creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

**저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.**

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학석사 학위논문

# 이중 경로 신경망과 적대적 생성 신경망을 이용한 문자 이미지 데이터 증강

2022년 2월

서울대학교 대학원

컴퓨터공학부

하 경 모

# 이중 경로 신경망과 적대적 생성 신경망을 이용한 문자 이미지 데이터 증강

지도교수 문 병 로

이 논문을 공학석사 학위논문으로 제출함
2021년 12월

서울대학교 대학원
컴퓨터공학부
하 경 모

하경모의 석사 학위논문을 인준함
2022년 1월

위 원 장 _____ 전화숙 _____ (인)

부위원장 _____ 문병로 _____ (인)

위 원 _____ 권태경 _____ (인)

# 초록

대부분의 딥러닝 모델들을 학습시키는 데에는 다양한 품질의 많은 데이터가 필요하며, 광학 문자 인식을 위해 필요한 모델도 그중 하나다. 하지만 이에 필요한 학습 데이터들을 구하는 것에는 너무나도 큰 비용이 필요하기에 이것을 극복하기 위해 데이터 증강이 시도된다. 안타깝게도, 광학 문자 인식의 특성상 일부 제한된 증강 기법들만 사용 할 수 있으며 이마저도 다양한 품질의 데이터를 얻는 데 한계를 지니고 있다. 본 논문에서는 이러한 한계점을 극복하고자 이중 경로 신경망을 이용한 적대적 생성 신경망을 구축하여 데이터를 증강하는 방안을 제안하였다. 또한, 실제로 제안한 방식이 인식 성능에 어느 정도 영향을 미치는지 기존 방법들과 비교 분석하였다. 연구 결과, 적대적 생성 신경망이 기존의 증강과는 다른 방식으로 다양하게 이미지들을 변환하는 것을 관찰할 수 있었고, 이에 따라 문자 인식 모델의 성능도 개선되는 것을 알 수 있었다. 또한, 기존의 데이터 증강과 함께 적용한다면 큰 폭으로 성능이 향상되는 결과를 보여주어 적대적 생성 신경망을 통한 데이터 증강이 광학 문자 인식에 있어 효용성이 있다는 것을 확인하였다.

**주요어 : 문자 이미지 데이터 증강, 적대적 생성 신경망, 이중 경로 신경망, 광학 문자 인식, 컴퓨터 비전**
**학 번 : 2020-27787**

# 목 차

초록 ····· i
목차 ····· ii
표 목차 ····· iii
그림 목차 ····· iv
제 1 장 서론 ····· 1
제 2 장 이론적 배경 ····· 3
제 1 절 적대적 생성 신경망 ····· 3
제 2 절 이중 경로 신경망 ····· 4
제 3 절 관련 연구 ····· 6
제 3 장 연구 방법 ····· 8
제 1 절 모델 ····· 8
제 2 절 데이터 ····· 9
제 3 절 실험 ····· 12
제 4 장 연구 결과 및 분석 ····· 14
제 5 장 논의 ····· 18
제 6 장 결론 ····· 21
참고문헌 ····· 22
부록 ····· 26
Abstract ····· 27

# 표 목 차

[표 3-1] 인식 모델 성능 비교를 위한 데이터 종류 ····· 12
[표 4-1] 각 데이터에 따른 인식 모델의 정확도 ····· 15

# 그 림 목 차

[그림 2-1] 적대적 생성 신경망 ····· 3
[그림 2-2] 이중 경로 신경망 ····· 5
[그림 3-1] 생성자의 세부 구조 ····· 8
[그림 3-2] 판별자의 세부 구조 ····· 9
[그림 3-3] FUNSD 단어 이미지 ····· 10
[그림 3-4] 자체 제작 단어 이미지 ····· 11
[그림 4-1] 학습된 GAN으로 증강시킨 문자 이미지 ····· 14
[그림 4-2] GAN 증강을 통해 정답을 맞힌 이미지 모음 ·· 16
[그림 5-1] 학습에 실패한 두 생성자가 만든 이미지 모음 19
[그림 5-2] 한글에 데이터 증강을 시도한 결과 ····· 20

# 제 1 장 서론

딥러닝 모델의 성능을 향상하는 데 있어 다양한 품질의 많은 데이터가 필요하다는 것은 널리 알려진 정설이다. 특히나 지도 학습을 기반으로 모델을 학습할 때 사용해야 할 데이터에 반드시 정답지가 포함되어야 한다. 이러한 제약 조건 때문에 지도 학습에 활용할 수 있는 데이터는 전체 데이터에 비해 일부분이고, 원하는 모델의 학습 목표에 부합되는 충분한 데이터를 구하기에 어려움이 따른다. 만약 적절한 공개 데이터를 구하지 못해 정답지까지 직접 제작해야 한다면 값비싼 비용(노동력, 재화, 시간)이 들뿐더러, 때에 따라서는 현실적으로 불가능할 수도 있다.

광학 문자 인식 (Optical Character Recognition, 이하 OCR) 분야도 위 문제점에서 벗어날 수 없다. OCR은 종이 위에 사람이 쓰거나, 기계가 인쇄한 문자들을 스캔, 사진을 찍는 등의 방법으로 이미지를 얻어 이를 기계가 읽을 수 있게 변환하는 것인데, 크게 문자의 위치를 특정하고, 해당 문자를 인식하는 두 부분으로 나뉜다[1]. 이 중 문자를 인식하는 모델을 학습하기 위해서는 문서 이미지뿐만 아니라 이미지 내 문자들의 위치와 각 문자의 정답이 포함된 데이터가 필요하다. 앞서 언급했듯, 이러한 데이터를 대량으로 제작하는 것은 현실적으로 불가능하기에 그 양이 많지 않고, 그에 따라 주어진 데이터만으로는 문자 인식 모델을 학습시키는 데 어려움이 따를 수밖에 없다.

이것을 극복하기 위해 시도할 수 있는 방안 중 하나가 데이터 증강(augmentation)이다. 부족한 데이터양을 보충함과 동시에 다양한 품질의 데이터도 획득할 수 있어 학습 데이터를 마련하는 데 있어 필수적인 관

문이다. 이러한 데이터 증강은 OCR뿐만 아니라 다른 컴퓨터 비전 분야에서도 많이 쓰이기에 여러 가지 기법들이 존재한다[2, 3]. 하지만 주로 작은 흑백 문자를 인식해야 하는 문자 인식의 특성상 색상 변환이나, 대칭 이동, 혼합 등의 여러 증강 기법들은 적용할 수 없어 제한적인 회전, 수평 이동, 노이즈 추가 등의 일부 증강 기법들만 활용할 수 있다. 이런 제약 조건 때문에 문자 인식에 활용할 수 있는 다른 증강 기법에 관한 연구도 이루어졌지만[4], 실제로 문자를 인식하는 데 걸림돌이 되는 부분이 인쇄 시의 잉크 번짐이나 스캔 시의 화질 저하 등이라는 것을 고려한다면, 이러한 증강 기법들은 이미지 내 좌표의 상대 거리만 변형한다는 한계를 벗어나지 못하였다.

본 논문에서는 이러한 한계점을 적대적 생성 신경망 (Generative Adversarial Nets, 이하 GAN)[5]을 통해 극복하였다. 실제 영어 문서에서 일어난 손상을 GAN을 이용해 학습, 재현했으며, 이러한 방식으로 증강된 데이터가 기존의 방법들에 비해 인식 모델의 성능을 향상시킬 수 있다는 것을 확인하였다.

# 제 2 장 이론적 배경

## 제 1 절 적대적 생성 신경망

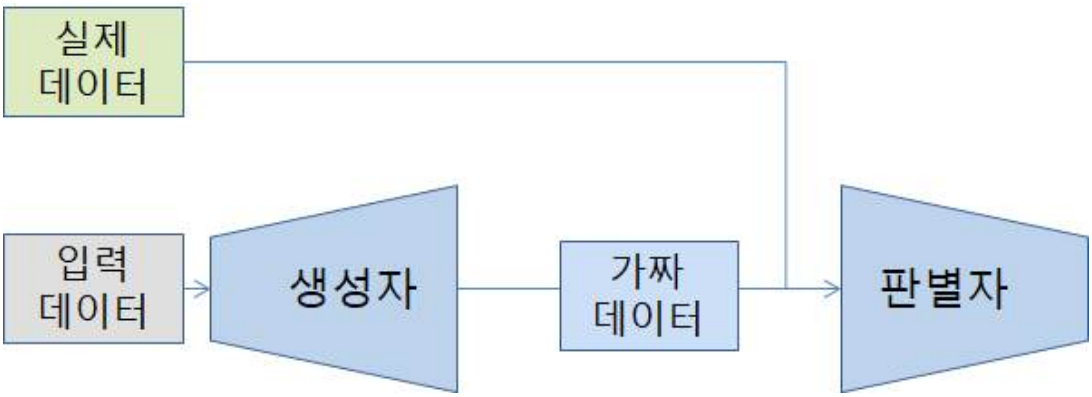


그림 2-1 적대적 생성 신경망

일반적으로 한가지의 목표를 가지고 하나의 모델을 학습시키는 것과 달리 GAN[5]은 그림 2-1과 같이 두 가지 모델, 생성자(Generator)와 판별자(Discriminator)를 동시에 학습시킨다. 두 모델은 각자 다른 목표를 가지고 학습하는데, 생성자는 입력 데이터를 바탕으로 판별자를 속일 데이터를 생성하는 것을 목표로, 판별자는 실제 데이터와 생성자가 만들어낸 가짜 데이터를 판별하는 것을 목표로 학습한다.

좀 더 자세히 보면, z를 입력 데이터, G를 생성자, D를 판별자라고 했을 때, 생성자의 경우 수식 2.1을 최소화하는 것을 목표로, 판별자의 경우 수식 2.2를 최대화하는 것을 목표로 삼기에 최종적인 GAN의 목표 함수는 수식 2.1, 2.2를 합친 형태인 수식 2.3과 같다.

이러한 구조로 인해 GAN은 종전의 생성 모델들[6, 7]에 비해 MNIST[8] 같은 데이터 세트에서 더 좋은 성과를 보였지만, 그보다 복잡한 결과물을 만들어내는 것에는 아직 미흡했고, 별개의 두 모델을 같이 학습시키다 보니 학습의 난이도가 높은 등, 여러 단점도 존재하였다. 이런 몇몇 단점들을 극복하기 위해 손실함수를 개선하거나[9, 10], 모델의 세부 구조를 복잡하게 변경하거나[11], 데이터와 모델에 조건을 추가해 학습하는 [12, 13, 14] 식으로 후속 연구가 이어졌다.

$$E_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (2.1)$$

$$E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (2.2)$$

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (2.3)$$

본 논문에서는 DCGAN[11]의 구조를 바탕으로 연구 주제와 맞게 변형, CycleGAN[14]의 일부를 차용해 생성 모델을 설계했으며, 이에 대한 자세한 사항은 3장에서 언급하도록 한다.

## 제 2 절 이중 경로 신경망

비전 분야에서 Deep Convolutional Neural Networks(DCNN)]가[15] 등장한 이후 모델들은 더욱 복잡한 임무를 잘 수행해내기 위해 점차 깊어

졌다. 문제는 과도하게 깊어진 모델들은 그렇지 않은 것들에 비해 오히려 성능이 떨어지는 모습을 보여주었는데, 이 문제점을 개선한 대표적인 두 모델이 Residual Network(ResNet)와[16] Densely Connected Convolutional Networks(DenseNet)다[17]. 전자인 ResNet은 앞서 언급한 degradation 문제를 특유의 'shortcut connection'을 통해 개선했으며[16], 후자인 DenseNet 또한 이러한 문제를 앞선 층을 후속 층에 모두 연결하는 'dense connectivity pattern'을 이용해 극복하려고 시도하였다[17].

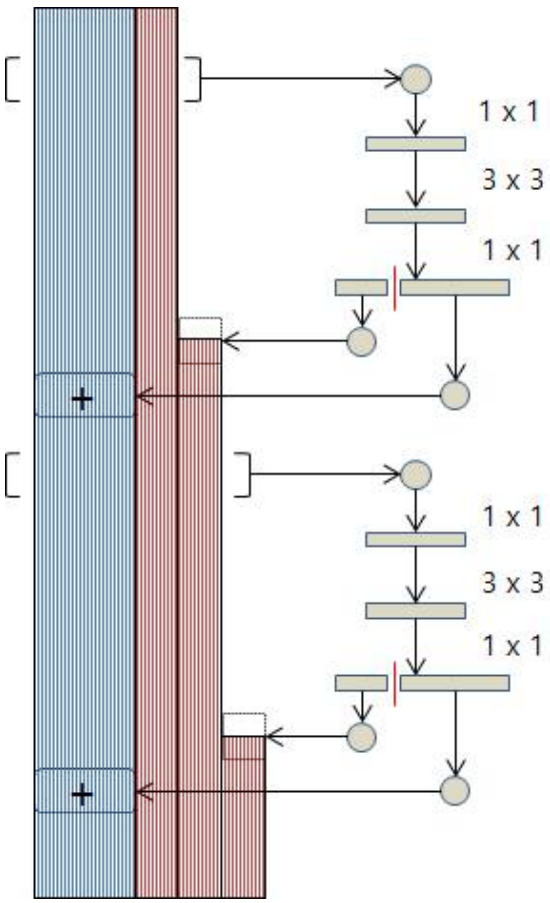


그림 2-2 이중 경로 신경망

이중 경로 신경망 (Dual Path Networks, 이하 DPN)은[18] 이 두 모델

각각의 장점을 취하려고 시도한 모델이다. 그림 2-2에서 볼 수 있듯이 DPN은 ResNet의 ‘shortcut connection′(파란색)을 유지해 나가며, DenseNet의 ′dense connectivity pattern′(빨간색) 또한 구성하는 방법을 택해 종전의 모델들보다 더 좋은 성능을 보였다[18].

본 논문에서는 DPN의 구조를 바탕으로 생성 모델의 생성자를 설계했으며, 세부 사항은 3장에서 이어 나가도록 한다.

## 제 3 절 관련 연구

Luo et al.[4] 또한 문자의 다양한 모양과 왜곡 패턴이 이를 인식하는데 있어 문제점이라고 제기하였다. 이를 개선하기 위해 저자들은 본인들이 명명한 ‘joint learning′이라는 방법론으로 데이터 증강 모델과 문자 인식 모델을 동시에 학습시켰다. 해당 방식은 학습 주기마다 이미지 내 특정 위치의 점들을 기준으로 이미지를 뒤틀면서 증강하는데, 이때 CNN과 FCN으로 이루어진 ′learnable agent′와 무작위, 두 가지 방법으로 뒤트는 정도를 결정한다. 이렇게 두 방법으로 증강이 된 두 데이터는 미리 준비된 문자 인식 모델에 입력되며, 인식 모델은 새로이 들어온 이 데이터를 학습한다. 학습이 끝나면 인식 모델은 새로 적용된 증강 데이터가 전 학습 주기보다 인식이 어려워졌는지 피드백을 주고, 이 피드백을 ’learnable agent′가 배우면서 학습이 진행된다[4]. 결과적으로 어느 정도 학습이 수렴된 ′learnable agent′를 통해 저자들은 새로운 데이터 증강 방식을 제안했고, 더불어 문자 인식 모델의 성능을 개선하는 데 성공하였다. 증강 모델 학습과 동시에 인식 모델의 성능도 개선된다는 점이 인상 깊었지만, 미리 인식 모델을 준비해야 한다는 점, 제시한 증강 기법이

문자의 형태만 왜곡시킬 뿐 그 이상의 다양한 증강을 가하지는 못했다는 한계점이 있었다.

본 논문에서는 생성 모델을 통해 이러한 한계점을 개선하고, 문자의 왜곡뿐만 아니라 실제 데이터에서 일어날 법한 손상을 재현해내 더 다양한 품질의 데이터를 생성해내는 것에 목표를 둔다.

# 제 3 장 연구 방법

## 제 1 절 모델

본 논문에서 쓰일 생성 모델은 앞서 언급했듯 DCGAN의[11] 형태를 띠고 있기에 하나의 생성자, 하나의 판별자로 동일하지만, 그 외 나머지 부분에서는 연구 주제에 맞게 변경하였다.

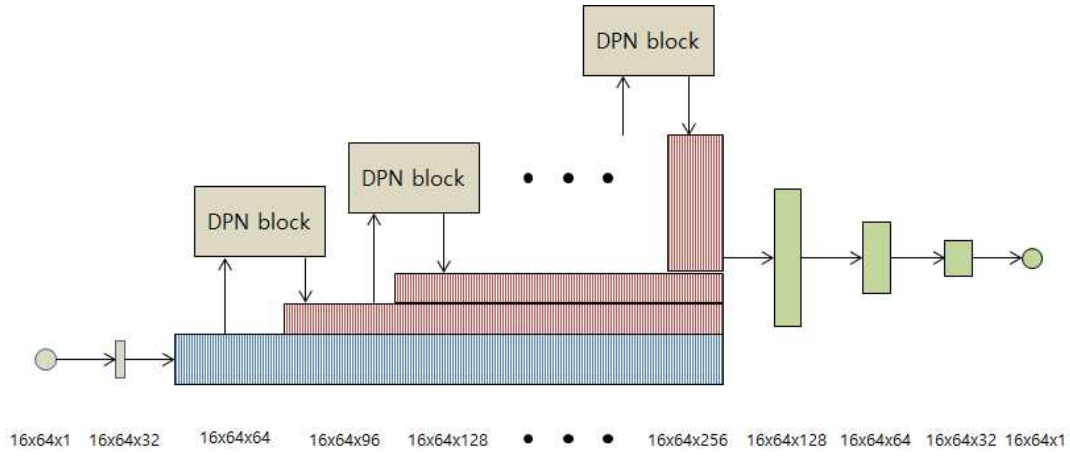


그림 3-1 생성자의 세부 구조

기존 DCGAN의 생성자 같은 경우 일련의 노이즈를 입력 데이터로 받아 이미지를 생성해내지만, 이번 연구의 방향성은 이미지를 생성해내는 것이 아니라 증강, 혹은 변형시키는 것이기에 증강 전의 영어단어 이미지들을 입력 데이터로 제공한다. 입력 데이터가 하나의 벡터에서 출력물과 같은 크기의 이미지로 대체된 만큼, 모델의 뒤로 갈수록 흐르는 데이터의 크기가 커져야 하는 기존 모델과 달리 본 모델의 생성자는 같은 크

기를 유지하고 채널 수만 변경하며 이미지를 생성한다. 이러한 형태를 띠기 위해 생성자의 모델을 적절히 변형된 DPN으로[18] 대체하였다(그림 3-1).

판별자의 경우 생성자가 만들어낸 데이터와(가짜) 실제 데이터(진짜)를 가려내는 역할로, 생성자의 수행 목표 대비 간단한 만큼 GAN 학습 시 모델 간의 균형을 고려해 생성자보다 간단한 구조를 띠고 있다. 세부 구조는 그림 3-2와 같이 생겼으며, 일반적인 DCNN[15] 구조와 유사하게 CNN, pooling, FCN으로 이루어져 있다.

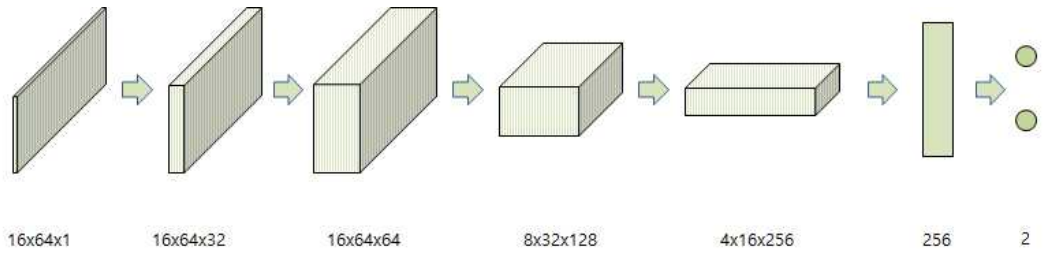


그림 3-2 판별자의 세부 구조

생성자와 판별자에 활용한 DPN과 DCNN 모두 기본적으로 활성화 함수에 ReLU를[19] 활용하지만, 본 모델에는 비전 분야에서 활용하기에 조금 더 성능이 좋은 Swish를[20] 적용하였다.

## 제 2 절 데이터

가장 먼저 필요한 데이터는 GAN이 학습할 데이터, 즉 손상이 가해진 영어 문서 이미지들이다. Form Understanding in Noisy Scanned Documents (이하 FUNSD)는[21] OCR과 문서 구조 파악 학습을 위한

데이터 세트로, 이름에서 알 수 있다시피 상태가 좋지 않은 실제 문서들로 구성되어 있어 본 연구에 적합한 데이터 세트다.

해당 데이터 세트는 이미 학습, 시험용(train, test)으로 데이터가 구분되어 있기에 총 199개의 문서 중 학습용으로 구비된 149개의 문서에서 단어 이미지를 잘라내 GAN의 학습 데이터로 만들었다. 일반적인 GAN은 기존 데이터에 존재하지 않는 무언가를 생성하기 때문에 그 품질을 정량적으로 평가하기 어려워 평가 데이터가(validation set) 존재하지 않지만, 모드 붕괴(mode collapse)를 방지하기 위해 본 연구에서는 예외적으로 평가 데이터를 활용하였다. 이 때문에 학습 데이터의 10% 정도를 평가 데이터에 할당, 나머지를 학습 데이터에 할당하였다. 나머지 시험용으로 구비된 50개의 문서는 생성이나 인식 모델의 학습에는 전혀 관여하지 않고 마지막으로 학습이 완료된 인식 모델의 성능을 측정하는 테스트 데이터로 사용하였다. 즉, 정리하자면 FUNSD의 문서에서 추출한 단어 이미지 중 GAN의 학습 데이터로 12722장, 평가 데이터로 1270장, 인식 모델의 성능 파악을 위한 테스트 데이터로 5241장으로 구성하였다. 그림 3-3이 이 중 일부를 나열한 것인데, 전반적으로 화질이 떨어지며 군데군데 획이 손상된 것을 알 수 있다.

그림 3-3 FUNSD 단어 이미지

그다음 필요한 데이터는 GAN이 변형시킬 재료, 즉 손상되지 않은 영어단어 이미지들이다. python의 PIL 라이브러리를[1] 이용하면 구비한 폰트에서 바로 문자 이미지들을 생성해낼 수 있으므로, 상업적 사용이 가능한 폰트 10개(부록 A) 중 단어마다 무작위로 골라 준비된 FUNSD의 학습 데이터와 쌍을 이루도록 만든다. 그림 3-4가 만들어낸 데이터이며, 그림 3-3의 FUNSD 데이터보다 확연히 깔끔한 것을 알 수 있다.

그림 3-4 자체 제작
단어 이미지

마지막으로 필요한 데이터는 인식 모델을 학습시킬 영어단어 이미지 데이터다. Google ngram viewer[2]는 본인들이 보유하고 있는 1조개의 말뭉치 데이터(trillion word corpus)를 분석하고 제공해주는데, 본 논문에서는 해당 데이터를 빈도수로 나열, 상위 6만개의 단어를 추출해 데이터로 가공하였다. 이때, 영어의 대문자, 소문자를 구분해서 학습시키기 위해 단어를 다음과 같이 3 분류로 나누어 가공하였다. 1. 대문자로만 이루어진 영어단어, 2. 앞글자 하나만 대문자인 영어단어, 3. 소문자로만 이루어진 영어단어. 물론 인식 모델을 학습시키기 위해서는 평가 데이터가 필요하기에 모든 단어 중 무작위로 선별한 7169개의 단어는 평가 데이터

1) https://pillow.readthedocs.io/en/stable/
2) https://books.google.com/ngrams/info

로 할당, 나머지 단어는 학습 데이터로 할당하였다.

이렇게 만들어진 인식 모델 학습 데이터는 기존 데이터 증강 기법과 본 논문에서 제시한 방식과의 차이점을 파악하기 위해 다시 5종류의 데이터로 재가공 되었으며, 자세한 구성은 표 3-1과 같다. 이때, python의 imgaug 라이브러리를[3] 이용해 기존 증강 기법을 적용했으며, 그중 4개, 회전, 평행 이동, 확대 축소, 기울이기를 선택해 활용하였다. 즉, 데이터 B 같은 경우 원본 데이터 A에 기존 증강 기법 중 이미지마다 무작위로 고른 하나만을 적용했고, 데이터 C는 이미지마다 기존 증강 기법 중 하나에서 네 기법까지 무작위로 선택해 구성했다. 데이터 D 같은 경우 본 논문에서 제안한 GAN을 통한 증강 기법을 적용했으며, 데이터 E 같은 경우 데이터 C와 D에 가한 증강을 모두 사용해 구성한 데이터다.

| 데이터 명칭 | 데이터 구성 | 데이터 내 이미지 개수 |
|---|---|---|
| A | 제작된 인식 모델 학습 데이터 | 60,000 |
| B | A + (A * 기존 증강 중 택 1) | 120,000 |
| C | A + (A * 기존 증강 중 택 1~4) | 120,000 |
| D | A + (A * GAN 증강) | 120,000 |
| E | A + (A * GAN 증강 * 기존 증강 중 택 1~4) | 120,000 |

표 3-1 인식 모델 성능 비교를 위한 데이터 종류

## 제 3 절 실험

모델 간의 균형을 위해 구조의 복잡도 외에도 학습 시의 학습률도 고려해야 하기에 판별자는 생성자에 비해 절반의 학습률로 학습하게 지정해 주었다. 이렇게 두 모델 간의 균형을 고려해 GAN을 설계했다 하더라도

3) https://imgaug.readthedocs.io/en/latest/

여전히 GAN의 학습 난이도는 낮지 않은 편이다. 특히 OCR의 대상이 되는 작은 흑백 글자는 의도한 손상과 달리 다른 글자로 변형되는 등 문제가 되는 형태로 증강이 적용되기 쉽다. 따라서 생성자의 학습 시 CycleGAN의[14] 학습에 활용되는 identical loss를 추가로 적용해 모드 붕괴를 방지하였다. 또한, 일반적으로 GAN에서 생성자를 선택할 시 일관적으로 학습 loss가 가장 작은 것을 고르는 것과 달리, 두 구간으로 나누어 학습을 진행하였다. 첫 구간에서 학습할 때는 본 장의 2절에서 언급했던 평가 데이터를 따로 두어 해당 데이터의 identical loss가 가장 작은 모델을 골라냈다. 이렇게 선택된 모델로 두 번째 구간에서는 일반적인 방식과 같이 학습 loss가 작은 것을 고르는 대신, 첫 구간에 비해 학습률을 3분의 1로 낮추어 학습시켰다.

학습이 완료된 생성 모델의 결과를 비교하기 위해서는 문자 인식 모델이 필요해 문자 인식 모델 중 하나인 TRBA를[22] 2절에서 언급된 데이터로 학습시켰다. TRBA는 모델을 구성하는 각 부분의 앞 글자를 따온 이름인데, 순서대로 TPS, ResNet, BiLSTM, Attention으로 구성되어 있다[22]. 다만, TRBA의 본 목적이 Scene Text Recognition(STR)이었던 점을 고려해 모델 가장 앞부분에서 이미지의 변형(transformation)을 담당하는 TPS를 제외하고 사용하였다. 이 모델을 인식 모델로 5종류의 데이터를 사용해 학습시켰으며, 해당 결과는 4장에서 언급하도록 한다.

# 제 4 장 연구 결과 및 분석

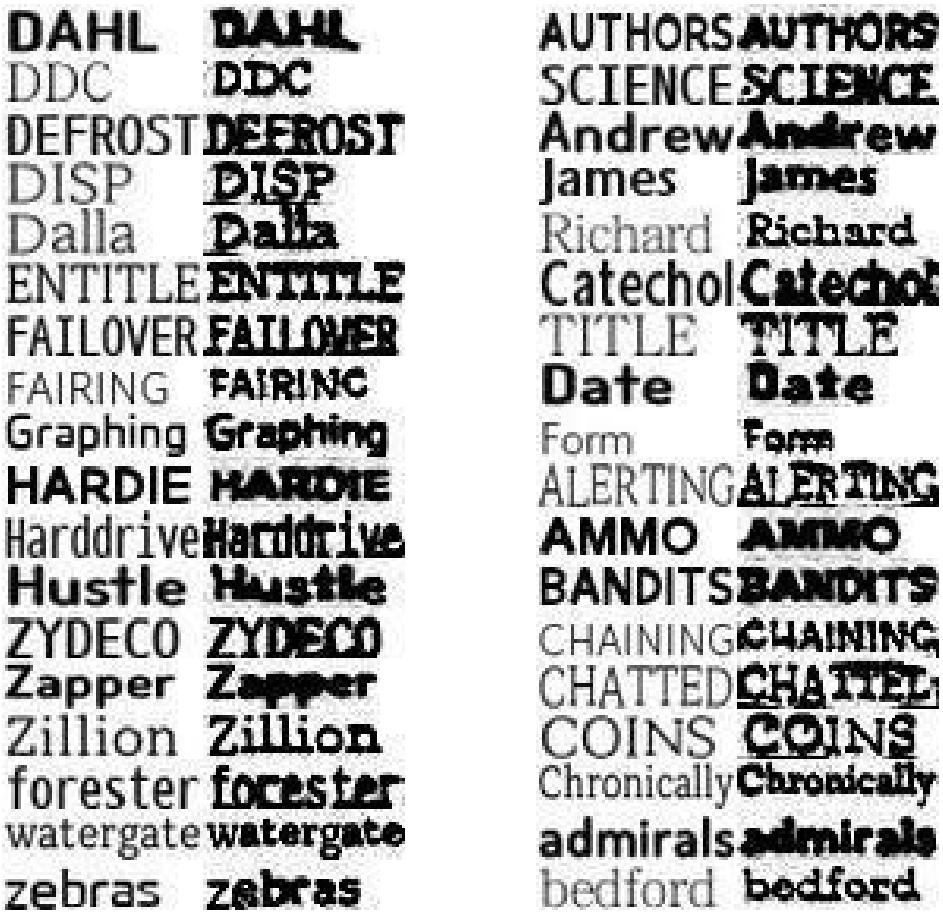


그림 4-1 학습된 GAN으로 증강시킨 문자 이미지

그림 4-1은 학습 완료된 GAN으로 데이터 증강을 적용한 문자 이미지들의 일부다. 각 그림의 왼쪽 열에 나열된 이미지는 3장에서 언급했던 데이터 A의 일부이며 GAN의 학습에는 일절 관여하지 않고 인식 모델의 학습에만 쓰이는 데이터다. 각 그림의 오른쪽 열에 나열된 이미지는 같은 행의 왼쪽 열의 데이터 A를 증강한 결과다.

증강 결과 전체적으로 화질이 조금 떨어진 듯한 모습을 보이며, 실제

문서에서 잉크가 번져서 생긴 손상을 재현하려는 것 같이 획이 굵어지거나 거칠어지는 현상을 관찰할 수 있다. 특히 그림 4-1의 DAHL, Zapper, Andrew 같은 단어에서 획 중간의 빈 곳이 메워지는 것으로 이런 현상을 좀 더 확연히 볼 수 있다.

글씨체에서도 약간의 변화가 생긴 대상들이 있다. 그림 4-1의 Hustle, James, Richard 같은 경우 왼쪽의 원본 글씨체와 오른쪽의 증강 후 글씨체가 조금 다른 형태를 띠는 것을 관찰할 수 있다.

그림 3-3에서 볼 수 있듯이 FUNSD 데이터에는 표 혹은 밑줄 때문에 단어의 밑부분에 획이 그어진 경우가 있는데, GAN에서 그러한 손상을 어느 정도 재현하려는 것을 확인할 수 있다. 그림 4-1의 DISP, FAILOVER, forester, ALTERING, CHATTED, COINS가 이러한 경향성을 보여주었다.

여기에 더해, 본 논문에서 제시한 GAN의 데이터 증강이 실제로 문자 인식에 긍정적인 효과를 내는지 검증해보기 위해 3장에서 언급한 5가지 종류의 데이터로 실험해보았다. 모든 인식 모델은 학습 데이터를 제외하고 모두 같은 조건에서 학습되었으며, FUNSD 테스트 데이터 5241자를 대상으로 한 인식 성능은 아래 표 4-1과 같다. 이때, 데이터마다 모델 3개를 학습시켜 중앙값에 위치하는 모델의 정확도를 표기하였다.

| 학습 데이터 명칭 | 정확도 |
|---|---|
| A | 28.77 % |
| B | 36.73 % |
| C | 40.45 % |
| **D** | **45.88 %** |
| **E** | **57.83 %** |

표 4-1 각 데이터에 따른 인식 모델의 정확도

당연하게도, 아무런 증강을 적용하지 않은 A를 학습한 모델이 가장 정확도가 낮은 모습을 보여주었다. 다만, A의 데이터 수가 다른 데이터들에 비해 절반밖에 되지 않기에 정확도의 큰 차이는 이를 참작해서 생각할 필요가 있다.

기존 증강 기법을 하나만 적용한 B와 모두를 적용한 C를 비교했을 때는 C가 조금 우위를 차지하였다. 과도한 증강이 오히려 성능을 떨어뜨리는 경우도 존재하지만, 실험에 사용한 FUNSD 데이터의 상태를 고려해 보았을 때 증강을 더 가한 데이터가 더 좋은 성능을 이끌어내는 것은 타당하다고 보인다.

본 논문에서 제시한 방법을 적용한 데이터 D가 기존 방식의 증강만 가했던 B, C 모두에 비해 좋은 성능을 가졌으며, GAN을 통한 데이터 증강 시도가 효용이 있음을 알 수 있다. 주목할 만한 것은 그전까지의 증강이 모두 적용된 데이터 E인데, 다른 데이터 간의 성능 차이보다 상대적으로 훨씬 더 개선되었기 때문이다. 이를 통해 본 논문에서 제시한 방안이 기존의 방안보다 뛰어난 것뿐만 아니라 같이 적용하면 상승효과를 일으켜 성능 개선에 더 도움이 되는 것을 확인하였다.

그림 4-2 GAN 증강을 통해
정답을 맞힌 이미지 모음

수치상으로 확인하는 것 말고도 GAN 증강이 어떠한 이미지를 맞추는데 도움이 되었는지 알아보기 위해 데이터 C를 학습한 모델에서는 맞추지 못했지만, 데이터 E를 학습한 모델에서는 맞춘 단어들 일부를 나열하였다(그림 4-2). GAN 증강을 통해 일부 이미지에서 밑줄이 그어지는 듯한 변환이 일어난 것 때문인지 그전까지 맞추지 못했던 밑줄이 그어진 단어, Duggan, Merchandising, receiving을 맞춘 것을 볼 수 있다. 또한, 글자가 뭉개져 덩어리진 듯한 느낌을 주는 may, facsimile, Review 등을 더 잘 맞추었으며, Acknowledgement 같이 단어의 글자 수가 많아 촘촘하게 생긴 이미지에서도 인식 성능에 어느 정도 개선이 있는 것을 확인할 수 있다.

# 제 5 장 논의

앞선 결과에서 알 수 있듯이 DCGAN의[11] 구조를 바탕으로 한 생성 모델의 데이터 증강으로도 인식 모델의 성능을 개선할 수 있었다. 하지만 여기서 그치지 않고 한 단계 더 개선할 방안이 있는데, CycleGAN[14]의 구조를 차용하는 것이다. CycleGAN의 경우 DCGAN과 달리 서로의 역으로 수행하는 2개의 생성자, 판별자로 이루어져 있다. 예로 본 논문의 주제에 적용해보면, 손상을 가하는 생성자, 이러한 손상된 이미지의 참, 거짓을 가리는 판별자 한 쌍과, 손상을 되돌리는 생성자, 이런 손상되지 않은 이미지의 참, 거짓을 가리는 판별자 한 쌍, 이렇게 네 모델로 구성할 수 있다. 학습이 의도대로 잘 된다면 이러한 생성 모델을 통해 본 논문에서 사용한 손상을 가하는 생성자와 더불어 손상을 되돌리는 생성자까지 얻을 수 있다. 이 뜻은 본 연구에서 적용한 데이터 증강 외에 테스트할 이미지의 저화질이나, 손상 등을 개선함으로써 인식 모델의 성능을 한층 더 끌어올릴 가능성이 있다는 것이다. 하지만 안타깝게도 이 방안은 학습 시간의 증가 외에도 OCR 데이터 증강의 특수성 때문에 마냥 긍정적으로 바라볼 수만은 없다.

여느 GAN이 그러하듯 CycleGAN 또한 내부 모델 4개가 모두 균형을 이루며 학습해야 한다. 여기서 균형을 이루어야 하는 모델의 수가 늘어난 것도 문제지만, 더 큰 문제는 모델 간의 수행 난이도 편차가 더 늘어났다는 것이다. 문제의 시작은 가장 낮은 난이도의 수행이 요구되는 손상되지 않은 이미지를 가리는 판별자로부터 시작된다. 해당 판별자는 여타 세 모델에 비해 학습 난이도가 낮은 만큼 학습 시 빠른 속도로 수렴한다. 아이러니하게도 한 쌍으로 엮인, 네 개의 모델 중 가장 높은 난이

도의 수행을 요구받는 손상을 되돌리는 생성자는 앞선 판별자의 빠른 학습 속도를 따라가지 못해 결국 판별자로부터 제대로 된 학습 피드백을 받지 못하게 된다. 이렇게 문제가 생긴 생성자는 반대편의 생성자에게도 영향을 미치게 되고 종국에는 생성 모델 전체의 학습 실패로 연결된다.

그림 5-1 학습에 실패한 두 생성자가 만든 이미지 모음

그림 5-1은 이렇게 학습에 실패한 두 생성자가 생성해낸 이미지들 일부를 나열한 것이다. 왼쪽 그림의 오른쪽 열은 왼쪽 열의 손상되지 않은 이미지를 손상을 가하는 생성자가 생성해낸 결과고, 오른쪽 그림의 오른쪽 열은 왼쪽 열의 손상된 이미지를 손상을 되돌리는 생성자가 생성해낸 결과다. 볼 수 있다시피 모두 적절치 못한 학습의 결과로 일종의 모드 붕괴가 발생해 문자 자체를 바꾸어버렸으며, 생성해낸 이미지의 품질 자체도 떨어지는 것을 볼 수 있다.

그럼에도 여전히 두 단계의 과정을 통해 인식 성능을 개선할 수 있다는 것은 매력적이다. 비록 본 연구에서는 이를 통한 개선이 성공적이지 못했지만, 추후 연구를 통해 더 발전할 가능성이 있다고 본다.

다른 방식으로 학습을 시도하는 것 외에 타 언어와의 호환성도 고려해 볼 만한 주제다. 많은 분야가 그러하듯 OCR 또한 대부분 데이터가 영어로 구성되어 있고, 한글 OCR 데이터의 경우 대표적으로 AI hub[4]와 Ko. AI OCR[5] 에서 데이터를 제공하고 있지만, 영어에 비해 데이터의 품질,

4) https://aihub.or.kr/

양은 턱없이 부족한 것이 현실이다. 이렇게 한글 데이터가 부족한 상황에 기인해 학습이 수월한, 데이터를 얻기 비교적 쉬운 언어인 영어로 학습한 데이터 증강이 친척 언어가 아닌 문자에도 효력을 보이는지 실험하였다.

그림 5-2 한글에 데이터
증강을 시도한 결과

그림 5-2는 본 연구에 사용한 데이터 증강 생성 모델을 한글에 적용해 보았을 때의 결과 이미지다. 안타깝게도 비슷한 생김새의 문자가 아니어서 그런 것인지 그림 4-1의 알파벳을 대상으로 한 증강과는 달리 다양한 변화를 관찰할 수는 없었다. 또한, 현시점에서 이렇게 변환된 데이터가 얼마나 더 성능을 개선할지 시험할 수 있는 마땅한 한글 데이터가 없는 관계로 이에 관한 연구는 여건이 좀 더 개선된 후속 연구들에서 진행되기를 바라는 바이다.

5) http://101.101.175.217:8080/page/main

# 제 6 장 결론

본 논문에서는 OCR 모델 학습을 위한 데이터 취득이 제한적이라는 것에 문제로 삼고 이를 해결하기 위해 GAN을 이용한 데이터 증강을 제안하였다. 이를 위해 본 연구 주제에 맞게 적절히 변형된 생성자와 판별자를 통해 증강 생성 모델을 설계했고, 제안된 모델과 기존의 방식에 따른 문자 인식 성능을 비교하였다.

실험 결과, 기존의 방식에 비해 GAN을 통한 데이터 증강이 더 다채로운 품질의 데이터를 생성해낸다는 사실을 확인할 수 있었으며, 그에 따라 인식 모델의 성능 또한 기존에 비해 개선되었다. 주목할만한 점은 기존의 방식과 본 논문에서 제안한 방식을 같이 적용할 때 일종의 상승효과를 일으켜 큰 폭으로 인식 성능을 개선했다는 것이다.

다만, 생성자, 판별자 간의 균형을 맞추기 어려워 다른 모델을 적용하지 못한 점이나, 현실적인 요건으로 인해 타 언어 간의 호환성을 정확히 검증해보지 못한 아쉬움도 있었다. 이러한 부분들은 후속 연구를 통해서 개선되어 OCR 기술을 연구하는 데 있어 더 편의성을 줄 것으로 기대한다.

# 참 고 문 헌

[1] SUBRAMANI, Nishant, et al. A Survey of Deep Learning Approaches for OCR and Document Understanding. *arXiv preprint arXiv:2011.13534*, 2020.

[2] SHORTEN, Connor; KHOSHGOFTAAR, Taghi M. A survey on image data augmentation for deep learning. *Journal of Big Data*, 2019, 6.1: 1-48.

[3] BLOICE, Marcus D.; STOCKER, Christof; HOLZINGER, Andreas. Augmentor: an image augmentation library for machine learning. *arXiv preprint arXiv:1708.04680*, 2017.

[4] LUO, Canjie, et al. Learn to augment: Joint data augmentation and network optimization for text recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. p. 13746-13755.

[5] GOODFELLOW, Ian, et al. Generative adversarial nets. *Advances in neural information processing systems*, 2014, 27.

[6] BENGIO, Yoshua, et al. Better mixing via deep representations.

In: *International conference on machine learning*. PMLR, 2013. p. 552-560.

[7] BENGIO, Yoshua, et al. Deep generative stochastic networks trainable by backprop. In: *International Conference on Machine Learning*. PMLR, 2014. p. 226-234.

[8] LECUN, Yann, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, 86.11: 2278-2324.

[9] MAO, Xudong, et al. Least squares generative adversarial networks. In: *Proceedings of the IEEE international conference on computer vision*. 2017. p. 2794-2802.

[10] ARJOVSKY, Martin; CHINTALA, Soumith; BOTTOU, Léon. Wasserstein generative adversarial networks. In: *International conference on machine learning*. PMLR, 2017. p. 214-223.

[11] RADFORD, Alec; METZ, Luke; CHINTALA, Soumith. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[12] MIRZA, Mehdi; OSINDERO, Simon. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[13] ISOLA, Phillip, et al. Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017. p. 1125-1134.

[14] ZHU, Jun-Yan, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE international conference on computer vision*. 2017. p. 2223-2232.

[15] KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012, 25: 1097-1105.

[16] HE, Kaiming, et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 770-778.

[17] HUANG, Gao, et al. Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017. p. 4700-4708.

[18] CHEN, Yunpeng, et al. Dual path networks. *arXiv preprint arXiv:1707.01629*, 2017.

[19] NAIR, Vinod; HINTON, Geoffrey E. Rectified linear units improve restricted boltzmann machines. In: *Icml*. 2010.

[20] RAMACHANDRAN, Prajit; ZOPH, Barret; LE, Quoc V. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[21] JAUME, Guillaume; EKENEL, Hazim Kemal; THIRAN, Jean-Philippe. Funsd: A dataset for form understanding in noisy scanned documents. In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. IEEE, 2019. p. 1-6.

[22] BAEK, Jeonghun, et al. What is wrong with scene text recognition model comparisons? dataset and model analysis. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019. p. 4715-4723.

# 부록 A

## 글씨체 목록

마루부리 글꼴
https://hangeul.naver.com/2017/nanum

카페 24 서라운드
https://fonts.cafe24.com/

고운 돋움
https://github.com/yangheeryu/Gowun-Dodum

원스토어 모바일 POP
https://www.gg-onestore.com/Font

고운 바탕
https://github.com/yangheeryu/Gowun-Batang

페이북
https://paybooc.co.kr/app/paybooc/RPastEventDetail.do?evntNo=2020040025

티웨이 항공
https://www.twayair.com/app/serviceInfo/contents/1320

리디바탕
https://ridicorp.com/ridibatang/

조선일보명조
https://event.chosun.com/100/100font.html

D2 코딩
https://github.com/naver/d2codingfont

## Abstract

# Text Image Data Augmentation via Generative Adversarial Nets



Kyungmo Ha

Department of Computer Science and Engineering

College of Engineering

The Graduate School

Seoul National University



Numerous data of various qualities are required to train most of deep learning models, and model for optical character recognition is one of them. However, it costs a lot to obtain these data. So, data augmentation is attempted to overcome this problem. Unfortunately, only some limited augmented techniques are available due to the characteristics of optical character recognition. Even worse, these techniques have limitations in obtaining data of various qualities. In this paper, we proposed to improve these limitations by data augmentation using Generative Adversarial Nets(GAN) with Dual Path Networks(DPN) and compared with existing methods how much the proposed method actually affects recognition performance. As a result of the study, trained GAN with DPN can transforms images in

various ways compared with the existing augmentation. And accordingly, the performance of the character recognition model was also improved. Moreover, there was significant improvement in performance when conventional and proposed data augmentation are applied together. By these, we confirmed that data augmentation through GAN is effective in optical character recognition.

keywords : text image data augmentation, generative adversarial nets, dual path networks, optical character recognition, computer vision
*Student Number* : 2020-27787