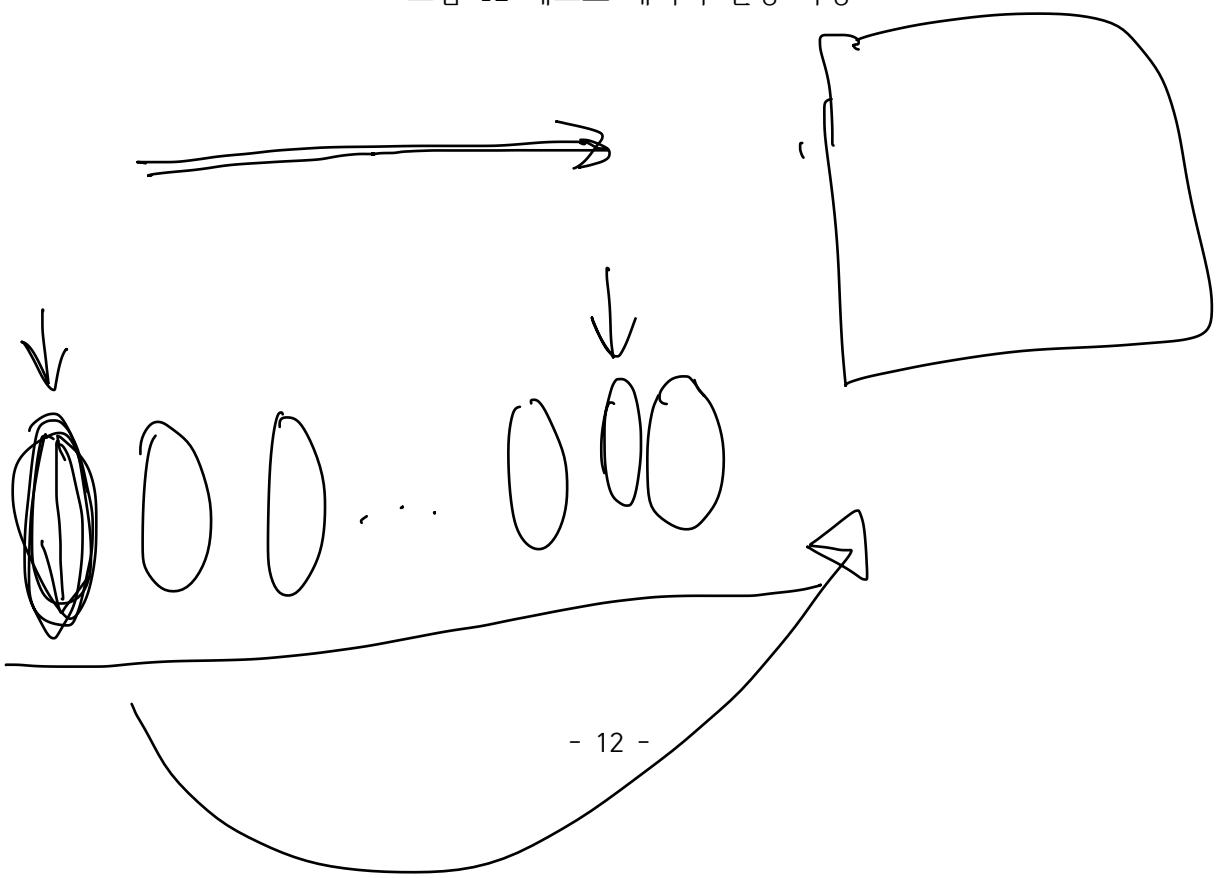


그림 12 테스트 데이터 선정 과정



IV. 사례 연구

본 장에서는 사례 연구의 실험 환경, 실험 결과를 설명한다. 그림 13은 이미지 분류 모델의 테스트에 효과적인 데이터 선정을 위한 피처 맵 기반 테스트 데이터 선정 방법 실험 과정을 표현한다. 본 연구에서는 데이터 증강을 적용한 테스트 데이터 셋에 대하여 FMD 테스트 데이터를 선정하고, 데이터 증강을 적용한 테스트 데이터 셋의 테스트 효과성과 FMD 테스트 데이터의 테스트 효과성을 측정하고 비교하였다,

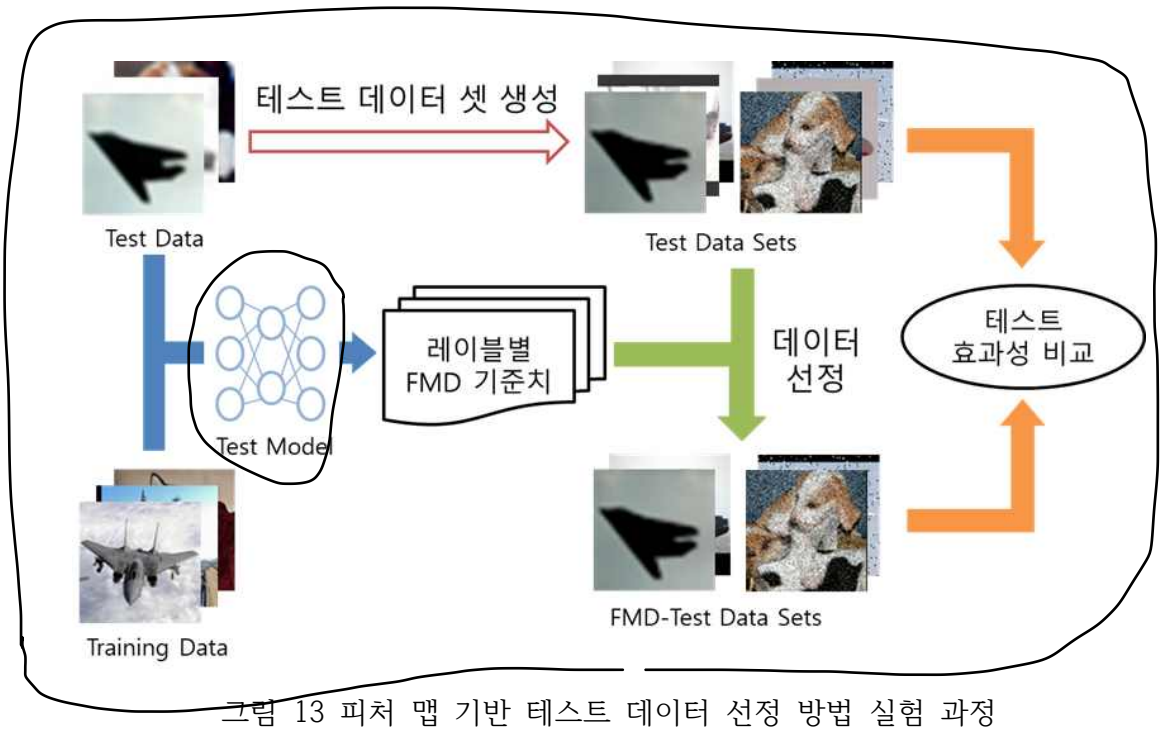


그림 13 피처 맵 기반 테스트 데이터 선정 방법 실험 과정

1. 실험 환경

가. 데이터 셋

본 연구에서는 STL10 데이터 셋을 사용한다[9]. STL10은 이미지 분류 모델을 위한 데이터 셋으로 Airplane과 Bird, Car, Cat, Deer, Dog, Horse, Monkey, Ship, Truck 총 10개 레이블로 구성되어 있다. 표 1은 STL10 데이터 셋의 특정을 정리한 표이며, 그림 14는 STL10 데이터 셋의 레이블별 예시 이미지이다.

표 1 STL10 데이터 셋 특성

특성	설명
컬러/흑백 여부	컬러
학습 데이터 수	10 label × 500 images/label = 5,000 images
테스트 데이터 수	10 label × 800 images/label = 8,000 images
이미지 크기	96 pixel × 96 pixel

그림 14 STL10 데이터 셋 데이터 예시

나. 테스트 대상 모델

본 연구에서 테스트 대상 모델로 ResNet-20 모델[3]을 사용하였다. ResNet은 Skip Connection을 사용하는 이미지 분류 모델이다. Skip Connection은 입력값이 레이어를 건너뛰는 지름길이다. 기존 CNN 모델들은 성능을 향상시키기 위해 레이어 수를 증가시켰을 때 오히려 정확도가 감소하는 기울기 소실 문제(Vanishing Gradient Problem)가 발생하였다[3]. ResNet은 Skip Connection을 사용하여 입력값 X를 레이어를 건너뛰어 다시 더함으로써 기울기 소실 문제를 해결하고 심층 신경망의 성능을 향상시켰다. 표 2는 본 연구에서 사용하는 ResNet-20 모델의 학습 파라미터 정보를 정리한다.

본 연구에서는 Skip Connection을 적용하는 Hidden Layer의 최후방 레이어에서 피처 맵을 출력한다. Hidden Layer의 최후방 레이어의 뉴런 활성화 값이 입력 데이터에 대한 레이블 추론에 영향을 미치며, Hidden Layer의 전방부 레이어들의 피처 맵은 최후방 레이어의 피처 맵에 비해 레이블 추론에 영향력이 작다.

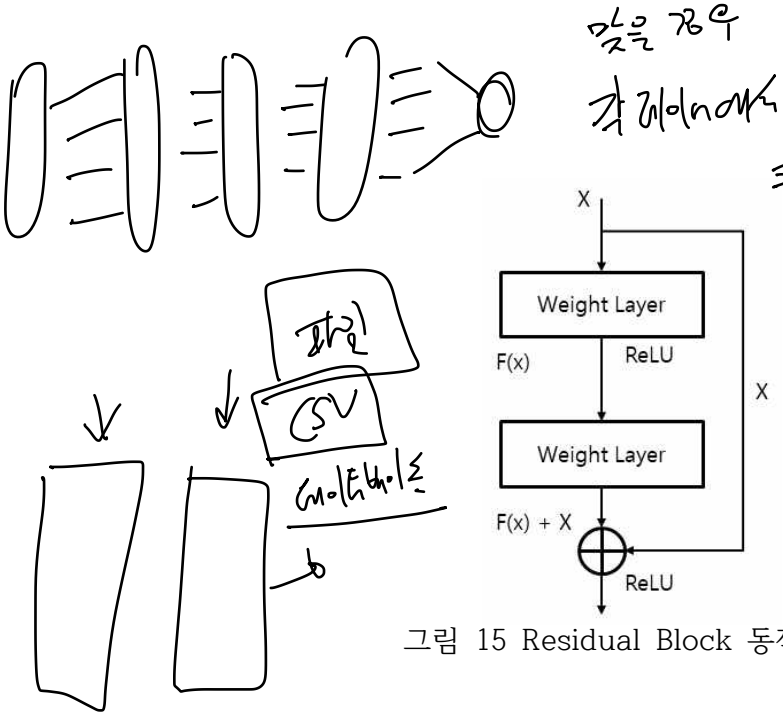


그림 15 Residual Block 동작 구조

표 2 ResNet-20 모델의 학습 파라미터

Data Set	Epoch	Batch Size	Learning Rate	Accuracy
STL10	100	32	0.001	65.1%

다. 테스트 데이터 생성

이미지 분류 모델의 테스트를 위해 데이터 증강을 적용한 테스트 데이터를 생성한다. 표 3은 본 연구에서 적용한 데이터 증강 기법들의 기능과 적용 강도를 정리한다. 6가지 데이터 증강 기법들을 적용 강도를 다르게 적용하여 한 테스트 데이터에 대하여 25가지 변형된 테스트 데이터를 생성한다. 데이터 증강 기법은 imgaug 라이브러리[10]를 활용하였다. 그림 16는 데이터 증강 기법들을 적용한 예시 이미지들이다.

본 연구에서는 데이터 증강 기법을 적용하여 테스트 데이터 셋들을 만든다. 25가지 데이터 증강 기법이 적용된 테스트 데이터 셋은 레이블별로 테스트 데이터 20,000개를 가진다. 실험 데이터들이 특정 데이터 증강 기법에 편향되지 않도록 데이터 증강 기법이 적용된 테스트 데이터 셋으로부터 레이블별로 테스트 데이터 1,000개를 무작위로 선택하여 테스트 데이터 셋을 20개 생성한다.

표 3 적용 데이터 증강 기법 목록

데이터 증강 기법	설명	적용 강도
밝기 증가/감소	이미지의 HSV 공간에서 적용 강도만큼 명도를 증가/감소시킨다.	+30, +60, +90, -30, -60, -90
Gaussian 노이즈	이미지에 적용 Scale 비율에 따른 Gaussian 노이즈를 적용한다.	0.1, 0.2
Salt&Pepper 노이즈	이미지에 적용 점 비율에 따른 Salt&Pepper 노이즈를 적용한다.	3%, 6%
Cutout	이미지의 너비 10%, 높이 10%의 회색 사각형을 적용 강도 개수만큼 이미지 내 임의의 위치에 부착한다.	1, 2, 3, 4, 5
확대/축소	이미지를 적용 Scale 비율만큼 확대/축소한다. 축소 시 빈공간은 흑색 배경을 적용한다.	0.8, 0.9, 1.1, 1.2
회전	이미지를 적용 각도만큼 시계 방향으로 회전시킨다. 회전 시 빈 공간은 흑색 배경을 적용한다.	30˚, 60˚, 90˚, 120˚, 150˚, 180˚

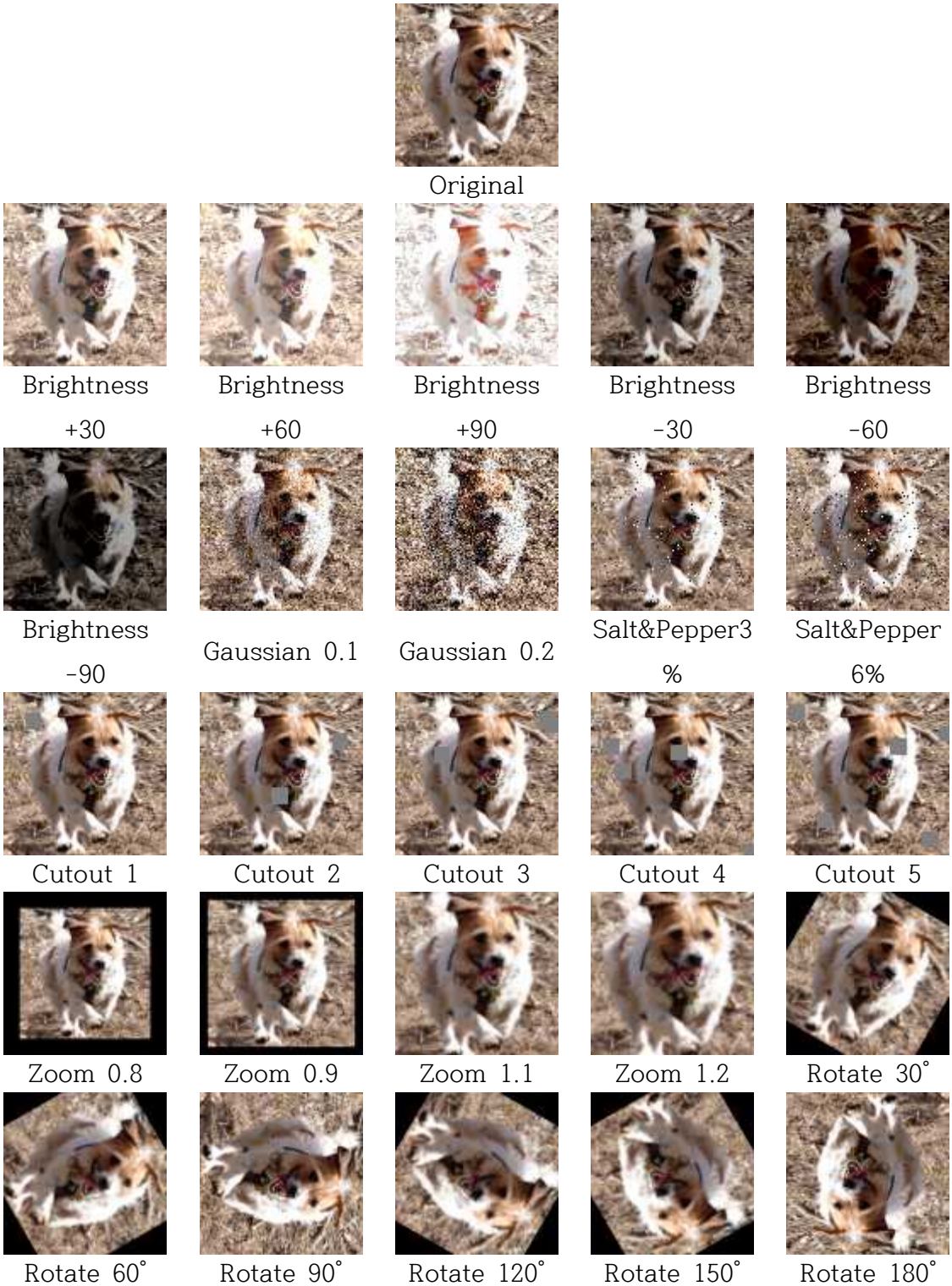


그림 16 데이터 증강 기법 적용 강도별 이미지 예시

라. 테스트 데이터 효과성 평가 기준

본 연구에서는 FMD 기준치로 선정한 테스트 데이터의 효과성을 측정하기 위해 테스트 효과성(Test Effectiveness)을 측정한다. 테스트 효과성은 데이터 셋 중에서 오분류 데이터의 비율이다. 그림 17은 테스트 데이터 셋의 데이터 영역을 표현한다. 식 (11)과 (12)는 테스트 데이터 셋의 데이터들의 관계를 정의하고, 식 (13)과 (14)는 각각 테스트 효과성을 정의한다. FMD 기준치로 FMD 테스트 데이터를 선정하기 이전의 전체 테스트 데이터 셋을 U 테스트 데이터 셋 TS^U로 정의한다. U 테스트 데이터 셋의 데이터 t의 피처 맵 거리 $fmd(t)$가 FMD 기준치 $FMDC$ 이상의 값을 가질 때 데이터 t를 FMD 테스트 데이터 셋 TS^{FMD}으로 선정한다. U 테스트 데이터의 테스트 효과성 TE_{TS^U}는 전체 U 테스트 데이터 중에서 오분류되는 데이터 $Misclassified_{TS^U}$의 비율이고, FMD 테스트 데이터의 테스트 효과성 TE_{FMD}는 FMD 테스트 데이터 중에서 오분류되는 데이터 $Misclassified_{TS^{FMD}}$의 비율이다.

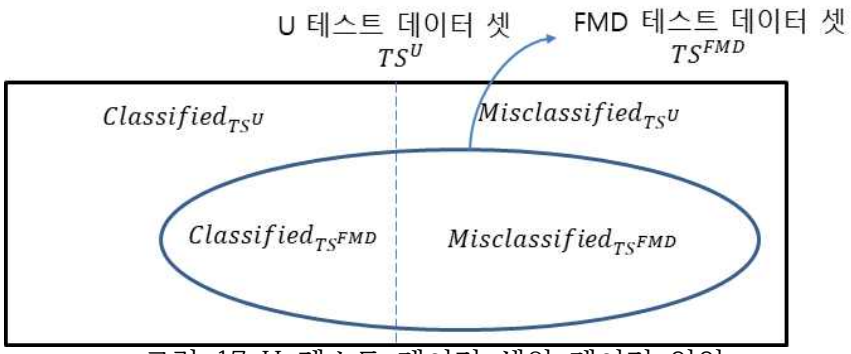


그림 17 U 테스트 데이터 셋의 데이터 영역

$$TS^U = Classified_{TS^U} \cup Misclassified_{TS^U} \quad (11)$$

$$TS^{FMD} = \{t \mid fmd(t) \geq FMDC, t \in TS^U\} \quad (12)$$

$$TE_{TS^U} = \frac{|Misclassified_{TS^U}|}{|TS^U|} \quad (13)$$

$$TE_{TS^{FMD}} = \frac{|Misclassified_{TS^{FMD}}|}{|TS^{FMD}|} \quad (14)$$

본 연구에서는 FMD 테스트 데이터 셋의 테스트 효과성이 U 테스트 데이터 셋의 테스트 효과성보다 더 클 것을 예상한다. FMD 테스트 데이터 셋의 데이터 수는 U 테스트 데이터 셋에 FMD 기준치 미만의 피처 맵 거리를 가지는 테스트 데이터가 존재하는 경

우 U 테스트 데이터 셋의 데이터 수보다 작다. U 테스트 데이터 셋의 모든 데이터가 FMD 기준치 이상의 피처 맵 거리를 가지는 경우에만 FMD 테스트 데이터의 수와 U-테스트 데이터의 수는 동일하다. FMD 테스트 효과성이 U 테스트 효과성보다 큰 경우 FMD-테스트 데이터는 U 테스트 데이터보다 적은 수의 테스트 데이터로 더 많은 비율의 오분류 테스트 데이터들을 포함하고 있다. 따라서 U 테스트 효과성보다 FMD-테스트 효과성이 더 크게 측정된 FMD 테스트 데이터는 U 테스트 데이터보다 심층 신경망을 더 효율적으로 테스트할 수 있다.

2. 실험 결과

가. FMD 기준치 선정

1) 피처 맵 생성

FMD 기준치 선정을 위해 테스트 데이터들의 피처 맵 거리를 측정해야 하며, 피처 맵 거리 측정을 위해 베이스 피처 맵과 테스트 데이터의 피처 맵이 필요하다. 그림 18과 19, 20은 각각 STL10 데이터 셋의 학습 데이터에 대해 ResNet-20 모델이 레이블별로 출력한 베이스 피처 맵, 정분류 테스트 데이터 평균 피처 맵, 오분류 테스트 데이터 평균 피처 맵이다. STL10의 정분류 테스트 데이터의 피처 맵은 베이스 피처 맵과 유사하지만, 오분류 테스트 데이터의 피처 맵은 베이스 피처 맵과 다름을 확인하였다.

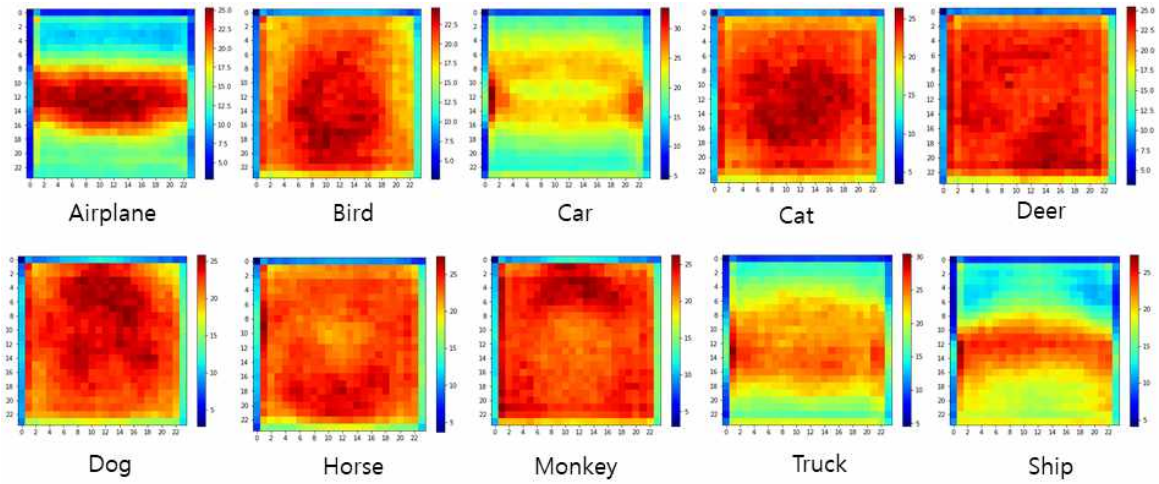


그림 18 STL10 레이블별 베이스 피처 맵

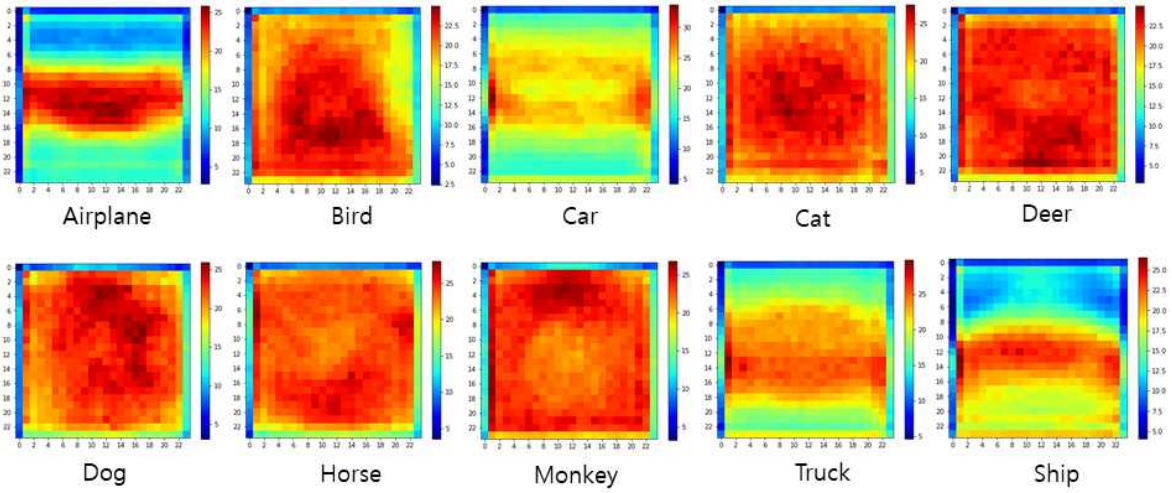


그림 19 STL10 레이블별 정분류 테스트 데이터 평균 피처 맵

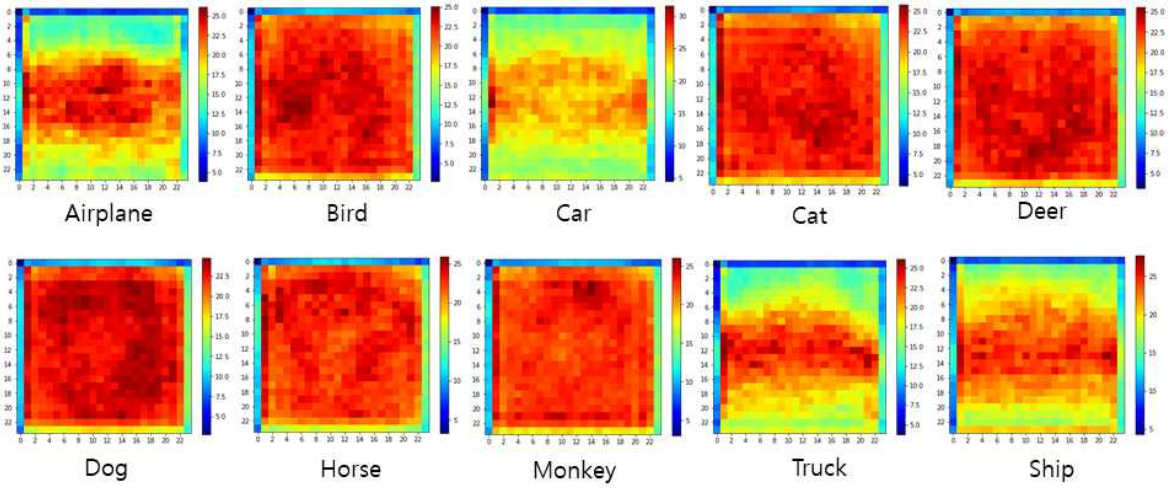


그림 20 STL10 레이블별 오분류 테스트 데이터 평균 피처 맵

2) 피처 맵 거리 측정

피처 맵 거리 측정 방법으로는 L_1 Norm 거리와 L_2 Norm 거리가 있다. 그림 21과 그림 22는 각각 STL10 데이터 셋의 레이블별 원본 테스트 데이터의 L_1 피처 맵 거리 분포 Box plot과 L_2 피처 맵 거리 분포 Box plot이다. 레이블 Airplane과 Car, Deer, Horse, Monkey, Ship, Truck의 테스트 데이터 셋에서 정분류 테스트 데이터의 피처 맵 거리의 중앙값이 오분류 테스트 데이터의 중앙값보다 작게 측정되었다. 피처 맵 거리 측정 방법에 따라 피처 맵 거리 값의 차이가 있었으나 레이블별 정분류 테스트 데이터의 중앙값과 오분류 테스트 데이터의 중앙값 간의 대소 관계는 변하지 않았다.

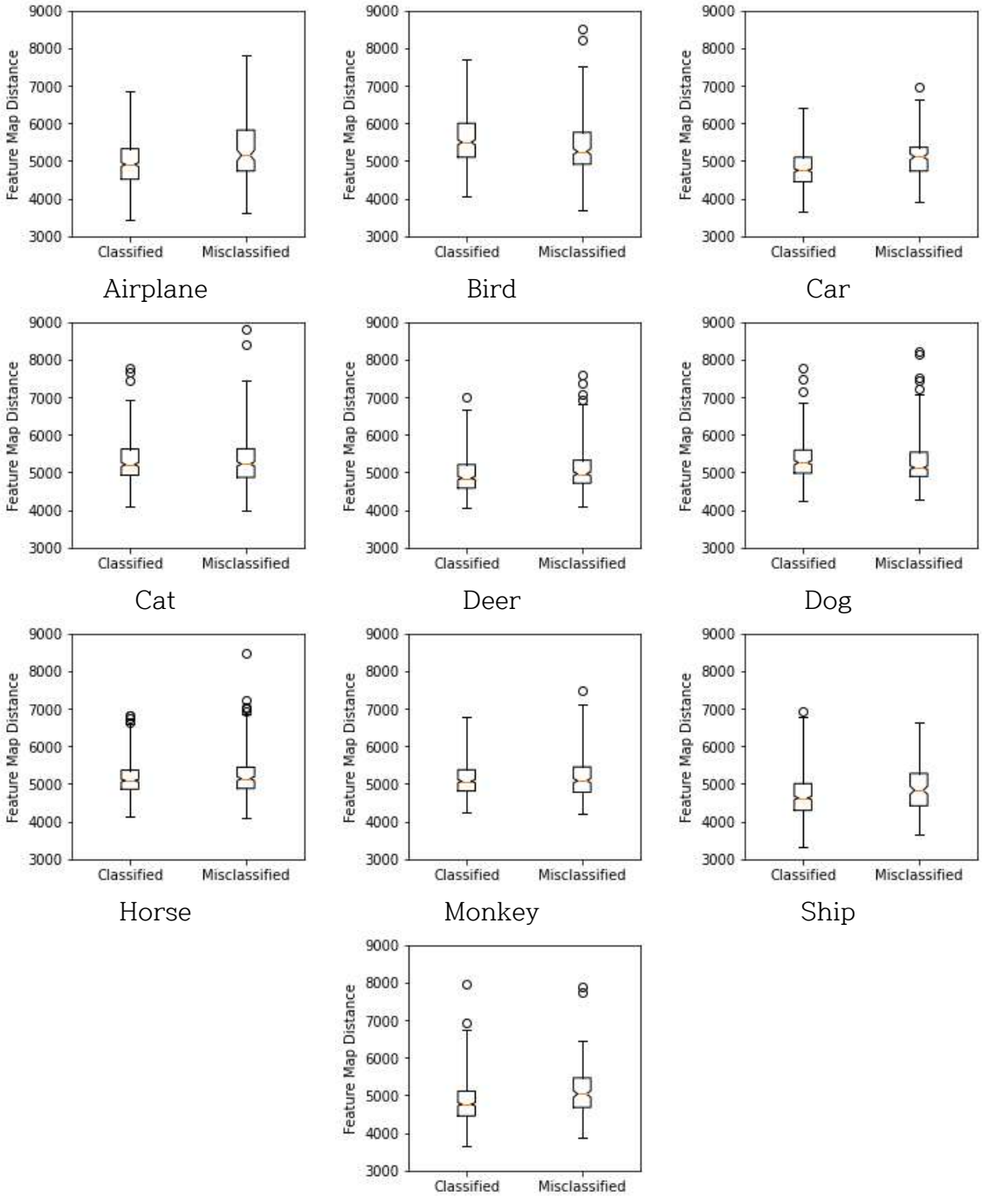


그림 21 STL10 레이블별 L_1 피처 맵 거리 분포

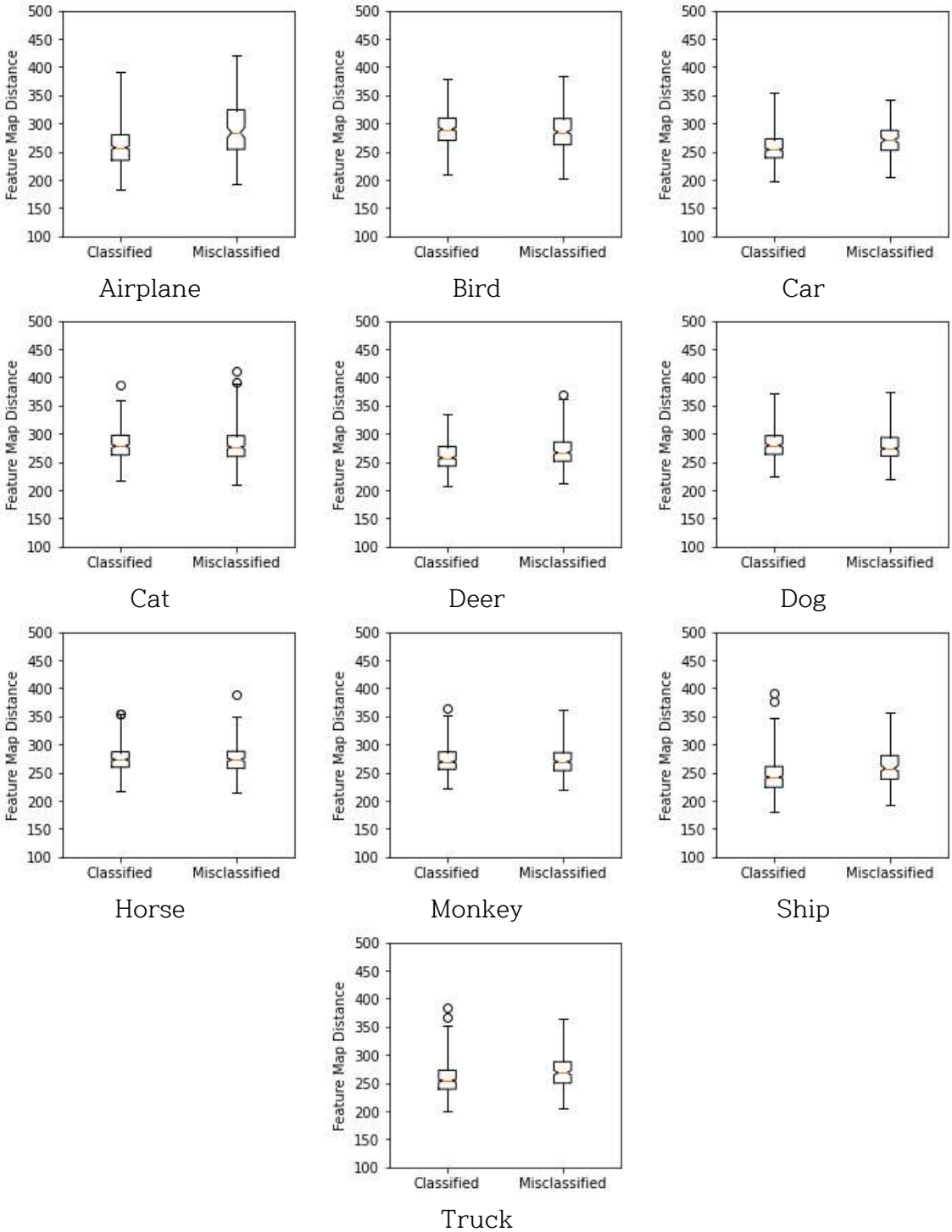


그림 22 STL10 레이블별 L_2 피처 맵 거리 분포

피처 맵 기반 테스트 데이터 선정 방법의 실효성을 확인하기 위해 측정한 테스트 데이터의 피처 맵 거리를 활용하여 테스트 데이터 셋의 피처 맵 거리에 따른 정확도를 분석하였다. 그림 23와 24는 각각 STL10 테스트 데이터의 L_1 피처 맵 거리에 따른 정분류 데이터 비율 그래프와 L_2 피처 맵 거리에 따른 정분류 데이터 비율 그래프이다. 레이블별 테스트 데이터들을 피처 맵 거리에 따라 정렬하여 일정한 수의 데이터 셋으로 나누어 정분류 데이터 비율을 측정하고 상관관계를 분석하였다. 그래프의 추세선은 선형으로 그려졌으며 결정계수 R^2가 0.3 이상인 경우에만 그래프에 추세선을 표시를 하였다. L_1 피처 맵 거리의 경우 6개 레이블에서 테스트 데이터 피처 맵 거리가 커질수록 올바르게 분류되는 데이터의 비율이 낮아지는 추세를 보인다. 반면에 L_1 피처 맵 거리의 경우 10개 레이블 중 5개 레이블에서만 피처 맵 거리가 커질수록 정분류 데이터 비율이 감소하는 추세를 보였다.

피처 맵 거리에 따른 정분류 데이터 비율 그래프를 활용하여 레이블별 테스트 데이터의 피처 맵 거리와 정확도 간의 상관관계를 분석하였다. 표 4와 5는 각각 레이블별 테스트 데이터의 L_1 피처 맵 거리와 정분류 데이터 비율의 피어슨 상관계수와 L_2 피처 맵 거리와 정분류 데이터 비율의 피어슨 상관계수를 측정한 결과이다. L_1 피처 맵 거리로 측정한 경우 Car와 Ship, Truck은 -0.7 이하의 강한 음의 상관관계를 가지며, Airplane과 Deer, Horse는 -0.7 초과 -0.4 이하의 음의 상관관계를 가진다. Bird는 0.4 이상 0.7 이하의 양의 상관관계를 가진다. Cat과 Dog, Monkey는 절대값 0.4 이하의 상관계수를 가져 상관관계를 가진다고 보기 어렵다. 이를 통해 일부 레이블에서 피처 맵 기반 테스트 데이터 선정 방법의 실효성을 확인하였다.

L_2 피처 맵 거리로 측정한 경우 Airplane과 Car, Deer, Ship, Truck 레이블에서 -0.7 이하의 상관계수를 가진다. Horse와 Monkey 레이블에서는 절대값 0.1 이하의 상관계수를 가지므로 상관관계를 가진다고 주장하기 어려우며 나머지 Bird와 Cat, Dog 레이블에서는 양의 상관관계를 가진다. L_2 피처 맵 거리에서는 10개 레이블에서 5개 레이블만 음의 상관관계를 가지므로 L_2 피처 맵 거리를 활용한 테스트 데이터 선정의 실효성을 주장하기 어렵다. 따라서 본 연구에서는 L_1 피처 맵 거리로 FMD 테스트 데이터의 기준이 될 FMD 기준치를 선정한다.

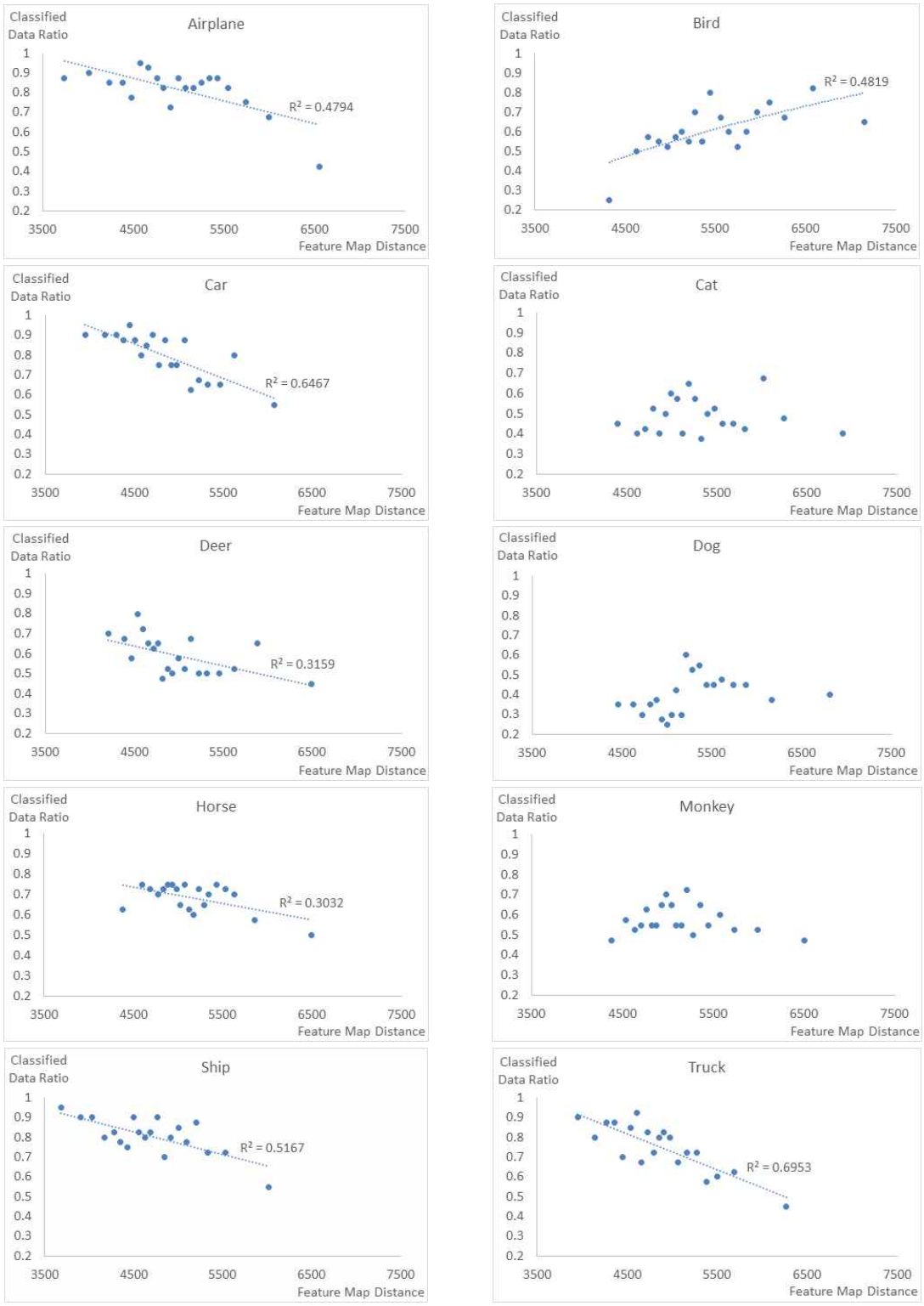


그림 23 STL10 레이블별 L_1 피처 맵 거리-정분류 데이터 분포

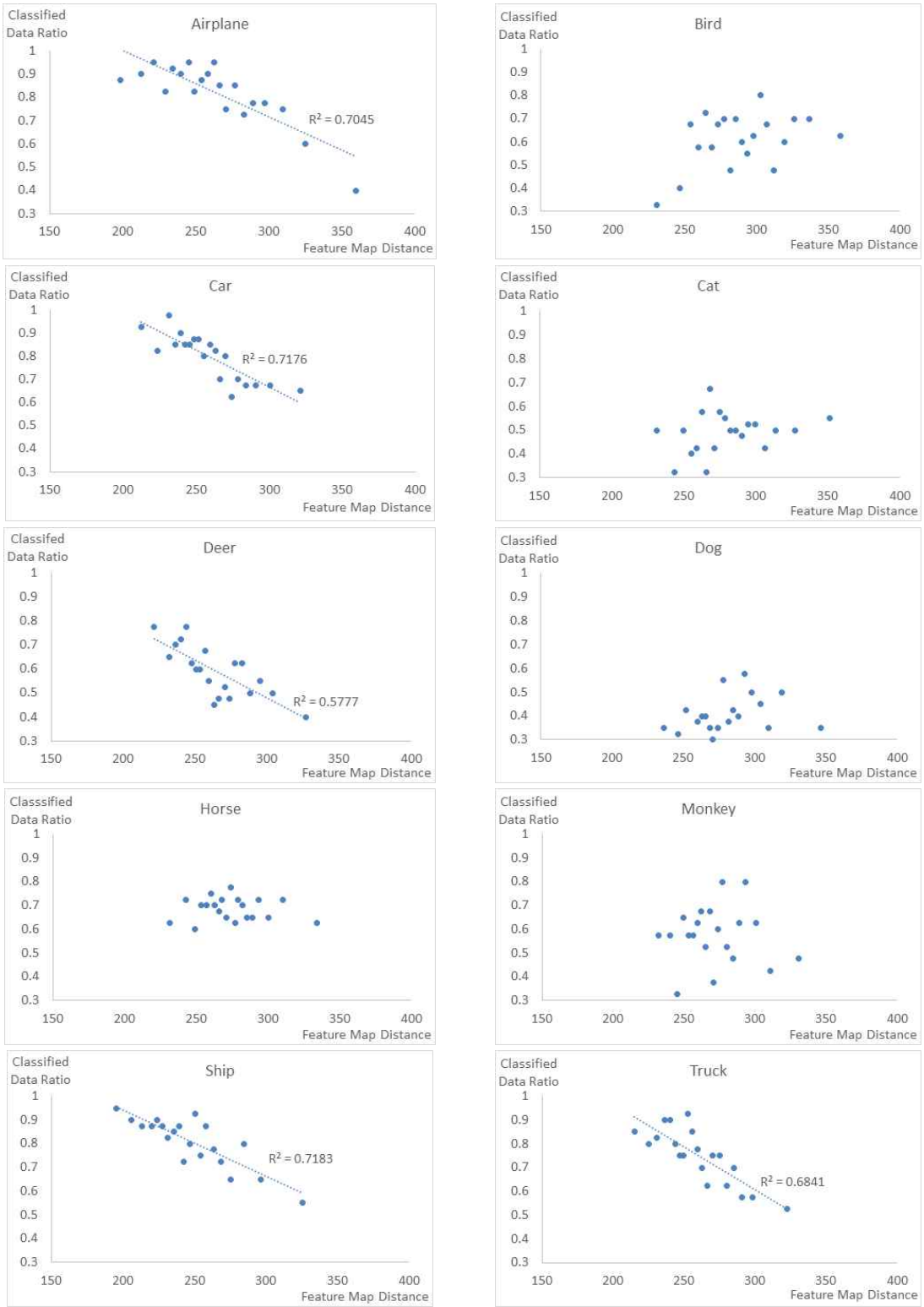


그림 24 STL10 레이블별 L_2 피처 맵 거리-정분류 데이터 분포

표 4 STL10 레이블별 테스트 데이터 L_1 피처 맵 거리-정분류 데이터 비율 상관계수

레이블	Airplane	Bird	Car	Cat	Deer
상관계수	-0.692	0.664	-0.804	0.014	-0.562
레이블	Dog	Horse	Monkey	Ship	Truck
상관계수	0.346	-0.551	-0.177	-0.719	-0.834

표 5 STL10 레이블별 테스트 데이터 L_2 피처 맵 거리-정분류 데이터 비율 상관계수

레이블	Airplane	Bird	Car	Cat	Deer
상관계수	-0.839	0.419	-0.847	0.259	-0.760
레이블	Dog	Horse	Monkey	Ship	Truck
상관계수	0.340	-0.059	-0.026	-0.848	-0.834

3) FMD 기준치 선정

FMD 기준치는 STL10의 오분류된 테스트 데이터의 평균 L_1 피처 맵 거리로 선정하였다. 표 6은 STL10 레이블별 FMD 기준치이다.

표 6 STL10 레이블별 FMD 기준치

레이블	Airplane	Bird	Car	Cat	Deer
FMD 기준치	5284.47	5359.35	5087.53	5315.33	5085.69
레이블	Dog	Horse	Monkey	Ship	Truck
FMD 기준치	5260.51	**5226.36**	5173.53	4883.76	5097.22

나. FMD 테스트 데이터 선정

데이터 증강 기법이 적용된 20개의 테스트 데이터 셋에서 FMD 기준치 이상의 피처 맵 거리를 가지는 데이터들을 FMD 테스트 데이터로 선정한다. 표 7은 테스트 데이터 셋 20개에 대하여 FMD 기준치에 따라 선정된 레이블별 평균 FMD 테스트 데이터 수이다. 레이블별 테스트 데이터 1,000개 중에서 최소 398개의 FMD 테스트 데이터를 선정하였다. Ship 레이블에서 평균 547개로 FMD 테스트 데이터가 가장 많이 선정되었고, Car 레이블에서 평균 426개로 가정 적게 선정되었다.

표 7 레이블별 FMD 테스트 데이터 수

레이블 / 테스트 데이터 셋	Airplane	Bird	Car	Cat	Deer	Dog	Horse	Monkey	Ship	Truck
Data Set1	490	515	404	425	429	476	427	430	517	430
Data Set2	543	497	429	448	417	478	439	431	544	404
Data Set3	510	487	453	420	442	458	454	431	552	429
Data Set4	492	494	418	439	428	461	416	432	537	426
Data Set5	492	505	414	427	422	460	418	439	568	425
Data Set6	496	480	435	430	431	496	396	447	560	445
Data Set7	487	503	441	414	449	478	426	434	544	457
Data Set8	498	455	421	438	414	447	433	458	521	428
Data Set9	409	493	398	452	433	446	424	447	536	409
Data Set10	512	524	443	412	420	490	427	412	546	408
Data Set11	511	481	429	404	425	497	426	429	547	402
Data Set12	500	510	414	427	407	464	420	423	549	427
Data Set13	513	482	455	461	417	469	454	445	546	424
Data Set14	519	497	418	409	425	472	459	462	569	408
Data Set15	506	494	423	421	422	477	437	452	561	427
Data Set16	481	522	427	415	429	479	414	434	547	454
Data Set17	494	458	430	441	429	457	429	460	553	441
Data Set18	481	485	425	414	441	480	475	419	550	443
Data Set19	484	495	434	473	458	476	432	439	531	427
Data Set20	514	541	401	406	438	417	426	435	559	424
평균	497	496	426	429	429	469	432	438	547	427

다. 테스트 효과성 측정

FMD 테스트 데이터의 효과성을 확인하기 위하여 U 테스트 데이터의 테스트 효과성과 FMD 테스트 데이터의 테스트 효과성을 측정하여 비교한다. 표 8은 데이터 증강 기법들이 적용된 20개의 테스트 데이터 셋의 U 테스트 데이터 셋의 테스트 효과성과 FMD 테스트 데이터의 테스트 효과성을 측정한 결과이다. 20개 테스트 데이터 셋에서 U 테스트 효과성보다 FMD 테스트 효과성이 더 크게 측정되었다. 따라서 테스트 데이터 셋 20개의 각 데이터 셋의 전체 데이터보다 FMD 테스트 데이터 셋이 심층 신경망 테스트에 더 적은 데이터 수로 이미지 분류 모델 테스트에 효과적으로 활용할 수 있다.

추가적으로 레이블별로 U 테스트 데이터 셋의 테스트 효과성과 FMD 테스트 데이터의 테스트 데이터 효과성을 측정하여 비교하였다. 그림 25는 레이블별 테스트 데이터 셋의 평균 테스트 효과성 그래프이다. Airplane과 Car, Deer, Horse, Ship, Truck 레이블에서 FMD 테스트 효과성이 U 테스트 효과성보다 크게 측정된 것을 확인할 수 있다.

반면에 Bird와 Cat, Dog, Monkey 레이블에서 FMD 테스트 효과성이 U 테스트 효과성보다 감소한 것을 확인할 수 있다. 이는 L_1 피처 맵 거리의 상관관계 분석에서 Bird와 Cat, Dog, Monkey가 음의 상관관계를 보이지 않은 것과 연관되어 있다고 해석된다.

테스트 효과성 / 테스트 데이터 셋	U 테스트 데이터 셋의 테스트 효과성	FMD 테스트 데이터 셋의 테스트 효과성	변화량
Data Set1	0.537	0.613	+0.076
Data Set2	0.539	0.599	+0.060
Data Set3	0.544	0.609	+0.065
Data Set4	0.540	0.611	+0.071
Data Set5	0.539	0.617	+0.078
Data Set6	0.543	0.606	+0.063
Data Set7	0.543	0.613	+0.070
Data Set8	0.545	0.607	+0.062
Data Set9	0.542	0.603	+0.061
Data Set10	0.544	0.616	+0.072
Data Set11	0.549	0.617	+0.068
Data Set12	0.535	0.602	+0.067
Data Set13	0.549	0.616	+0.067
Data Set14	0.541	0.602	+0.061
Data Set15	0.539	0.601	+0.062
Data Set16	0.538	0.611	+0.073
Data Set17	0.548	0.613	+0.065
Data Set18	0.545	0.611	+0.066
Data Set19	0.540	0.601	+0.061
Data Set20	0.545	0.603	+0.058
평균	0.542	0.609	+0.067

표 8 STL10 데이터 셋별 테스트 효과성

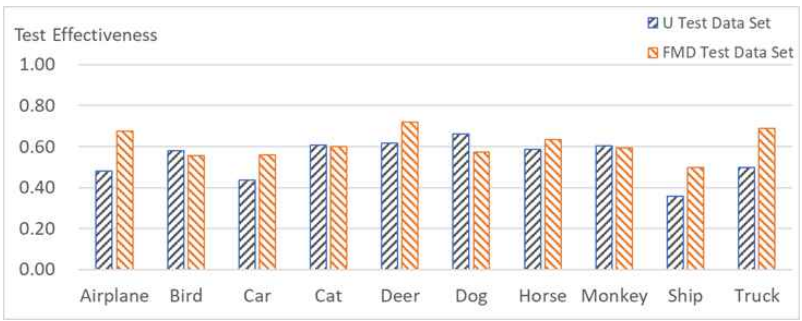


그림 25 레이블별 테스트 데이터 셋 평균 테스트 효과성

V. 관련 연구

본 장에서는 이미지 분류 모델 테스트를 위한 데이터 생성 방법에 대한 기존 연구들을 정리한다. 기존 심층 신경망 테스트 데이터 생성 방법으로 적대적 공격 기법과 데이터 증강 기법 등 여러 데이터 변형 기법들을 기존 테스트 데이터에 적용하거나, 적대적 생성 신경망(Generative Adversarial Network)을 활용하였다.

적대적 공격 기법은 사람이 인지 못하는 미세한 수준으로 데이터를 변형하여 적대적 데이터(Adversarial Example)를 생성하는 기법이다[4]. 적대적 공격 기법의 종류로 FGSM과 C&W 등이 있다. FGSM은 적대적 데이터에 대한 심층 신경망의 취약성의 원인이 모델의 선형성이라고 주장하였다[6]. FGSM은 이미지의 픽셀 값에 대한 손실 함수의 기울기를 기반으로 이미지를 변형하여 적대적 데이터를 생성한다[6]. FGSM이 적용된 MNIST 데이터 셋에 대하여 이미지 분류 모델의 에러율이 증가하여 FGSM의 테스트 데이터 생성 방법의 효과성을 확인하였다[6].

C&W는 원본 데이터로부터 변형된 정도를 L_p Norm 거리로 측정하면서 미세하게 변형시키는 기법이다[7]. 적대적 공격 기법에 대한 방어 기법 Defensive Distillation[11]을 적용한 모델이 C&W를 적용한 테스트 데이터에 대하여 정확도가 감소함으로써 C&W는 이미지 분류 모델 테스트에 효과적인 데이터를 생성함을 확인하였다[7].

데이터 증강 기법은 데이터에 특성을 추가하는 데이터 변형 기법이다. 이미지에 대한 데이터 증강 기법은 이미지의 픽셀 값을 변형하는 기법과 이미지에 아핀 변환을 적용하는 기법으로 분류된다[12]. 픽셀 값을 변형하는 데이터 증강 기법은 밝기 변화와 색상 대조, 노이즈 등 이미지 픽셀의 RGB값이 변경되는 기법이다. 아핀 변환을 적용하는 데이터 증강 기법은 이미지 좌우 반전, 이미지 회전, 이미지 확대/축소 등 이미지의 형태가 변형되는 기법이다.

적대적 생성 신경망은 생성자(Generator)와 판별자(Discriminator) 2가지 심층 신경망 모델을 경쟁시켜 데이터를 생성하는 모델이다[13]. 생성자는 실제 데이터와 유사한 거짓 데이터를 생성하는 모델이고, 판별자는 실제 데이터와 생성자가 생성한 거짓 데이터를 판별하는 모델이다. 생성자와 판별자가 서로 적대적으로 거짓 데이터를 생성하고 판별하는 과정을 반복하여 학습을 하면 생성자는 판별자가 거짓 데이터로 판별하지 못하는 데이터를 생성할 수 있다. 그러나 적대적 생성 신경망은 적대적 공격 기법과 데이터 증강 기법에 비해 데이터 생성 과정이 복잡하며 생성자 학습이 잘못된 경우 유사한 데이터들만 생성하거나 생성하고자 하는 레이블의 특성이 없는 데이터가 생성될 수 있다.

Surprise Adequacy는 학습 데이터의 활성화 자취(Activation Trace)와 테스트 데이터의 활성화 자취의 차이를 Surprise Adequacy로 측정하는 방법이다[14]. 측정 방법으로 Likelihood-based Surprise Adequacy(LSA)와 Distance-based Surprise Adequacy(DSA)가 있다. LSA는 Keneral Density Estimation(KDE)[15]을 사용하여 입력 데이터와 학습 데이터의 확률 밀도를 측정하고, DSA는 입력 데이터와 근접한 2개 레이블의 학습 데이터들 간의 일부 뉴런 활성화 값들의 거리의 비율을 비교한다. 데이터들의 LSA와 DSA로 높은 Surprise Coverage가 측정된 데이터들을 선정한다.

본 논문에서 제안한 피처 맵 기반 테스트 데이터 선정 방법은 학습 데이터와 테스트 데이터 간의 피처 맵 거리를 기준으로 테스트 데이터를 선정한다. 기존 테스트 데이터 선정 방법 Surprise Adequacy와는 뉴런 활성화 값을 기반으로 데이터를 선정하는 공통점이 있으나, Surprise Adequacy는 특정 뉴런들을 대상으로 입력 데이터의 Surprise Coverage를 측정하여 선정하고 피처 맵 기반 테스트 데이터 선정 방법은 특정 레이어의 뉴런 활성화 값을 기반으로 FMD 기준치에 따라 데이터를 선정하는 차이점이 있다.

VI. 결론 및 향후 연구

이미지 분류 모델 테스트를 위한 테스트 데이터 생성 과정에는 효과적인 테스트 데이터를 선별할 기준이 필요하다. 기존 이미지 분류 모델 테스트에서는 테스트 데이터 생성 기법으로 적대적 공격 기법과 데이터 증강 기법을 사용하였다. 그러나 무분별한 데이터 변형으로 이미지 분류 모델 테스트에 효과적이지 않은 데이터가 생성될 수 있다.

본 연구에서는 이미지 분류 모델의 성능에 뉴런 활성화 값이 영향을 미치므로 데이터의 피처 맵을 기반으로 테스트 데이터를 선정하는 방법을 제안하였다. 피처 맵은 레이어별로 출력된 뉴런 활성화 값의 집합이다. 학습 데이터의 피처 맵이 정분류 테스트 데이터의 피처 맵과 유사하지만, 오분류 테스트 데이터와는 뉴런 활성화 값의 크기와 분포에서 차이가 있음을 확인하였다. 학습 데이터의 피처 맵과 테스트 데이터의 피처 맵 간의 거리를 측정하고, 측정된 피처 맵 거리로부터 FMD 기준치를 선정한다. FMD 기준치보다 더 큰 피처 맵 거리를 가지는 테스트 데이터들을 이미지 분류 모델 테스트에 효과적인 데이터로 선정한다.

사례 연구로 STL10 데이터 셋과 ResNet-20 모델을 대상으로 FMD 테스트 데이터를 선정하고 테스트 효과성을 측정하였다. STL10 데이터 셋으로부터 베이스 피처 맵과 테스트 데이터들의 피처 맵을 생성하여 동일한 레이블의 베이스 피처 맵과 정분류 테스트 데이터의 피처 맵은 유사함을 보였지만, 베이스 피처 맵과 오분류 테스트 데이터의 피처 맵은 뉴런 활성화 값에서 차이가 있음을 확인하였다.

테스트 데이터의 피처 맵과 베이스 피처 맵 간의 피처 맵 거리는 피처 맵 간의 L_1 Norm 거리와 L_2 Norm 거리로 측정하였다. L_1 피처 맵 거리의 경우 STL10의 10개 레이블 중 6개 레이블에서 오분류 테스트 데이터의 피처 맵 거리 중앙값이 정분류 테스트 데이터의 피처 맵 거리 중앙값보다 더 크게 측정됨을 확인하였다. L_2 피처 맵 거리는 L_1 Norm 거리로 측정한 테스트 데이터의 피처 맵 거리보다 작아졌지만 L_1 Norm 거리의 경우와 동일하게 10개 레이블 중 6개 레이블에서 오분류 테스트 데이터의 피처 맵 거리 중앙값이 정분류 테스트 데이터의 피처 맵 거리 중앙값보다 더 크게 측정되었다.

피처 맵 기반 테스트 데이터가 검증하기 위하여 테스트 데이터의 피처 맵 거리와 정확도 간의 상관관계를 분석하였다. L_1 피처 맵 거리의 경우 6개 레이블에서 피처 맵 거리와 정확도 간에 음의 상관관계를 확인하였다. 그러나 L_2 피처 맵 거리의 경우 5개 레이블에서 피처 맵 거리와 정확도 간에 음의 상관관계를 확인하였다. L_2 피처 맵 거리는

10개 레이블 중 5개의 레이블만 음의 상관관계를 가지므로 피처 맵 거리 측정 방법으로 부적합하다고 판단하였다. 이에 FMD 기준치는 원본 테스트 데이터 중 오분류 테스트 데이터의 평균 L_1 피처 맵 거리로 선정하였다.

실험에 편향되지 않은 테스트 데이터 셋을 사용하기 위해 6종류 데이터 증강 기법이 여러 적용 강도로 적용된 STL10 테스트 데이터 셋으로부터 임의로 레이블별 1,000개 테스트 데이터로 구성된 20개 테스트 데이터 셋을 생성하였다. 20개 테스트 데이터 셋에 대하여 FMD 기준치 이상의 피처 맵 거리를 가지는 데이터들을 FMD 테스트 데이터로 선정하였고 U 테스트 효과성과 FMD 테스트 효과성을 비교하여 FMD 테스트 데이터의 테스트 효과성이 평균적으로 더 크게 측정됨을 확인하여 피처 맵 기반 테스트 데이터 선정 방법의 효과성을 확인하였다. 레이블별로 U 테스트 효과성과 FMD 테스트 효과성을 비교한 결과 10개 레이블 중 6개 레이블에서 테스트 효과성이 증가하였으며, 테스트 효과성이 감소한 4개 레이블은 테스트 데이터의 피처 맵 거리와 정확도가 음의 상관관계를 가지지 않았던 레이블들이다.

본 연구에서는 데이터 증강 기법이 적용된 20개 테스트 데이터 셋에서 FMD 테스트 데이터의 테스트 효과성이 증가하였지만, 레이블별로 테스트 효과성을 측정하였을 때 10개 레이블 중 6개 레이블에서만 FMD 테스트 데이터의 테스트 효과성을 확인할 수 있었다. 향후 연구로 FMD 테스트 효과성이 감소한 4개 레이블들에서 테스트 효과성을 향상시키기 위해 피처 맵에서 큰 뉴런 활성화 값을 가진 뉴런들을 선정하여 피처 맵 거리를 측정한다. 본 연구에서는 한정된 데이터 셋과 이미지 분류 모델을 대상으로 실험을 하여 다른 데이터 셋과 다른 이미지 분류 모델을 대상으로 피처 맵 기반 테스트 데이터 생성 방법의 효과성을 확인한다.

참고문헌

[1] A. Krizhevsky et al., "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, 2012.

[2] C. Szegedy et al., "Going Deeper with Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9, 2015.

[3] K. He ea al., "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[4] C. Szegedy et al. "Intriguing Properties of Neural Networks," *arXiv preprint arXiv:1312.6199*, 2013.

[5] K. Pei et al., "DeepXplore: Automated WhiteBox Testing of Deep Learning Systems," in *Proceedings of the 26th Symposium on Operating Systems Principles*, pp. 1-18, 2017.

[6] I. Goodfellow et al. "Explaining Harnessing Adversarial Examples," *arXiv preprint arXiv:1412.6572*, 2014.

[7] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," *IEEE Symposium on Security and Privacy*, pp. 39-57, 2017.

[8] L. Ma et al., "DeepGauge: Multi-Granularity Testing Criteria for Deep Learning Systems," in *The 33rd IEEE/ACM International Conference on Automated Software Engineering*, pp. 120-131, 2018.

[9] A. Coates et at., "An Analysis of Single Layer Networks in Unsupervised Feature Learning," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011.

[10] A. Jung, "imgaug Documentation," *Readthedocs.io*, 2019.

[11] N. Papernot et al., "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks," in *2016 IEEE Symposium on Security and Privarcy*, pp. 582-597, 2016.

[12] X. Xie et al., "DeepHunter: A Coverage-Guided Fuzz Testing Framework for Deep Neural Networks," in *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis*, pp. 146-157, 2019.

[13] I. Goodfellow et al., "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems 27*, 2014.

[14] J. Kim et al., "Guiding Deep Learning System Testing using Surprise Adequacy," in *The 41st ACM/IEEE International Conference on Software Engineering*, 2019.

[15] M. P. Wand and M. C. Jones, "Kernel Smoothing," *Chapman and Hall/CRC*, 1994.

Effective Data Selection Method for Image Classification Model Testing

Young-Won Choi

Department of Information Convergence Engineering
The Graduate School Pusan National University

Abstract

As the performance of Deep Neural Networks(DNNs) has been improved, DNNs have been utilized in variety fields. However, The vulnerability of image classification models against Adversarial Attack was discovered, and human/capital damages can be occurred when image classification models which have been used in Safety-Critical System are defected. Therefore, test data generation methods for testing image classification models have been researched.

Conventional researches about testing image classification model generated test data by adversarial attack techniques and data augmentation techniques. But effective data for testing image classification models could not be generated due to indiscreet data transformations. Therefore a criteria to select effective test data for testing image classification model is needed.

In this research, I suggested an feature map-based effective data selection method for testing image classification models. The neuron activation values of the model against input data affect the inference of image classification model, and neuron activation values in a layer of model could be formed a feature map. Feature map distance was defined as a calculation method for the distances of neuron activation values between training data's feature maps and test data's feature maps. FMD Criteria is a baseline feature map distance for selecting effective test data. FMD Test Data is test data which feature map distance is bigger than FMD Criteria.

As a case study, a research using STL10 data set and ResNet-20 proceeded. Negative correlations between feature map distance and accuracy were observed in 6 labels of STL10. Average feature map distance of misclassified original test data

of STL10 was selected as FMD Criteria. FMD test data were selected in 20 STL10 test data sets which were applied data augmentation techniques, and it was observed that test effectiveness of FMD test data was bigger than test effectiveness of augmented test data Set.