DBpia

논문 20-45-07-04 The Journal of Korean Institute of Communications and Information Sciences '20-07 Vol.45 No.07
https://doi.org/10.7840/kics.2020.45.7.1166

# 생성적 적대 신경망과 GSVD 기반 LDA를 이용한 효과적인 학습 데이터셋 제작 및 선별

양 윤 지•, 홍 용 기*, 박 재 현°

# Efficient Learning Dataset Generation and Data Selection Using Generative Adversarial Network and GSVD-Based Linear Discriminant Analysis

Yunji Yang•, Yong-gi Hong*, Jaehyun Park°

요 약

본 논문은 딥러닝과 같은 기계학습 시 방대한 양의 학습용 데이터셋이 필요한 상황에서 생성적 적대 신경망 (Generative Adversarial Network, GAN)과 선형 판별 분석을 통한 적절한 데이터셋의 생성 및 선별 기법을 제안한다. 생성적 적대 신경망으로 생성한 유사 가짜 이미지와 실제의 데이터셋의 이미지를 통합한 데이터셋에서 GSVD 기반 선형 판별 분석 (Linear Discriminant Analysis, LDA)을 통해 축소된 차원에서 통합 데이터셋 중 훈련에 활용할 데이터들을 클러스터 간 거리에 기반한 다양한 기준으로 선별하였다. 선별 기준에 따른 성능 검증을 위해 문자 인식을 위한 심층 신경망 (Deep Neural Network, DNN)을 구성하여 선별된 데이터셋을 활용하여 훈련하여 문자 인식 정확도를 비교하였으며, 축소된 차원에서의 데이터셋 선별 기준의 유효성을 검증하였다. 제안한 최적 학습 데이터셋 선별 선처리 기법은 GAN을 통한 훈련용 데이터 제작 및 딥러닝 등의 학습을 진행할 때 효율적인 학습 데이터셋의 선별 방식 선정에 대한 지표가 될 수 있다.

**Key Words** : Generative Adversarial Network, Linear Discriminant Analysis, Learning Dataset, Preprocessing, Classification

ABSTRACT

In this paper, we propose a dataset generation and screening method through generative adversarial neural networks (GAN) and GSVD based linear discriminant analysis (LDA) in situations where a large amount of learning datasets are required for the deep learning applications. We first generate the fake dataset through the GAN and include them to the unified training dataset together with real measurement data samples. To reduce the dimension of the unified dataset, GSVD based LDA is applied and the data samples are selected in the reduced-dimensional space to train the deep neural network. We develop the deep neural network for the character recognition and evaluate the recognition accuracy to verify the validity of the proposed dataset generation and data selection methods.

※ 이 논문은 부경대학교 자율창의학술연구비(2019년)에 의하여 연구되었습니다.
• First Author : Pukyong National University, Department of Electronics Engineering, yyj26@pukyong.ac.kr, 학생회원
° Corresponding Author : Pukyong National University, Department of Electronics Engineering, jaehyun@pknu.ac.kr, 종신회원
* Pukyong National University, Department of Electronics Engineering
논문번호 : 202002-032-A-RN, Received February 21, 2020; Revised April 8, 2020; Accepted May 18, 2020



www.dbpia.co.kr

## Ⅰ. 서 론

최근 딥러닝 (또는 심층학습) 등의 머신러닝이 다양한 분야에 적용 되고 있고 이에 따른 연구도 활발하다. 딥러닝은 데이터가 많아지면 많아질수록 성능이 더 좋아진다[1]. Ian Goodfellow의 "Deep Learning"에 따르면 카테고리 당 대략 5000개 정도의 학습 데이터가 있어야 허용 성능 (Acceptable Performance)를 보이며, 인간에 필적하거나 뛰어넘기 위해서는 적어도 100만장의 학습 데이터가 필요하다[2]. 따라서 딥러닝을 적용하여 높은 성능을 내기 위해서는 학습을 위한 방대한 양의 데이터셋이 필요하지만, 분야에 따라 학습용 데이터의 수집이 쉽지 않을 수 있어 어려움이 있다. 딥러닝에 필요한 데이터셋을 제작하는 방안으로써 딥러닝의 비지도 학습법 (Unsupervised Learning) 중 하나인 생성적 적대 신경망 (Generative Adversarial Network, GAN)을 활용할 수 있다[3].

GAN은 생성자인 Generator와 판별자인 Discriminator 두 네트워크로 구성되어 있고 서로 경쟁하며 생성자가 실제 이미지와 유사한 이미지를 만들도록 학습할 수 있다. 특히, GAN을 이용한 Data Augmentation 연구는 최근에 많이 진행되었으며, 기존 연구의 경우 크게 데이터셋의 클래스가 불균형한 경우 (Oversampling) 부족한 클래스를 보충하기 위해서, 그리고 데이터가 민감한 정보를 포함할 때 이를 합성 이미지로 극복하는데 초점을 맞추고 있다[4]. 또한 [5], [6]에서는 데이터 분류 문제에서 식별률을 높이기 위한 Feature를 선택하는데 있어 심층학습을 활용하고 있으며, [7]에서는 심층 신경망 (Deep Neural Network, DNN)의 Hyper Parameter들을 선택하는데 있어 진화알고리즘을 활용하고 있다.

본 논문에서는 GAN을 이용한 Data Augmentation 관련 기존 연구와 달리, 통합된 훈련 데이터셋에서 DNN 학습을 위한 훈련용 데이터 선별 기법을 제안하고자 한다. 훈련용 데이터 생성을 위해 기존에 확보해 둔 측정 데이터를 활용하여 GAN을 학습하여 유사 가짜 데이터를 생성하며, 이를 활용하여 유사 가짜 데이터와 기존 확보해 놓은 진짜 데이터를 통합한 훈련 데이터셋을 만들고 이 훈련 데이터셋에서 학습에 적합한 데이터셋을 선별하기 위해 일반적 고유값 분해 (Generalized Singular Value Decomposition, GSVD) 기반의 선별 알고리즘을 제안한다. 참고로, 딥러닝을 위한 방대한 양의 데이터를 수집하고 그에 따라 학습에 적합한 훈련용 데이터셋을 선별하는 기법에 대한 연구는 현재까지 많이 이루어지지 않았다. 대부분의 기존 심층학습에서는 수집된 데이터셋에서 임의적으로 (Random 하게) Batch를 구성하여 신경망을 학습시키고 있다[5,6].

이와 함께 본 논문에서는 훈련용 데이터 선별을 위해 GSVD 기반 선형 판별 분석 (Linear Discriminant Analysis, LDA) 방식[8-10]을 통해 차원이 축소된 공간에서의 기하학적 관계를 해석한다. 여기서 LDA는 고차원의 데이터를 해당 클러스터 별로 분류하여 Between-cluster 분산은 최대로 하고 Within-cluster 분산은 최소로 하여 클러스터의 수 수준의 낮은 차원에 선형 맵핑하는 방법이다. 즉 고차원의 데이터들의 기하학적 관계를 유지한 채 저차원에서 기하학적 관계를 분석할 수 있으며, 이 때 필요한 선형 변환 행렬을 효율적으로 계산하기 위해 본 논문에서는 GSVD 알고리즘을 활용하며, 이와 함께 축소된 차원에서 훈련용 데이터 선별 방식을 제안한다. 구체적으로, LDA/GSVD 알고리즘을 통해 차원 축소 뒤, 낮은 차원에서 각 클러스터 별 중심에서 멀리 떨어진 데이터를 효과적으로 선별하여 DNN 학습에 활용함으로써 신경망 학습 정확도를 높일 수 있다.

선별 기준에 따른 성능 검증을 위해 본 논문에서는 문자 인식을 위한 DNN을 구성하였으며, GAN에서 생성한 유사 가짜 데이터셋과 실제 데이터셋을 통합한 뒤 LDA를 통해 차원을 축소하고, 축소된 공간에서 각각의 클러스터에 대하여 클러스터 중심에 가까운 그룹, 클러스터 중심에서 먼 그룹 그리고 클러스터 내 무작위로 선발한 그룹으로 총 세 그룹으로 나누어 DNN을 학습하여 그에 따른 문자 인식 정확도를 비교하였다. 모의실험에 사용된 데이터는 MNIST (Mixed National Institute of Standards and Technology, MNIST) 데이터셋으로, 길쌈신경망 (Convolutional Neural Network, CNN)으로 학습 시 100%에 가까운 문자 인식 정확도를 달성할 수 있지만[11], 선별 기법에 따른 결과를 비교하기 위하여 본 논문에서는 비교적 간단한 구조의 DNN 심층학습을 고려하였다. 하지만 제안하는 알고리즘은 CNN을 포함한 다양한 구조의 신경망 학습에도 확장이 가능할 것으로 예상한다.

이 논문의 2장에서는 본 논문에서 사용한 문자 인식을 위한 DNN과 GAN의 네트워크 모델을 소개한다. 3장에서는 제안하는 LDA/GSVD 알고리즘 기반 훈련용 데이터 선처리 기법을 설명한다. 4장에서는 모의실험 결과를 확인하고 5장에서는 결론을 맺는다.



www.dbpia.co.kr

## Ⅱ. 문자 인식용 DNN 및 GAN 모델

### 2.1 문자 인식용 DNN

문자 인식을 위한 DNN은 기계학습법 중 지도학습에 해당하는 분류로써 널리 연구가 되고 있다. 지도학습은 라벨 (Label)이 정해져 있는 데이터셋를 바탕으로 신경망을 학습 시킨다. 지도학습의 방법은 크게 분류 (Classification)와 회귀 (Regression)로 나눌 수 있는데 본 논문의 모의실험에서 다룰 MNIST 손글씨 문자 인식용 데이터셋을 위해서는 다중 분류 방식을 사용한다. 딥러닝 모델의 구조는 그림 1과 같이 Fully Connected Network 기반 DNN을 사용하였다[1].

DNN은 입력층 (Input Layer)와 출력층(Output Layer) 사이에 여러 개의 은닉층 (Hidden Layer)들로 이뤄진 인공신경망 (Artificial Neutral Network, ANN)이다. DNN은 표준 오류 역전파 알고리즘으로 학습한다. 이때 가중치 (Weight)들은 아래 (1)을 이용한 확률적 경사 하강법 (Stochastic Gradient Descent)을 통하여 갱신될 수 있다.

$$\triangle w_{ij}(t+1)=\triangle w_{ij}(t)+\eta\frac{\partial C}{\partial w_{ij}} \tag{1}$$

여기서 $\eta$는 학습률 (Learning Rate), $C$는 비용함수 (Cost Function)를 의미한다. 비용함수는 학습의 형태와 활성 함수 (Activation Function)같은 요인들에 의해서 결정된다. 본 논문의 신경망에서는 활성 함수와 비용 함수는 각각 ReLu (Rectified Linear Unit, ReLu) 함수와 Sigmoid Cross-Entropy (Sigmoid CE) 함수를 사용하였다. ReLu 함수는 아래 (2)와 같이 입력 값이 0보다 작으면 0이고 0보다 크면 입력 값 그대로를 내보낸다.

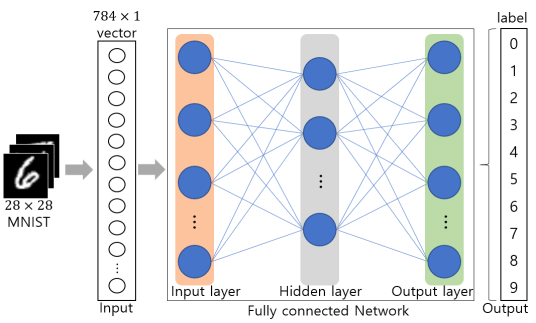



그림 1. 문자 인식을 위한 완전 연결 심층 신경망 구조
Fig. 1. Fully connected deep neural network structure for character recognition

$$f(x)=\max(x,0) \tag{2}$$

Sigmoid CE 함수는 (4)와 같이 주어지는 CE를 Sigmoid 함수의 입력으로 대입한 (3)과 같이 주어진다.

$$f(s_i)=\frac{1}{1+e^{-s_i}} \tag{3}$$

$$CE=-t_1\log(f(s_1))-(1-t_1)\log(1-f(s_1)) \tag{4}$$

본 논문에서는 간단한 형태의 문자 인식을 위한 DNN을 고려하지만 제안하고자 하는 훈련 데이터 제작 및 선별 방식은 일반적인 지도학습을 위한 신경망 응용 분야로 확장할 수 있다. 또한, 각 문자 인식은 0, 1,...,9의 문자를 고려하고 있으며 각 문자 별 훈련용 데이터의 개수는 $N_0=N_1=...=N_9=N$으로 같은 수의 훈련용 데이터를 활용한다고 가정한다.

### 2.2 생성적 적대 신경망

그림 2와 같이 GAN은 생성자 (Generator)와 판별자 (Discriminator) 두 네크워크가 서로 경쟁하는 구조로 구성되어 있다. 즉 $G(x)$를 생성자에서 생성되는 출력의 함수이고, $D(z)$는 판별자의 판별 함수라 두자. 그러면 GAN의 학습을 위한 가치 함수는 다음과 같이 설정할 수 있다[3].

$$\min_G\max_D V(D,\ G)$$
$$=E_{x\sim p_{data}}[\log D(x)]+E_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{5}$$

즉, 판별자는 생성자가 만든 데이터와 실제 데이터를 받아 그 데이터가 진짜인지 생성자가 만든 가짜 데이터인지를 판별하고 생성자는 판별자의 판별 확률을 보고 진짜 데이터와 유사한 데이터를 생성하도록 학습하며 서로 경쟁하도록 한다. (5)에서 판별자는 가치 함수 $V(D, G)$가 최대가 되도록 $D(z)$를 수렴하거나 학습하여야 하는 반면, 생성자는 $V(D, G)$가 최소가

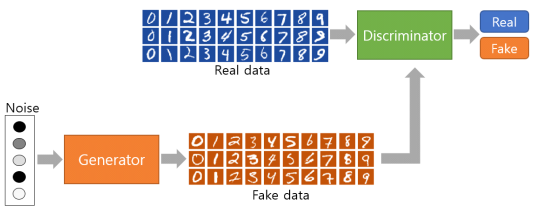



그림 2. 생성적 적대 신경망 구조
Fig. 2. GAN structure



www.dbpia.co.kr

되도록 $G$를 학습하여야 한다. 구체적으로 생성자는 임의의 잡음을 입력 $z$로 받아 출력 $x = G(z)$로 맵핑하여 유사 데이터를 생성한다. 이때 학습이 진행됨에 따라 $G$는 실제 데이터의 확률 분포를 학습하여 출력 데이터의 확률 분포가 실제 데이터의 것과 유사하게 가지도록 입력 잡음을 출력 데이터로 매핑하게 된다. 이에 따라 판별자의 분류 확률은 0.5에 수렴하게 되어 진짜 데이터와 생성자가 만든 가짜 데이터를 구분할 수 없게 된다. 학습이 완료된 GAN의 생성자를 활용하여 진짜 데이터와 유사한 데이터를 생성할 수 있게 된다[3]. 즉, GAN의 학습이 끝난다면, 그림 2에서 생성자 네트워크를 활용하여, GAN을 학습할 때 활용된 실제 데이터와 유사한 가짜 데이터를 생성할 수 있다.

## Ⅲ. 데이터 선처리 시 선형 판별 분석

본 절에서는 앞서 설명한 GAN에서 생성한 유사 가짜 데이터를 활용하여 문자 인식용 DNN을 학습시키는 방안에 대해서 설명한다. 제안하는 선처리 기법은 그림 3과 같다. 우선 0~9에 해당하는 실제 측정한 문자 데이터를 활용하여 가짜 데이터 생성을 위해 GAN을 학습한다. GAN의 학습이 끝나게 되면 GAN의 생성자 네트워크를 통해 유사 가짜 데이터를 0~9까지 생성할 수 있다. 이로 인해 각 라벨 별로 유사 가짜 데이터와 측정 데이터를 통합하여 기존 측정 데이터만을 활용할 때보다 많은 풀의 데이터셋을 가질 수 있게 된다. 이 통합 데이터셋에서 문자 인식용 DNN 학습을 위해 데이터를 선별할 수 있는데, 이를 위해 본 논문에서는 LDA/GSVD 알고리즘을 이용하여 실제 데이터의 차원보다 낮은 차원에서 훈련용 데이터를 선택하는 선처리 기법을 제안한다.

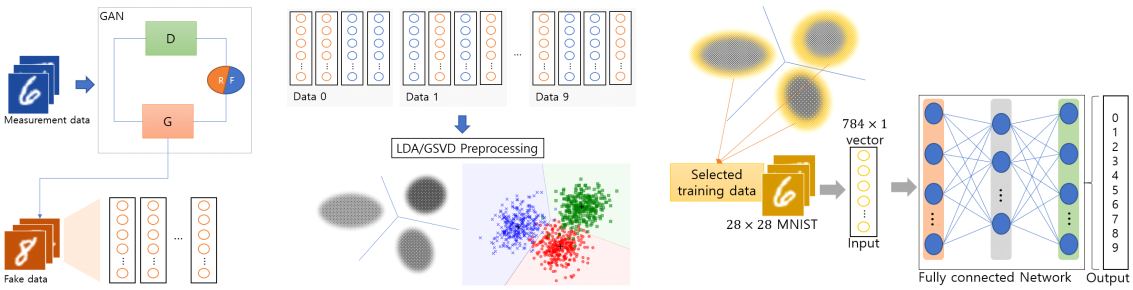
### 3.1 심층망 학습용 훈련 데이터 선별을 위한 LDA/GSVD 알고리즘

본 절에서는 DNN 학습용 훈련 데이터 선별을 위한 LDA/GSVD 알고리즘에 대하여 제안한다. LDA는 여러 클래스에 속하는 높은 차원의 데이터들을 클래스 간의 식별 거리 (Classification Distance)가 최대가 되면서 낮은 차원으로 축소하는 선형 변환(이를테면, 선형 변환 행렬)을 찾는 알고리즘이다[8].

$m$-차원의 데이터 벡터 $\boldsymbol{a}_i^{(k)}$를 $k$번째 클러스터의 $i$번째 데이터 샘플이라 두자. 그러면 $\boldsymbol{a}_i^{(k)}$를 $l$-차원 데이터 $\boldsymbol{y}_i^{(k)}$로 변환하는 선형변환 $\boldsymbol{G} \in \mathbb{R}^{m \times l}$를 찾는 것이 LDA/GSVD 알고리즘이다[8,9].

$$\boldsymbol{G}^T : \boldsymbol{a} \in R^{m \times 1} \rightarrow \boldsymbol{y} \in R^{l \times 1} \tag{6}$$

구체적인 설명을 위해 데이터 샘플 $\boldsymbol{a}_i^{(k)}$를 열벡터로 가지며, 열이 $K$개의 클러스터로 분할되는 데이터 행렬 $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_K] \in \mathbb{R}^{m \times \overline{N}}$을 정의하자. 이때 클러스터 별 통합 데이터셋의 샘플 개수는 $\overline{N_k}$개가 있으며 $\sum_{k=1}^{K} \overline{N_k} = \overline{N}$가 된다고 가정하자. 클러스터 간의 식별 정도를 나타내기 위해 클러스터 내 평균을 $\boldsymbol{c}^{(k)} = \frac{1}{\overline{N_k}} \sum_{i=1}^{\overline{N_k}} \boldsymbol{a}_i^{(k)}$으로 정의하고 전체 데이터 샘플의 평균을 $\boldsymbol{c} = \frac{1}{\overline{N}} \sum_{k=1}^{K} \sum_{i=1}^{\overline{N_k}} \boldsymbol{a}_i^{(k)}$로 정의한다. 이때 각 클러스터의 샘플이 클러스터 내 평균값 $\boldsymbol{c}^{(k)}$으로부터의 분산은 최소가 되면서, 각 클러스터 내 평균값 $\boldsymbol{c}^{(k)}$가 전체 평균 $\boldsymbol{c}$로부터의 분산은 최대가 될수록 클러스터간 데




(a) (b) (c)

그림 3. 제안한 심층 데이터셋 생성 및 훈련 데이터 선택 방법: (a) GAN을 활용한 유사 가짜 훈련 데이터 생성 (b) LDA/GSVD를 활용한 차원 축소 및 데이터 선택 (c) 선택된 데이터를 활용한 DNN 훈련

Fig. 3. Proposed learning dataset generation and training data selection method: (a) Pseudo fake data generation using GAN (b) Dimension reduction using LDA/GSVD and Data selection (c) DNN training with selected training data set



www.dbpia.co.kr

이터 식별이 용이해진다. 이를 판별하기 위한 선형변환 되기 이전의 식별 거리는 다음과 같이 정의된다[9].

$$J=\frac{trace(\boldsymbol{H}_b\boldsymbol{H}_b^T)}{trace(\boldsymbol{H}_w\boldsymbol{H}_w^T)} \tag{7}$$

여기서 $\boldsymbol{H}_w\in\mathbb{R}^{m\times\overline{N}}$와 $\boldsymbol{H}_b\in\mathbb{R}^{m\times K}$을 다음과 같이 나타낼 수 있다.

$$\boldsymbol{H}_w=\left[\boldsymbol{A}_1-\boldsymbol{c}^{(1)}\boldsymbol{e}^{(1)^T},\ \cdots,\boldsymbol{A}_K-\boldsymbol{c}^{(K)}\boldsymbol{e}^{(K)^T}\right]\in\mathbb{R}^{m\times\overline{N}}$$

$$\boldsymbol{H}_b=\left[\sqrt{\overline{N_1}}(\boldsymbol{c}^{(1)}-\boldsymbol{c}),\cdots,\ \sqrt{\overline{N_K}}(\boldsymbol{c}^{(K)}-\boldsymbol{c})\right]\in\mathbb{R}^{m\times K}$$

이때 $\boldsymbol{e}^{(k)}$와 $\boldsymbol{e}$는 다음과 같이 주어진다.

$$\boldsymbol{e}^{(k)}=(1,\ \cdots,1)^T\in\mathbb{R}^{\overline{N_k}\times 1},$$

$$\boldsymbol{e}=(1,\ \cdots,1)^T\in\mathbb{R}^{\overline{N}\times 1}.$$

(6)의 선형 변환 뒤, 식별 거리는 다음과 같이 표현할 수 있다.

$$J_{LDA}=\frac{trace(\boldsymbol{G}^T\boldsymbol{H}_b\boldsymbol{H}_b^T\boldsymbol{G})}{trace(\boldsymbol{G}^T\boldsymbol{H}_w\boldsymbol{H}_w^T\boldsymbol{G})} \tag{8}$$

따라서, LDA/GSVD 알고리즘은 차원 축소 후 식별 거리 $J_{LDA}$가 최대가 되는 선형 변환 $\boldsymbol{G}$를 찾는 것이며, 구체적으로 이러한 선형 변환 $\boldsymbol{G}$는 행렬 쌍 $(\boldsymbol{H}_b^T,\ \boldsymbol{H}_w^T)$의 GSVD의 가장 큰 Generalized Singular Value들에 해당하는 Generalized Singular Vector들로 이루어진다[8,9]. 또한 논문 [9]의 Theorem 2와 [12]로부터 행렬 쌍 $(\boldsymbol{H}_b^T,\ \boldsymbol{H}_w^T)$에 대해 다음을 만족하는 직교행렬 $\boldsymbol{U}\in\mathbb{R}^{K\times K}$, $\boldsymbol{V}\in\mathbb{R}^{\overline{N}\times\overline{N}}$, $\boldsymbol{W}\in\mathbb{R}^{t\times t}$, $\boldsymbol{Q}\in\mathbb{R}^{m\times m}$은 항상 찾을 수 있다고 알려져 있다.

$$\begin{aligned}\boldsymbol{U}^T\boldsymbol{H}_b^T\boldsymbol{Q}&=\boldsymbol{\Sigma}_b[\boldsymbol{W}^T\boldsymbol{R},\ \boldsymbol{0}],\\ \boldsymbol{U}^T\boldsymbol{H}_w^T\boldsymbol{Q}&=\boldsymbol{\Sigma}_w[\boldsymbol{W}^T\boldsymbol{R},\ \boldsymbol{0}]\end{aligned} \tag{9}$$

여기서 $t$는 행렬 $\boldsymbol{S}=\begin{pmatrix}\boldsymbol{H}_b^T\\ \boldsymbol{H}_w^T\end{pmatrix}\in\mathbb{R}^{(K+\overline{N})\times m}$의 Rank를 나타내며, $\boldsymbol{\Sigma}_b$와 $\boldsymbol{\Sigma}_w$는 각각 대각행렬이고, $\boldsymbol{R}\in\mathbb{R}^{t\times t}$은 Nonsingular 행렬이다. 따라서 (9)로부터 다음을 유도할 수 있다.

$$\begin{aligned}\boldsymbol{H}_b^T&=\boldsymbol{U}[\boldsymbol{\Sigma}_b,\ \boldsymbol{0}]\boldsymbol{X}^{-1},\\ \boldsymbol{H}_w^T&=\boldsymbol{U}[\boldsymbol{\Sigma}_w,\ \boldsymbol{0}]\boldsymbol{X}^{-1}\end{aligned} \tag{10}$$

여기서 $\boldsymbol{X}=\boldsymbol{Q}\begin{pmatrix}\boldsymbol{R}^{-1}\boldsymbol{W} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{pmatrix}\in\mathbb{R}^{m\times m}$로 주어진다. 즉, $\boldsymbol{X}$의 열벡터가 행렬 쌍 $(\boldsymbol{H}_b^T,\ \boldsymbol{H}_w^T)$의 Generalized Singular Vector임을 알 수 있으며, 위의 유도를 바탕으로 본 논문의 훈련 데이터 선별용 선형 변환 $\boldsymbol{G}$를 찾기 위한 알고리즘은 Algorithm 1의 내용과 같이 정리할 수 있다.

Algorithm 1을 통해 훈련용 데이터의 차원을 축소

Algorithm 1. 훈련 데이터 선별 위한 LDA/GSVD 알고리즘

| | |
|---|---|
| 열이 $K$개의 클러스터로 분할되는 데이터 행렬 $\boldsymbol{A}=[\boldsymbol{A}_1,\boldsymbol{A}_2,\ldots,\boldsymbol{A}_K]\in\mathbb{R}^{m\times\overline{N}}$가 주어지면 알고리즘은 차원 축소 변환 $\boldsymbol{G}\in\mathbb{R}^{m\times l}$를 계산한다. | |
| 1) | $\boldsymbol{A}$로부터 $\boldsymbol{H}_w\in\mathbb{R}^{m\times\overline{N}}$와 $\boldsymbol{H}_b\in\mathbb{R}^{m\times K}$을 다음과 같이 정리한다.<br>$\boldsymbol{H}_w=\left[\boldsymbol{A}_1-\boldsymbol{c}^{(1)}\boldsymbol{e}^{(1)^T},\ \cdots,\boldsymbol{A}_K-\boldsymbol{c}^{(K)}\boldsymbol{e}^{(K)^T}\right]\in\mathbb{R}^{m\times\overline{N}}$<br>$\boldsymbol{H}_b=\left[\sqrt{\overline{N_1}}(\boldsymbol{c}^{(1)}-\boldsymbol{c}),\cdots,\ \sqrt{\overline{N_K}}(\boldsymbol{c}^{(K)}-\boldsymbol{c})\right]\in\mathbb{R}^{m\times K}$ |
| 2) | $\boldsymbol{S}=\begin{pmatrix}\boldsymbol{H}_b^T\\ \boldsymbol{H}_w^T\end{pmatrix}\in\mathbb{R}^{(K+\overline{N})\times m}$를 다음과 같이 완전 직교 분해 (Complete Orthogonal Decomposition)한다.<br>$\boldsymbol{S}=\boldsymbol{P}\begin{pmatrix}\boldsymbol{R}\boldsymbol{0}\\ \boldsymbol{0}\,\boldsymbol{0}\end{pmatrix}\boldsymbol{Q}^T,\ \boldsymbol{R}\in\mathbb{R}^{t\times t}$ (11)<br>여기서 $t=rank(\boldsymbol{S})$ (일반적으로 $t>K$)이고, $\boldsymbol{P}\in\mathbb{R}^{(K+\overline{N})\times(K+\overline{N})}$과 $\boldsymbol{Q}\in\mathbb{R}^{m\times m}$은 직교행렬이며, 완전직교분해의 경우 $\boldsymbol{R}$은 상삼각행렬로 주어진다. |
| 3) | 행렬 $\boldsymbol{P}$의 처음 $t$개의 행을 가지는 부분 행렬 $\boldsymbol{P}_1$를 다음과 같이 Partition을 나눈다.<br>$\boldsymbol{P}_1=\begin{pmatrix}\boldsymbol{P}_{1b}\\ \boldsymbol{P}_{1w}\end{pmatrix}\in\mathbb{R}^{(K+\overline{N})\times t},\ \boldsymbol{P}_{1b}\in\mathbb{R}^{K\times t}$ (12) |
| 4) | $\boldsymbol{P}_{1b}$에 대해 다음과 같이 SVD를 수행한다.<br>$\boldsymbol{P}_{1b}=\boldsymbol{U}\boldsymbol{\Sigma}_b\boldsymbol{W}^T$ (13)<br>즉, $\boldsymbol{U}\in\mathbb{R}^{K\times K}$와 $\boldsymbol{W}\in\mathbb{R}^{t\times K}$는 직교한 열을 가지는 행렬이며, $\boldsymbol{\Sigma}_b$는 $\boldsymbol{P}_{1b}$의 고유값($\sigma_1\geq\sigma_2\geq\ldots\geq\sigma_K\geq 0$)을 대각원소로 가지는 대각행렬이 된다. |
| 5) | $\boldsymbol{Q}\begin{pmatrix}\boldsymbol{R}^{-1}\boldsymbol{W} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{pmatrix}$의 첫 $l$개의 열을 계산하여 $\boldsymbol{G}\in\mathbb{R}^{m\times l}$에 할당한다. |
| 6) | 계산된 $\boldsymbol{G}$를 활용하여 각 데이터들은 다음과 같이 $l$차원의 벡터로 선형 변환 된다.<br>$\boldsymbol{G}^T:\boldsymbol{a}\in\mathbb{R}^{m\times 1}\rightarrow\boldsymbol{y}\in\mathbb{R}^{l\times 1}$ |



www.dbpia.co.kr

하는 선형 변환 행렬 $\boldsymbol{G} \in \mathbb{R}^{m \times l}$를 구할 수 있으며, 입력 벡터 $\boldsymbol{x}(\in \mathbb{R}^{m \times 1})$에 대해 $\boldsymbol{y} = \boldsymbol{G}^T\boldsymbol{x}(\in \mathbb{R}^{l \times 1})$로 $l$-차원으로 표현할 수 있다. 다음 절에서는 Algorithm 1을 통해 변환된 낮은 차원에서 효과적인 심층학습용 훈련용 데이터들을 선택하기 위한 다양한 기준들을 제시한다.

### 3.2 LDA/GSVD 알고리즘을 통한 축소된 차원에서의 훈련용 데이터 선별 방식

2.2절에서 GAN 학습을 통해 가짜 데이터를 생성하는 방법을 제시하였으나, 문자 인식용 신경망을 학습하기 위한 최적의 훈련용 데이터를 선택하기 위한 기준은 명확하게 알려지지 않았다. 직관적으로 문자별 (일반적인 지도 학습의 경우, 각 클러스터 또는 라벨 별) 경계에 있는 데이터들에 대해 신경망이 학습이 잘 되면 우수한 성능을 낼 수 있을 것으로 예상한다. 하지만 고차원의 측정 데이터 (참고로, MNIST 손글씨 데이터의 경우 784차원을 가지고 있다.)의 경우 클러스터별 경계에 있는지에 대한 기하학적인 관계를 판단하기 어렵거나 기하학적인 관계를 판단하기 위한 계산 복잡도가 매우 높게 된다.

3.1절에서 제안한 LDA/GSVD 알고리즘을 활용해 여러 클래스 간의 식별거리가 최대가 되면서 데이터들의 차원을 축소한다면 클러스터 경계에 있는 데이터 선별을 용이하게 할 수 있다. 즉 축소된 차원 ($l << m$)에서의 평균값은 $\boldsymbol{G}^T\boldsymbol{c}^{(k)}$로 표현할 수 있으며, 축소된 데이터들 $\boldsymbol{y}_i^{(k)}$중 거리가 먼 순서대로 훈련용 데이터를 용이하게 선별할 수 있다. 이때, 클러스터 경계를 넘어 발생된 Outlier 데이터를 제거하기 위해 각 클러스터 중심 간의 평균 거리를 다음과 같이 계산할 수 있다.

$$D_{out} = \frac{1}{CLSUBK_2}\sum_{k \neq j}\|\boldsymbol{G}^T\boldsymbol{c}^{(k)} - \boldsymbol{G}^T\boldsymbol{c}^{(j)}\|_2$$

따라서 클러스터 경계를 넘어서는 데이터 (구체적으로 $\|\boldsymbol{y}_i^{(k)} - \boldsymbol{G}^T\boldsymbol{c}^{(k)}\|_2 > D_{out}/2$에 부합하는 데이터) $\boldsymbol{y}_i^{(k)}$는 Outlier로 지정하여 통합 학습 데이터셋에서 제거할 수 있다. 이 후 $k$번째 클러스터에 대해 $N$개의 훈련용 데이터 선별을 다음과 같이 진행할 수 있다.

벡터 $\boldsymbol{y}_{(i)}^{(k)}$는 길이 $\|\boldsymbol{y}_i^{(k)} - \boldsymbol{G}^T\boldsymbol{c}^{(k)}\|_2$가 $i$번째로 큰 벡터라고 정의하자. 따라서 다음의 수식을 만족하게 된다.

$$\|\boldsymbol{y}_{(i)}^{(k)} - \boldsymbol{G}^T\boldsymbol{c}^{(k)}\|_2 \geq \|\boldsymbol{y}_{(i+1)}^{(k)} - \boldsymbol{G}^T\boldsymbol{c}^{(k)}\|_2$$

낮은 차원에서 $\boldsymbol{y}_i^{(k)}$의 정렬을 통해 $\boldsymbol{y}_{(i)}^{(k)}$를 획득할 수 있으며, 훈련용 데이터는 $\boldsymbol{y}_{(1)}^{(k)}, \boldsymbol{y}_{(2)}^{(k)}, \ldots, \boldsymbol{y}_{(N)}^{(k)}$를 선택하면 DNN 학습 때 클러스터 경계에 대한 학습을 용이하게 할 수 있게 된다.

## Ⅳ. 모의실험 결과

제안하는 LDA/GSVD 알고리즘을 활용하여 축소된 차원에서의 선별 기법 성능을 검증하기 위해 간단한 문자 인식을 위한 신경망과 유사 가짜 데이터를 생성하기 위한 GAN을 제작하였다. 문자 인식을 위해 측정된 데이터로써 MNIST 손글씨 데이터셋을 활용하였으며, 이 데이터의 차원은 28x28 pixel로써 벡터로 나타내면 784x1의 차원을 가지게 된다.

GAN의 생성자는 노드의 수가 256인 은닉층 하나를 두었다. 활성 함수는 (2)의 ReLu 함수를 활용하였으며, 출력층은 (3)의 Sigmoid 함수를 활용하였다. 생성자의 입력으로는 임의의 균등한 분포를 가지는 잡음을 주었다. 판별자도 노드의 수가 256인 은닉층 하나를 두고 활성 함수는 (2)의 ReLu 함수를 활용하였고, 출력층은 (3)의 Sigmoid 함수를 활용하였다. 생성자와 판별자의 손실 함수는 (3)과 (4)로 나타나는 Sigmoid CE 함수를 활용하여 진짜 데이터를 이용하는 판별자의 값은 1에 가깝게, 생성자가 만든 데이터를 이용하는 판별자의 손실값은 0에 가깝게 만들게 한다. 생성자의 손실 값은 생성자가 만든 데이터를 이용하는 판별자의 값을 1에 가깝게 만드는 값으로 지정해주었다. ADAM (Adaptive Learning Rate Method, ADAM) 함수를 활용하여 학습하도록 하였으며 학습률은 $10^{-3}$으로 설정하였다.

문자 인식용 DNN은 은닉층은 5개를 두었으며, 층별 노드의 개수는 256, 128, 64, 32, 16로 줄어들도록 하였다. 활성 함수는 (2)의 ReLu 함수를 활용하였으며, 출력층에서의 손실 함수는 (3)과 (4)의 Sigmoid CE 함수를 활용하였다. 신경망학습은 ADAM 함수를 활용하여 학습하도록 하였으며 학습률은 $10^{-3}$으로 설정하였다.

그림 4에서는 간단한 성능을 검증하기 위해 MNIST 손글씨 중 두 개의 라벨 (0과 6, 2와 3, 3과 8, 3과 9)에 대해 인식률 성능을 검증하였다. 제안하는 기법의 우수성을 비교하기 위해 축소된 차원에서 통합 데이터셋에서 각 클러스터 중심에 가까운 순으



www.dbpia.co.kr

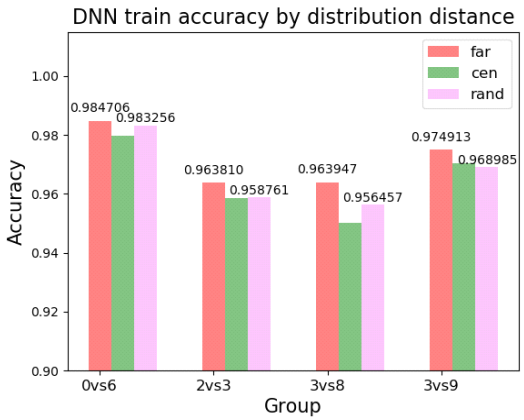



그림 4. 훈련 데이터 선택 방식에 따른 두 개 문자 인식에 대한 인식률 비교
Fig. 4. Accuracy comparison for two character recognition (0,6), (2,3), (3,8), (3,9) with three different training data selection methods.

로 훈련용 데이터를 선별하였을 때와 무작위로 훈련용 데이터를 선별하였을 때의 성능을 함께 비교하였다. 즉, MNIST 손글씨 데이터셋을 활용하여 GAN 학습 후, 그림 3 (a)와 같이 GAN에서 만든 가짜 이미지를 생성한 후 GAN 학습에 활용된 MNIST 손글씨 데이터셋과 유사 가짜 이미지를 섞어 통합 데이터셋을 만들었다. 그리고 서로 다른 세 가지 선별 방식에 따라 훈련용 데이터를 선별하여 문자 인식용 DNN 학습 후 분류 정확도를 비교하였다. 그림에서 확인할 수 있듯이 3.2절에서 제안하는 기법을 통해 훈련용 데이터를 선별하였을 때 우수한 분류 정확도를 가지는 것을 확인할 수 있다. 즉, 낮은 차원에서 샘플 평균값에서

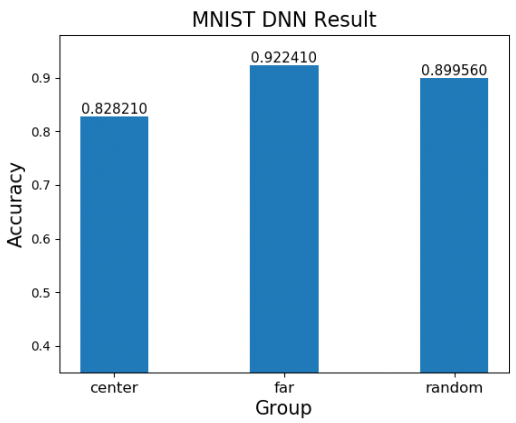



그림 5. 훈련 데이터 선택 방식에 따른 10개 문자 (0에서 9) 인식률 비교
Fig. 5. Accuracy comparison for ten character recognition (0 to 9) with three different training data selection methods.

멀수록, 문자별 (일반적인 지도 학습의 경우, 각 클러스터 또는 라벨별) 경계에 있는 데이터들에 가까우며, 이들을 선별하였을 때 문자 인식 학습이 더 잘되는 것으로 판단할 수 있다.

그림 5에서는 일반적인 0~9까지의 라벨에 대해 성능을 검증하기 위해 MNIST 손글씨 데이터셋을 활용하여 0에서 9까지의 10개의 라벨에 대하여 앞선 실험과 동일한 방식으로 진행하여 그 결과를 비교하였다. 그림에서 확인할 수 있듯이 제안하는 기법이 평균값 중심에서 선별된 데이터에 비해 10% 정도, 랜덤하게 선별하는 것에 비해 3% 정도 더 우수한 인식률을 보이는 것을 확인할 수 있다.

## V. 결 론

본 논문에서는 MNIST 손글씨 데이터셋에 대하여 GAN을 사용하여 만든 가짜 이미지와 기존의 데이터셋에 존재하는 진짜 이미지를 통합하여 훈련용 데이터를 늘릴 수 있는 방안에 대하여 연구하였다. 또한 훈련용 데이터를 선별하기 위한 선처리로 LDA를 통해 축소된 차원에서의 훈련용 데이터 선별 방식을 제안하였다. 즉, LDA/GSVD 알고리즘을 통해 차원 축소 뒤, 낮은 차원에서 각 클러스터 별 중심에서 멀리 떨어진 데이터를 선별하여 DNN의 심층학습 훈련에 활용하여 학습 정확도를 높일 수 있었다. 구체적인 문자 인식 DNN 학습에 적용하여 성능을 검증하였으며, 이를 바탕으로 일반적인 신경망 학습에도 확장할 수 있을 것으로 기대한다. 즉, 기존 측정 데이터에 GAN을 활용하여 생성한 유사 가짜 데이터를 늘려 LDA/GSVD 알고리즘으로 선처리를 해주어 클러스터 중심에서 거리가 먼 데이터를 효과적으로 선별하여 심층망 학습에 활용한다면 학습 효과를 더 높일 수 있을 것이다.

## References

[1] Andrew Ng, "*Machine Learning Yearning: Technical Strategy for AI Engineers, In the Era of Deep Learning*," 2018.

[2] I. Goodfellow, Y. Bengio, and A. Courville. "Deep Learning," *MIT Press*, 2016.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets," *NIPS*, Jun. 2014.



www.dbpia.co.kr

[4] F. H. K. S. Tanaka and C. Aranha, "*Data Augmentation with Using GANs*," arXiv:1904.09135v1, Apr. 2019.

[5] V. B. Semwal, K. Mondal, and G. C. Nandi, "Robust and accurate feature selection for humaniod push recovery and classification: Deep learning approach," *Neural Computing & Applications*, vol. 28, pp. 565-574, 2017.

[6] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. and Remote Sensing Lett.*, vol. 12, no. 11, pp. 2321-2325 Nov. 2015.

[7] S. R. Young, D. C. Rose, T. P. Karnowski, S. H. Lim, and R. M. Patton, "Optimizing deep learning hyper-parameters through an evolutionary algorithm," *SC15: The Int. Conf. High Performance Comput., Netw., Storage and Anal.*, Austin Texas, Nov. 2015.

[8] H. Park, B. Drake, S. Lee, and C. Park, "Fast linear discriminant analysis using QR decomposition and regularization," *Georgia Inst. Technol.*, Mar. 2007.

[9] P. Howland and H. Park, "Generalizing discriminant analysis using the generalized singular value decomposition," *IEEE TPAMI*, vol. 26, pp. 995-1006, Aug. 2004.

[10] J. Park, J. Chun, and H. Park, "Generalised singular value decomposition-based algorithm for multi-user multiple-input multiple-output linear precoding and antenna selection," *IET Commun.*, vol. 4, no. 16, pp. 1899-1907, Nov. 2010.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proc. IEEE*, vol. 86, pp. 2278-2324, 1998.

[12] C. C. Paige and M. A. Saunders, "Towards a generalized singular value decomposition," *SIAM J. Numerical Anal.*, vol. 18, no. 3, pp. 398-405, 1981.

**양 윤 지** (Yunji Yang)

2017년 3월~현재 : 부경대학교 전자공학과 학사 과정

<관심분야> 딥러닝, 통신신호처리, 디지털 신호처리

**홍 용 기** (Yong-gi Hong)

2020년 2월 : 부경대학교 전자공학과 학사

2020년 3월~현재 : 부경대학교 전자공학과 석사 과정

<관심분야> 통신공학, 디지털 신호처리, MIMO radar

**박 재 현** (Jaehyun Park)

2003년 2월 : KAIST 전기 및 전자 공학과 학사

2010년 2월 : KAIST 전기 및 전자 공학과 박사

2010년~2013년 : 한국전자 통신연구원, 선임연구원

2013년~2014년 : Imperial College London, Research Associate

2014년~현재 : 부경대학교 부교수

<관심분야> 다중안테나 기반 통신 신호처리, 동시 전력 및 정보 전송 기술, 인지무선기술

[ORCID:0000-0001-5327-9111]



www.dbpia.co.kr