Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.31, NO.1, Feb. 2021 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2021.31.1.51

입력 변이에 따른 딥러닝 모델 취약점 연구 및 검증*

김 재 욱,[1†] 박 래 현,[1] 권 태 경[2‡]
[1,2]연세대학교 정보대학원 정보보호 연구실 (대학원생, 교수)

Analysis of Deep Learning Model Vulnerability According to Input Mutation*

Jaeuk Kim,[1†] Leo Hyun Park,[1] Taekyoung Kwon[2‡]
[1,2]Information Security LAB, GSI, Yonsei University (Graduate student, Professor)

요 약

딥러닝 모델은 변이를 통해 훈련 데이터에서 벗어난 입력으로부터 잘못된 예측 결과를 산출할 수 있으며 이는 자율주행, 보안 분야 등에서 치명적인 사고로 이어질 수 있다. 딥러닝 모델의 신뢰성 보장을 위해서는 다양한 변이를 통해 예외적인 상황에 대한 모델의 처리 능력이 검증되어야 한다. 하지만, 기존 연구가 제한된 모델을 대상으로만 수행되었으며, 여러 입력 변이 유형에 구분을 짓지 않고 사용했다. 본 연구에서는 딥러닝 검증 데이터 세트로 널리 사용되고 있는 CIFAR10 데이터 세트를 기반으로 다양한 상용화된 모델과 추가 버전을 포함하여 총 6개의 모델에 대한 신뢰성 검증을 수행한다. 이를 위해 실생활에서 발생할 수 있는 6가지 유형의 입력 변이 알고리즘을 다양한 파라미터와 함께 데이터 세트에 개별적으로 적용하여 각각에 대한 모델의 정확도를 비교함으로써 특정 변이 유형과 관련된 모델의 취약점을 구체적으로 파악한다.

ABSTRACT

The deep learning model can produce false prediction results due to inputs that deviate from training data through variation, which leads to fatal accidents in areas such as autonomous driving and security. To ensure reliability of the model, the model's coping ability for exceptional situations should be verified through various mutations. However, previous studies were carried out on limited scope of models and used several mutation types without separating them. Based on the CIFAR10 data set, widely used dataset for deep learning verification, this study carries out reliability verification for total of six models including various commercialized models and their additional versions. To this end, six types of input mutation algorithms that may occur in real life are applied individually with their various parameters to the dataset to compare the accuracy of the models for each of them to rigorously identify vulnerabilities of the models associated with a particular mutation type.

Keywords: Deep learning, Mutation, Adversarial machine learning

I. 서 론

최근 딥러닝은 음성인식 [3], 이미지 분류 [1,2]를 포함한 다양한 연구 분야에서 인간이 인식하는 수준을 달성하거나 넘어서는 엄청난 발전이 진행되고 있다. 이러한 성과를 토대로 딥러닝은 자율주행 [5], 항공기 충돌 방지 시스템 [4], 악성 코드 탐지 [6]와 같은 보안 및 안전에 중요한 시스템에서 또한 광

Received(11. 26. 2020), Modified(01. 05. 2021), Accepted(01. 06. 2021)
* 본 연구는 2020년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2019R1A2C1088802).
† 주저자, freak0wk@yonsei.ac.kr
‡ 교신저자, taekyoung@yonsei.ac.kr(Corresponding author)

Fig. 1. Mutated samples with 6 mutation types from the same original image in CIFAR10 dataset

범위하게 사용되고 있다.

딥러닝 모델들은 일반적으로 많은 양의 데이터를 학습하여 수많은 레이어와 노드들이 업데이트를 진행함으로써 분류 정확도를 높이는 방식을 사용한다. 이러한 학습데이터에 의존적인 특성으로 인해 학습량이 적은 변이가 적용된 데이터가 입력되는 경우 잘못된 분류 결과가 발생할 수 있다. 예를 들어, 이미지 데이터 세트에서는 물체에 그림자가 드리워져 있는 경우, 강한 햇빛으로 인해 물체가 밝아지는 경우, 멀리 있는 물체가 원래의 형태보다 작게 보이는 경우, 물체가 기울어져 있는 경우 등에 의해 실생활에서 오분류가 발생할 수 있다. 실제 사례 [10]에서 볼 수 있듯이 딥러닝 시스템의 잘못된 분류는 생명 및 보안 관련 분야에서 더욱 치명적이다.

이에 따라, 입력 변이를 통해 모델을 검증하여 딥러닝 시스템의 성능을 확인하고 취약성을 개선하여 모델의 견고성을 높이는 것은 필수적이다. 기존의 변이를 활용한 모델 검증 연구 [7,8,9]는 다양한 딥러닝 모델을 사용하지 않았고 상용화되지 않은 모델을 선정하였으며 여러 변이들을 겹쳐 사용함으로써 변이별 모델의 결과를 확인할 수 없다는 점에서 한계가 있다.

이에 본 논문에서는 상용화된 여러 딥러닝 모델의 변이 별 분류 정확도를 측정하여 모델의 취약점을 비교 분석했다. 이를 수행하기 위해 DeepHunter[7]에서 제공하는 translation, scale, shear, rotation, contrast, brightness 총 6개의 변이 유형을 사용하여 샘플을 만들었고 나아가 다양한 변이 파라미터를 적용함으로써 그에 따른 모델의 분류 결과를 확인하였다. 데이터 세트는 딥러닝 연구 분야에서 학습데이터로 통용되고 있는 CIFAR10 데이터 세트를 사용하였다. 검증 대상 딥러닝 모델은 VGG, LeNet, MobileNet로 구성하여 각각의 차이를 확인하였다. VGG 모델은 파악하기 쉬운 아키텍쳐로 구성되어 구현이 쉽다는 장점에도 불구하고 좋은 성능을 보여주고 있으며, LeNet은 CNN 알고리즘을 최초로 사용한 모델이기 때문에 두 모델이 성능 비교의 기준으로 사용되고 있다. MobileNet은 자원이 제한적인 곳에서 사용 목적으로 개발되어 자동차 및 드론과 같은 분야에서 사용되고 있다. 이에 따라, 검증 모델로 MobileNet을 선정하여 자율주행 및 항공기 충돌 방지 시스템에서 모델의 신뢰도를 제공할 수 있다. 또한 각 동일한 모델에서 추가적인 노드 및 레이어에 따른 분류 결과 차이를 확인하기 위해 모델 버전을 추가하여 실험을 수행했다. Fig. 1.은 총 6개의 변이 유형에 의한 이미지 변화 예시를 보여준다. 이러한 다양한 조건을 설정하고 실험을 수행하여 다양한 결과를 통해 변이에 따른 모델의 정확도를 분석했다. 이를 통해 검증 모델에 정확도를 낮추는 변이를 선별하여 특정 모델에 취약한 변이를 구체적으로 파악할 수 있어 딥러닝 모델 자체에 대한 공격 기법의 안정성을 검증할 수 있는 기반연구로 가치가 있다.

본 논문의 구성은 다음과 같다. 2장에서는 변이에 대한 모델 분석의 필요성 및 기존 연구에 대해 설명한다. 3장에서는 연구 질문 및 진행한 실험 환경을 다루고 4장에서 연구 질문에 따른 실험 결과 및 분석을 설명한다. 5장에서는 기존 연구들과 본 논문을 비교하고 6장에서 결론 및 향후 연구를 제시한다.

II. 연구 배경

2.1 딥러닝 모델 분석의 필요성

딥러닝 모델의 동작은 주로 네트워크의 가중치로 부터 영향을 받아 분류를 수행한다. 네트워크의 가중치는 훈련 데이터 세트에 대한 학습을 통해 얻어지기 때문에 훈련 데이터 세트는 모델 분류 결과에 큰 영향을 미친다 [8]. 이는 학습된 모델에 대해 훈련되지 않은 변이된 데이터를 입력하여 잘못된 분류 결과가 발생하는 문제로 이어질 수 있다. 이에 대한 예시로 Fig. 2.는 Airplane 클래스의 원본 이미지에 Shear 변이를 적용하여 Ship 클래스로 오분류됨을

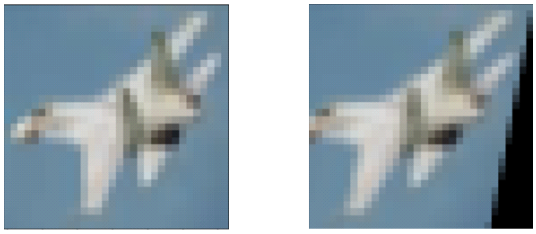

Fig. 2. An original image in CIFAR10 dataset and its mutated samples causing misclassification through shear mutation

보여준다.

위의 예시와 같은 변이로 인한 오분류 발생 가능성으로부터 딥러닝 모델을 보호하고자 다양한 방어 기법들이 소개되었다. 그 중에서도 적대적 재학습과 적대적 탐지는 선제적으로 변이 이미지를 탐지하는 과정을 거친다. 적대적 재학습은 탐지된 변이 이미지를 학습 데이터 세트에 포함시켜 딥러닝 모델을 재학습시키는 기법이다 [11]. 적대적 탐지는 변이 이미지를 통해 탐지기를 학습시킴으로써 모델 입력 전 단계에서 의심스러운 입력을 차단하는 기법이다 [12]. 하지만 기존의 방어 기법은 모델의 다양한 취약점을 고려하지 않고 무분별하게 재학습 데이터를 생성했다. 딥러닝 모델의 취약점이 변이 유형에 따라 다르게 나타날 수 있기 때문에 보다 효율적인 방어를 위하여 다양한 실험을 통해 더욱 심각한 취약점을 우선적으로 발견하고 대응하는 노력이 필요하다.

2.2 이미지 변이에 기반한 딥러닝 모델 분석

소프트웨어 검증에서 변이 테스트는 프로그램의 품질을 평가하고 취약점을 찾아내는 중요한 기법이다. 기존 검증 방식을 착안하여 딥러닝 연구 분야에서도 딥러닝 모델을 검증할 수 있다. 다시 말하면, 학습된 모델에 입력 변이를 주입하여 모델의 정확도를 분석함으로써 취약점을 발견할 수 있다. 이미지 변이는 이미지 분류 작업에 많은 가변성을 제공한다. 이를 활용하여 DeepMutation [8]에서는 테스트 데이터 품질을 평가하기 위해 source level 또는 model level에서 입력 변이 테스트를 수행하였다. 그러나 일반적으로 사용되고 있지 않은 3개의 모델에 대해 검증을 수행하였다. 이는 여러 딥러닝 연구 분야에서 상용화되고 있는 모델의 입력 변이 결과를 확인하기 어렵고 다양한 모델에 대한 취약점을 파악하기 어려운 한계점이 있다.

DeepTest [9]는 테스트세트에서 1,000개의 입력 이미지를 무작위로 선택하고 Translation, Scale, Shear, Rotation, Contrast, Brightness 등의 변이를 사용하여 각 입력 이미지를 변환하였다. 또한 변이의 매개 변수를 변경하여 새로운 합성 이미지를 생성하였다. 변이 이미지를 입력으로 사용하여 3개 모델에서 실험을 실행하고 각 입력에 의한 활성화 된 뉴런을 기록하였다. DeepTest와 달리 우리는 더 다양한 모델을 검증하였고 학습데이터로 통용되고 있는 CIFAR10을 사용하여 여러 경우의 모델 정확도를 확인할 수 있는 차이점이 있다.

DeepHunter [7]는 Translation, Scale, Shear 등 아핀 변환과 픽셀값 변환을 적용하고 퍼징을 활용하여 딥러닝 모델 검증을 수행하였다. 하지만 여러 변이를 겹쳐 이미지에 적용하였기 때문에 각 변이에 대한 모델의 결과를 확인할 수 없다.

III. 입력 변이에 따른 모델 취약점 분석 방법

3.1 연구 질문 및 방법

본 연구에서는 입력 변이와 딥러닝 모델의 취약점 사이의 관계를 확인하기 위하여 세 가지 연구 질문을 수립하고 그에 따른 구체적인 실험 방법을 설계하여 연구를 진행한다.

연구 질문 1) 변이 유형에 따른 모델의 취약점 차이가 존재하는가?

Fig. 1.에서 확인할 수 있듯이 각 변이에 따라 이미지는 다양하게 변환하고 학습데이터에 의존적인 딥러닝 모델은 이미지 변화에 의해 오분류를 일으킬 수 있다. 더불어 이미지 변환 중 모델의 결과에 더 큰 영향을 미치는 변환이 존재할 것이다. 이 두 가지 질문에 대한 답을 확인하기 위해 변이의 유형을 구분하여 실험을 수행하였고 가장 큰 영향을 미치는 변이를 확인하기 위해 총 6개의 변이를 적용한 결과를 비교 및 분석한다.

연구 질문 2) 각 변이 유형에 대해 모델의 서로 다른 버전 간의 정확도 차이가 존재하는가?

Table 1. Details of the dataset, DNN models and mutation types

Dataset	Model	Mutation
CIFAR 10	VGG16, VGG19, LeNet-1, LeNet-5, MobileNetV1, MobileNetV2	Scale, Shear, Rotation, Contrast, Brightness, Translation

모델의 새로운 버전은 기존 모델의 구조에서 노드 및 레이어를 추가하거나 새로운 기법을 적용하여 모델 구조를 일부 수정하는 것이 일반적이다. 딥러닝 모델의 노드 및 레이어 추가, 모델 일부 수정이 입력 변이의 모델 정확도에 긍정적인 영향을 끼치는지 확인하기 위해 3개의 모델에 대해 각 1개의 버전을 추가하여 총 6개의 딥러닝 모델을 대상으로 실험을 수행한다.

연구 질문 3) 변이 파라미터에 따른 모델의 취약점 차이가 존재하는가?

변이 파라미터에 변동을 주어 이미지 변화의 강도를 제어할 수 있다. 변이는 밝기 조절, 크기 조절 등으로 이미지의 픽셀값에 변화를 줌으로써 픽셀값을 입력 받아 이미지 분류를 수행하는 딥러닝 모델의 정확도를 파라미터의 변화에 따라 변화시킬 것이다. 이에 따라 6개의 변이의 파라미터에 변화를 주어 모델의 정확도 변화를 분석한다.

3.2 실험 설계

실험은 CIFAR10의 테스트 데이터 세트의 각 샘플에 6개의 변이를 적용하여 모델에 입력 후 정확도

차이를 각 연구 질문에 맞게 비교 분석함으로써 수행되었다. 총 6개의 변이는 Table 1.에서 확인할 수 있다. Scale은 이미지의 크기를 변화시키는 변이이며 이미지를 확대 혹은 축소 시킬 수 있다. Shear는 x, y 축을 기준으로 가로 또는 세로의 방향으로 변화를 준다. Rotation은 원점을 기준으로 회전을 주는 변이를 말한다. Contrast, Brightness는 대비와 명도를 말하며, Translation은 이미지의 위치를 변화시키는 변이를 뜻한다. 과도한 파라미터는 원본 이미지의 의미를 훼손시킬 수 있으므로 파라미터 범위에 제한을 두어 실

Table 2. Type of mutations and their parameter ranges

Mutation	Parameter ranges
Scale	0.7 ~ 1.2
Shear	-0.6 ~ 0.6
Rotation	-50 ~ 50
Contrast	0.2 ~ 2
Brightness	-80 ~ 80
Translation	-3 ~ 3

험을 수행하였다. 개별적인 변이와 파라미터에 대한 모델의 정확도를 확인하고자 했기 때문에 변이의 중복은 허용하지 않았으며, 동일한 변이와 파라미터는 1개의 원본 이미지에 한 번만 적용하였다.

3.3 실험 환경

Table 2.에서 실험에 사용한 데이터 세트, 딥러닝 모델, 변이 유형을 확인할 수 있다. 데이터 세트는 오픈소스로 제공되고 있고 여러 딥러닝 연구 분야에서 학습 데이터로 많이 사용되고 있는 CIFAR10을 사용하였다. CIFAR10은 10개의 클래스와 각 클래스당 6000개의 이미지로 구성되어 있다. 총 60,000개의 이미지는 32x32x3의 차원수를 가진 RGB 컬러 이미지이며 5개의 학습 세트와 1개의 테스트 세트로 구성된다. 우리는 10,000개의 테스트 이미지에 변이를 적용하여 모델의 정확도를 분석하였다.

모델 간의 차이와 동일 모델의 버전에 따른 차이를 분석하기 위해 VGG, LeNet, MobileNet 모델의 버전을 추가하여 실험을 수행하였다. 모델 학습 및 테스트 정확도는 Table 3.에서 확인할 수 있다. 학습 세트에서는 LeNet-1이 99.46%로 가장 높게 나왔으며, 테스트 세트에서는 VGG19가 92.62%로 다른 모델들 보다 월등히 높은 정확도를 보여주었다.

실험은 모든 모델에서 테스트 세트 10,000의 이미지 중 공통적으로 정상 분류된 이미지 1,970개를 추출하여 변이를 적용하고 실험을 수행하였다. 변이 유형으로는 DeepHunter [7]에서 제공하는 6개를 사용하였으며 각 변이에 대해 하나의 파라미터값을 적용하고 변환된 하나의 이미지를 모델에 입력하여 결과를 확인하였다. 파라미터는 최소값부터 최대값까지Scale, Shear, Contrast에서 0.1씩 Rotation, Brightness, Translation에서 1씩 증가시킴으로써 정확도를 측정하였다. 변화시킨 파라미터의 값의

Table 3. Model classification accuracy for train set and test set (%)

Data / Model	Train Set	Test Set
VGG16	98.27	47.29
VGG19	99.38	92.62
LeNet-1	99.46	68.17
LeNet-5	94.32	57.22
MobileNetV1	98.43	67.18
MobileNetV2	95.95	66.77

범위는 Table 3.에서 확인할 수 있다.

IV. 실험 및 분석 결과

실험 및 분석 결과는 3개의 연구 질문과 동일하게 구성하여 각 절에서 연구 질문에 대한 결과를 확인할 수 있다.

4.1 변이에 따른 모델 정확도

Table 4.는 각 변이마다 모든 파라미터 값을 적용하여 생성된 변이 이미지에 대한 모델의 정확도 평균값을 보여주고 괄호안의 수치는 표준편차를 말한다.

Fig. 1.의 Rotation 변이를 통해 볼 수 있듯이 이미지에 변화를 주면 여백이 발생하게 되고 여백은 검정색으로 채워지며 물체의 위치가 달라진다. 이 세 가지의 원인으로 Rotation 변이는 다른 변이들 보다 일반적으로 낮은 정확도를 보여주었다. 전반적으로 딥러닝 모델들은 Contrast, Translation, Brightness 변이에 대해 다른 변이 유형보다 높은 정확도를 보여주었다. 특히 이미지의 밝기를 조절하는 Brightness는 모든 모델의 정확도가 88% 이상을 보여줌으로써 모델의 정확도에 영향을 가장 적게 미침을 확인할 수 있어, 이미지를 구성하는 픽셀의 배치 구조를 변경함으로써 이미지 모양을 변경하는 작업을 뜻하는 기하학적 변화가 좀 더 모델의 분류 정확도를 낮추는 것을 확인하였다.

일반적으로 낮은 정확도를 보여주는 Rotation 변이에서 모델마다 결과가 상이함을 확인하였다. 이는 모델의 구조에 따라 입력 변이에 따른 정확도 차이가 있음을 말하고, 변이가 적용된 이미지에 대해 모델 신뢰성을 제공하기 위해서는 모델 구조 또한 중요한 요소인 것을 보여준다. 또한 MobileNet에서는 Shear 변이가 64% 이상이고 LeNet은 55% 이하였으며, VGG는 다른 모델들에 비해 Contrast 변이에서 74%이하로 정확도가 낮았다. 이를 통해 변이의 방식에 따라 모델의 취약점 차이가 발생한다는 연구 질문 1과 동일한 답을 확인하였다.

4.2 버전에 따른 모델 정확도

Table 4.에서 모델의 정확도를 통해 VGG16과 이후 버전 VGG19에서 변이 결과를 제외한 각 모델

Table 4. Average model classification accuracy (and standard deviation) according to each of 6 mutation types in which input samples are created from test set samples correctly classified by all models (%)

Model / Mutation	VGG16	VGG19	LeNet-1	LeNet-5	MobileNetV1	MobileNetV2
Original	100.00	100.00	100.00	100.00	100.00	100.00
Scale	63.60 (24.02)	98.78 (01.79)	68.86 (31.91)	70.26 (26.00)	74.43 (26.77)	73.71 (25.47)
Shear	51.23 (20.20)	87.04 (21.11)	55.19 (24.70)	53.36 (25.36)	66.42 (21.05)	64.51 (20.92)
Contrast	74.91 (32.57)	73.55 (32.57)	93.91 (12.54)	85.21 (14.79)	85.60 (19.64)	80.63 (17.12)
Rotation	38.30 (17.53)	71.90 (18.94)	50.40 (26.79)	46.29 (26.83)	56.38 (24.68)	58.80 (22.42)
Brightness	89.99 (08.80)	99.12 (00.94)	96.09 (03.02)	88.46 (08.08)	91.95 (06.52)	91.36 (05.93)
Translation	77.29 (12.13)	99.77 (00.16)	72.64 (20.92)	73.91 (20.53)	77.82 (15.38)	78.14 (15.17)

에서 버전에 따른 정확도의 큰 차이는 발견할 수 없었다. 다만, VGG 모델에서 Scale, Shear, Rotation, Brightness, Translation 변이와 LeNet 모델에서 Scale, Translation 변이와 MobileNet 모델에서 Rotation, Translaiton 변이를 제외한 모든 변이에서 이전 버전보다 이후 버전에서 변이된 이미지의 모델 정확도가 더 낮음을 발견하였다. 이후 버전은 기존 모델의 구조에서 새로운 기법 혹은 레이어 및 노드가 추가된다. 결과적으로 모델 구조의 작은 변화와 추가되는 노드 및 레이어는 변이된 입력에 대해 모델의 유의미한 신뢰성 향상을 제공하기 어렵다는 것을 시사한다. 또한 테스트 세트에서 가장 높은 정확도를 보여준 VGG19에서 다른 모델들 보다 변이에 따른 입력에 높은 정확도를 보여줌으로써 모델의 신뢰성을 제공할 것으로 예상하였다. 예상대로 VGG19는 Contarast, Rotation 변이를 제외한 Scale, Shear, Brightness, Translation 변이에 대해 87% 이상의 정확도를 보여주었다. 이는 VGG19가 높은 정확도를 보여주는 4개의 변이에 대해 나머지 5개의 모델보다 높은 신뢰성을 제공하는 것을 확인할 수 있었다. VGG19가 VGG16와 비교해 유의미한 신뢰성 향상을 불러일으킬 수 있었으나 LeNet과 MobileNet 모델에서는 해당 현상을 확인할 수 없었다. 이를 통해 연구 질문 2에 대한 모델 버전 간의 큰 차이는 확인하지 못하였다.

4.3 변이 파라미터에 따른 모델 정확도

Fig. 3.에서는 이미지를 변이의 개별적인 파라미터에 따라 구분하여 입력하였을 때 MobileNetV2

모델의 정확도의 변화를 보여준다. 각 그래프의 중간 값은 변이가 적용되지 않았을 때의 정확도를 말하며 좌우측으로 파라미터 값이 변화하면 이미지를 밝게 혹은 어둡게, 확대 혹은 축소 등으로 이미지가 변화한다.

전반적으로 파라미터를 증감시킬수록 모델이 더 취약해짐을 확인했다. Translation, Rotation, Shear, Brightness 변이에서는 정확도가 대칭 구조를 보여주었다. 이와 달리, Contrast와 Scale 변이에서는 낮은 파라미터의 값이 적용될수록 정확도 감소폭이 증가하는 것을 보여주었다.

Contrast 변이는 파라미터의 값을 높이면 이미지의 색상 경계들이 더욱 뚜렷해지고 파라미터 값을

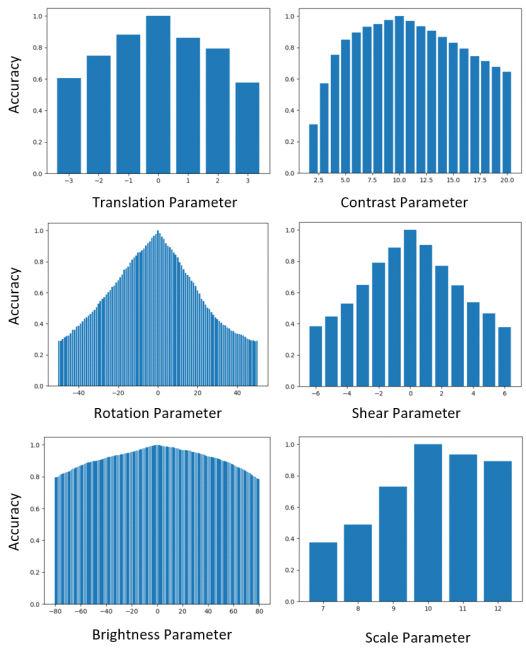

Fig. 3. Classification accuracy of MobileNetV2 according to 6 mutations and their all parameter values (%)

낮게 하면 이미지의 색상 경계가 부드러워지게 된다. 다시 말해, Contrast 변이의 파라미터가 증가하면 픽셀값들의 차이가 더욱 벌어지게 되고 이미지는 선명해지게 되는데, 모델은 이러한 이미지에 대해 낮은 파라미터를 값을 적용한 이미지보다 높은 정확도를 보여주었다. 또한 Contrast 변이의 낮은 파라미터 값에서 정확도 변화량이 급격하게 낮아지는 것을 확인할 수 있었다. 이는 모델이 이미지 색상 경계가 모호해지면 모델이 변이에 대해 더욱 취약하다는 것을 확인할 수 있었다.

Scale 변이는 낮은 파라미터 적용 시, 이미지가 축소되며 이미지 축소는 여백이 검정으로 채워지게 된다. 반대로 높은 파라미터 적용 시, 이미지는 확대되며 이미지 가장자리가 제거된다. 이에 따라 확대를 통한 이미지의 가장자리 변화는 모델에 큰 영향을 미치지 않고 이미지 축소를 통해 여백이 검정으로 채워지는 것에 좀 더 취약하다는 것을 확인하였다. 결과적으로 변이 파라미터에 따른 정확도 변화량이 차이가 있음을 확인하여 연구 질문 3과 동일한 답을 확인하였다.

V. 토의 및 시사점

DeepMutation[8]와 DeepTest[9]는 3개의 모델을 선정하여 검증을 수행하였다. 하지만 본 논문에서는 3개의 모델을 선정하였고 버전을 추가하여 총 6개의 모델에 대해 검증을 수행하였다. 이를 통해 더욱 다양한 모델의 결과를 분석하여 변이의 따른 모델들의 정확도 차이와 모델 버전 간의 정확도 차이를 확인할 수 있었다.

또한 DeepHunter[7]는 여러 변이를 겹쳐 이미지에 적용함으로써 각 변이의 모델의 정확도를 확인하기 어렵다. 이를 개선하기 위해 본 논문에서는 각 변이를 구분하여 적용하고 정확도를 확인함으로써 모델에 큰 영향을 미치는 Shear, Rotation 변이, 작은 영향을 미치는 Brightness 변이를 확인할 수 있었고 이에 더해 변이의 파라미터를 조정함으로써 파라미터의 따른 정확도의 감소량을 확인할 수 있었다. 결과적으로 모델의 정확도에 큰 영향을 미치는 변이를 파악함으로써 해당 변이와 관련된 모델 취약점 개선을 우선적으로 수행할 수 있다.

한편, 본 연구의 실험 설정에서 고려하지 못한 부분을 추가함으로써 검증을 확장할 수 있다. 먼저, 변이 파라미터에 따른 정확도 측정 시, 검증 모델은 MobileNetv2로만 수행하였기 때문에 다른 모델에서도 동일한 현상이 나타날지 검증이 필요하다. 파라미터의 값을 증감시킬 시 변화량의 차이는 있겠지만 동일한 구조의 결과가 나올 것으로 예상된다. 또한 본 연구에서 CIFAR10을 대상으로 데이터 세트를 구성한 한계점이 있으나 향후 연구에서는 1,000개의 클래스로 구성된 ImageNet을 사용하여 더 많은 물체에 대해 실험을 수행하고 실제 표지판 데이터 세트인 GTSRB(German Traffic sign Benchmark)를 사용하여 자율주행 분야에 사용되는 모델의 신뢰성을 제공할 것이다.

VI. 결론 및 향후 연구

본 논문에서는 변이에 따른 딥러닝 모델 정확도를 분석하는 기존 연구들의 한계점을 지적하고, 모델 및 변이를 추가하여 더 많은 실험을 통해 다양한 정확도를 확인하였다. 6개의 모델과 6개의 변이로 실험을 수행하여 총 36가지의 결과를 도출하였으며, 각 변이의 파라미터에 변동을 주어 파라미터에 따른 모델의 정확도 변화를 확인하였다. 결과적으로 기존의 이미지 변이 기반 딥러닝 모델 검증 방법보다 다양한 결과를 확인함으로써 다양한 경우의 모델 신뢰성을 제공할 수 있었고, 각 변이에 따른 모델 평균이 LeNet은 71.22%, VGG는 77.12%, MobileNet은 74.98%로 모든 모델의 평균 정확도가 80%를 넘지 못하여 입력 변이가 모델의 정확도에 많은 영향을 끼친다는 점을 파악할 수 있었다. 향후 연구로 본 연구에서 수립한 세 가지의 연구 질문에 기반한 실험 설계를 확장하고 변이가 모델에 미치는 영향에 대한 논리적인 분석을 진행하여 딥러닝 모델 공격의 방어 기법 연구를 진행할 것이다. 마지막으로 논문에서의 결과를 통해 딥러닝 모델 자체의 공격 기법에 대한 모델 검증 연구의 기반이 되어 더 많은 딥러닝 모델의 안전성 연구로 유도될 수 있기를 기대한다.

References

[1] He, Kaiming, et al., "Deep residual learning for image recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, Jun., 2016.

[2] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Proceedings of the Advances in Neural Information Processing Systems (NIPS), 25, pp. 1097-1105, Dec., 2012.

[3] Xiong, Wayne, et al. "Achieving human parity in conversational speech recognition." arXiv preprint arXiv: 1610.05256, 2016.

[4] Julian, Kyle D., et al. "Policy compression for aircraft collision avoidance systems." Proceedings of the IEEE/AIAA Digital Avionics Systems Conference (DASC), pp. 1-10, Aug., 2016.

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

[6] Yuan, Zhenlong, et al. "Droid-sec:

deep learning in android malware detection." Proceedings of the ACM Conference on SIGCOMM, pp. 371-372, Aug., 2014.

[7] Xie, Xiaofei, et al. "Deephunter: A coverage-guided fuzz testing framework for deep neural networks." Proceedings of the ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA), pp. 146-157, Jul., 2019.

[8] Ma, Lei, et al. "Deepmutation: Mutation testing of deep learning systems." Proceedings of the IEEE International Symposium on Software Reliability Engineering (ISSRE), pp. 100-111, Oct., 2018.

[9] Tian, Yuchi, et al. "Deeptest: Automated testing of deep-neural-network-driven autonomous cars." Proceedings of the International Conference on Software Engineering (ICSE), pp. 303-314, May, 2018.

[10] Who's responsible when an autonomous car crashes?, https://money.cnn.com/2016/07/07/technology/tesla-liability-risk/index.html, 2016.

[11] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083, 2017.

[12] Meng, Dongyu, and Hao Chen. "Magnet: a two-pronged defense against adversarial examples." Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), pp. 135-147, Oct., 2017.

〈저자소개〉

김 재 욱 (Jaeuk Kim) 학생회원
2018년 2월: 세명대학교 정보통신학부 졸업
2018년 6월~현재: 연세대학교 전보대학원 석사과정
〈관심분야〉 정보보호, 디지털 포렌식, 기계학습, Adversarial Machine Learning 등

박 래 현 (Leo Hyun Park) 학생회원
2017년 2월: 광운대학교 컴퓨터공학 졸업
2017년 3월~현재: 연세대학교 정보보호 연구실 통합 과정
〈관심분야〉 기계학습, 딥러닝 보안, Adversarial Machine Learning 등

권 태 경 (Taekyoung Kwon) 종신회원
1992년 2월: 연세대학교 컴퓨터과학과 학사
1995년 2월: 연세대학교 컴퓨터과학과 석사
1999년 8월: 연세대학교 컴퓨터과학과 박사
1999년~2000년: U.C. Berkely Post-Doc
2001년~2013년 8월: 세종대학교 컴퓨터공학과 교수
2007년~2008년: Univ. Maryland at College Park 교환교수
2013년 9월~현재: 연세대학교 정보대학원 교수
〈관심분야〉 암호프로토콜, Usable Security, 소프트웨어/시스템보안, 기계학습과보안 등