https://doi.org/10.7840/kics.2023.48.12.1549

심층 에너지 기반 은닉변수 모델을 위한 스코어 매칭 학습

곽 봉◆, 김 동 국°

Score Matching Training for Deep Energy-Based Latent Variable Model

Guo Peng◆, Doon Kook Kim°

요 약

본 논문에서는 최근 제안한 심층 에너지 기반 은닉변수 모델(DELVM)에 대한 score matching(SM)을 이용한 학습기법을 제시한다. DELVM의 에너지 함수는 연속적인 입력과 은닉층을 갖으며 심층신경망에 의해 정의된다. 이러한 에너지 함수에 근거하여 SM 기반 DELVM 학습의 목적함수는 Fisher divergence을 통해 유도하고 경사하강법을 이용한 파라미터의 갱신법을 제시한다. Fashion MNIST와 CIFAR10 데이터를 사용한 비지도 특징학습을 위한 이미지 인식 실험에서 제안된 기법은 contrastive divergence 기반 DELVM와 기존 모델보다 향상된 성능을 나타낸다. 또한 ECG와 CIFAR10 데이터를 이용한 이상탐지 실험에서도 기존 모델보다 더 효과적인 F1-score를 나타낸다.

Key Words : energy-based latent variable model, score matching, deep neural networks, gradient descent

ABSTRACT

In this paper, we present a learning technique using score matching (SM) for the recently proposed deep energy-based latent variable model (DELVM). The energy function of DELVM has continuous inputs and hidden layers and are defined by deep neural networks. Based on this energy function, the objective function of SM-based DELVM learning is derived through Fisher divergence and presents a parameter update method using the gradient descent algorithm. In an image recognition experiment for unsupervised feature learning using Fashion MNIST and CIFAR10 data, the proposed technique demonstrates an improved performance than contrastive divergence-based DELVM and existing models. In addition, abnormal detection experiments using ECG and CIFAR10 data also show more effective F1-score over existing models.

Ⅰ. 서 론

최근 에너지 기반 모델 (Energy-based Model, EBM)[1-8]은 기계학습의 중요한 분야로 연구가 활발히 진행되고 있다. EBM은 에너지 함수(energy function)를 통해 확률밀도함수를 정의하기 위해 사용되는 확률적인 모델이다. 에너지 함수로 주로 다양한 형태의 심층신경망을 사용하기 때문에 EBM의 확률모델 형태는 일반적으로 비정규화된(unnormalized) 형태을 갖고 있는 것이 특징이다. 이런 EBM은 여러 가지 형태의 데이터

◆ First Author : Chonnam National University, School of Electronic and Computer Engineering, bongkua0379@gmail.com, 학생회원
° Corresponding Autho : Chonnam National University, Dept. of Electronic Engineering, dkim@jnu.ac.kr, 종신회원
논문번호 : 202308-038-A-RE, Received August 9, 2023; Revised September 6, 2023; Accepted September 11, 2023

를 확률적으로 모델링하는데 있어서 매우 뛰어난 성능을 나타낸다. 이에 따라 EBM은 특징학습, 이상탐지, 이미지 생성, 밀도추정 등과 같은 다양한 응용 분야에 적용되고 있다[1].

많은 EBM 중에서 가장 많이 사용되는 형태는 에너지 기반 은닉변수 모델(Energy-based Latent-Variable Model, EBLVM)[1]-[8]이다. EBLVM은 EBM의 한 종류로 에너지 함수를 입력 데이터와 은닉변수의 결합 확률밀도함수로 정의한다. EBLVM 중에 가장 대표적인 모델은 단층 구조를 갖는 제한된 볼츠만머신(Restricted Boltzmann Machine, RBM)[9], 여러 개의 RBM을 계층적으로 쌓아 구성한 DBN(Deep Belief Network)[10], DBM(Deep Boltzmann Machine)[11], DEM(Deep Energy Model)[7] 그리고 최근에 제안된 CEBM (Conjugate EBM)[6]와 DELVM(Deep Energy-based Latent Variable Model)[8] 등이 있다. RBM[9]은 입력 데이터에 대한 확률분포를 모델링할 수 있는 에너지 기반의 단층 신경망이다. RBM의 입력 형태는 벡터와 행렬이 있으며, 이미지와 같은 행렬 형태일 때 Convolution RBM[12]을 사용한다. DBN[10]은 최상위 계층이 무방향(undirected) 연결이고 나머지 계층이 유향 연결(directed)인 그래프 확률모델이다. DBN의 학습은 RBM을 계층적으로 쌓아서 layer-wise 방식으로 학습한다. DBM[11]은 네트워크의 모든 계층이 무방향으로 연결되어 있고 학습은 layer-wise 방식으로 학습한 후 모든 층을 연합해서 훈련한다. DEM[7]은 여러 개의 deterministic 은닉층과 마지막에 하나의 stochastic 은닉층으로 구성된 심층 모델이다. CEBM[6]은 입력과 은닉변수의 결합밀도함수가 intractable한 데이터 분포와 tractable한 posterior 은닉 분포로 분해되도록 제안되었다. DELVM[8]은 DEM과 동일한 구조와 훈련 방식을 가진 최근에 제안된 모델이다. DEM은 입력층이 연속적 값이 갖고 은닉층이 binary 값이 갖는 것과 달리 DELVM은 입력층과 은닉층 모두 연속적 값을 갖는다. 이런 특성 때문에 DELVM은 DEM보다 다양한 비선형 활성화 함수를 사용할 수 있어 데이터의 확률분포를 모델링하는데 더 큰 장점을 갖고 있다.

EBM의 비정규화된 확률형태 때문에 정확한 유사도 계산이나 모델로부터 데이터를 정확히 생성하는 것이 불가능하여 EBM의 파라메터 학습이 매우 어려운 것으로 알려졌다[1]. EBM을 학습하는데 크게 최대우도비(Maximum－Likelihood, ML)와 스코어 매칭(Score Matching, SM) 기법이 주로 사용된다[2]. ML 기법은 좋은 이론적 성질을 가지고 있으며 확률모델의 파라미터를 추정하는 가장 기본적인 기법이다. 그러나 ML은 EBM에 존재하는 비정규화된 파티션 함수(partition function) 때문에 학습이 어렵다는 단점이 있다. 따라서 이러한 단점을 극복하기 위해 Markov chain monte carlo(MCMC)[15] 기법과 결합된 ML 기법이 사용되고 있다. MCMC을 사용한 ML 기법으로는 contrastive divergence(CD)[2,9-11]가 있으며, 기존의 EBM의 학습은 대부분 이를 기반으로 학습을 수행한다. 최근에 파티션 함수를 계산하지 않고 EBM을 학습할 수 있는 SM 기법이 제안되어 여러 분야에서 우수한 성능을 나타내고 있다[2,5,13,14]. SM의 목적함수는 입력에 대해 데이터 분포와 모델분포 사이에 Fisher divergence(FD)라 불리는 두 분포의 로그 미분값 차이의 유클리드 거리(Euclidean distance)에 의해 정의된다. 이 기법은 EBM 파티션 함수가 학습에 포함되지 않아 ML 기법보다 학습에 더 효율적인 장점을 갖고 있다.

본 논문에서는 SM 기법을 사용하여 최근에 제안된 DELVM을 학습하는 알고리즘을 제안한다. DELVM의 에너지 함수는 연속적인 입력과 은닉층을 갖으며 여러 형태의 심층신경망에 의해 정의된다. 제안된 기법은 DELVM의 에너지 함수로부터 학습을 위한 SM의 목적함수를 유도하고 gradient descent(GD) 알고리즘을 이용하여 파라미터를 갱신한다. 제안된 학습기법의 성능을 평가하기 위해 비지도(unsupervised) 특징학습을 통한 이미지 인식 및 이상탐지 실험을 진행한다. 특징학습을 위해 Fashion-MNIST와 CIFAR10을, 이상탐지를 위해 ECG과 CIFAR10 데이터를 사용하였다. 제안된 기법은 두 가지 실험에서 기존의 CD 기반 DELVM보다 더 높은 성능을 나타내었다. 또한 SM 기반 DELVM이 SM 기법으로 학습된 기존 기법들보다 더 향상된 성능을 나타내었다.

본 논문 본론 Ⅱ장에서는 기존의 DELVM과 SM 기법을 소개하고, SM 기반 DELVM 학습기법을 제시한다. Ⅲ장에서는 실험 및 결과를 나타내고, Ⅳ에서는 결론을 맺는다.

Ⅱ. 본 론

2.1 DELVM

DELVM[2]은 일반적인 EBM 형태 중에 하나이다. 따라서 이 단원에서는 기존의 EBM과 DELVM의 구조를 살펴보고, 이를 학습하기 위한 ML 기법에 대해 간단히 소개한다. 알려지지 않은 데이터 분포 $p_d(\boldsymbol{x})$로부터 발생된 i.i.d 데이터들의 집합, $\{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$가 주어진 경

우, EBM의 목적은 모델분포 $p_\theta(\boldsymbol{x}) = \frac{e^{-E_\theta(\boldsymbol{x})}}{Z(\theta)}$, $Z_\theta = \int \exp\{-E_\theta(\boldsymbol{x})\} d\boldsymbol{x}$ 을 사용하여 실제 데이터의 분포 $p_d(\boldsymbol{x})$를 근사하기 위한 모델분포 파라메터 θ을 구하는 것이다[2]. 여기서 $E_\theta(\boldsymbol{x})$는 에너지 함수이며, Z_θ는 파티션 함수라 한다. EBM을 학습하기 위한 대표적인 방법으로 ML 기법이 있다. ML 기법은 경사하강법을 통해 아래와 같이 $p_d(\boldsymbol{x})$와 $p_\theta(\boldsymbol{x})$ 사이의 Kullback-Leibler divergence(KLD), $D_{KL}(p_d \| p_\theta)$을 최소화하는 방식으로 θ를 구한다[2].

$$\begin{aligned} D_{KL}(p_d \| p_\theta) &= -\mathbf{E}_{\boldsymbol{x} \sim p_d(\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})] \\ &= -\mathbf{E}_{\boldsymbol{x} \sim p_d(\boldsymbol{x})}[E_\theta(\boldsymbol{x})] - \log Z_\theta \end{aligned} \tag{1}$$

위 식에서 두 번째 항인 $\log Z_\theta$의 값은 해석적으로 구할 수 없지만, 이에 대한 경사도(gradient)는 $\nabla_\theta \log Z_\theta = \mathbf{E}_{\boldsymbol{x}' \sim p_\theta(\boldsymbol{x}')}[-\nabla_\theta E_\theta(\boldsymbol{x}')]$ 형태로 구할 수 있다. 따라서 KLD의 경사도는 다음과 같다.

$$\begin{aligned} &\nabla_\theta D_{KL}(p_d \| p_\theta) \\ &= -\mathbf{E}_{\boldsymbol{x} \sim p_d(\boldsymbol{x})}[\nabla_\theta E_\theta(\boldsymbol{x})] + \mathbf{E}_{\boldsymbol{x}' \sim p_\theta(\boldsymbol{x}')}[\nabla_\theta E_\theta(\boldsymbol{x}')] \end{aligned} \tag{2}$$

위 식의 첫 번째 항은 학습 데이터로부터 쉽게 계산이 가능하지만, 두 번째 항의 기댓값은 계산할 수 없다. 따라서 이 계산을 위해 학습된 모델로부터 샘플링된 MCMC 추정치를 사용해 계산하는데 이를 CD 기법이라 한다. 이러한 CD 기법은 모델 분포로부터 얻어지는 샘플값을 통해 근사적으로 KLD의 경사도를 구하게 된다.

최근에 제안된 DELVM[8]은 입력 데이터 $\boldsymbol{x}$와 은닉변수 $\boldsymbol{h}$의 결합확률분포 $p_\theta(\boldsymbol{x},\boldsymbol{h}) = e^{-E_\theta(\boldsymbol{x},\boldsymbol{h})}/Z_\theta$로 정의되는 EBM의 한가지 형태이다. 여기서 $E_\theta(\boldsymbol{x},\boldsymbol{h})$는 $\boldsymbol{x}$와 $\boldsymbol{h}$의 결합 에너지함수로 심층신경망에 의해 정의된다. 이때 입력 데이터 $\boldsymbol{x}$와 stochastic 은닉변수 $\boldsymbol{h}$ 모두 연속적인 값을 갖는다고 가정한다. 이런 가정하에서 DELVM의 에너지 함수는 다음과 같다[8].

$$E_\theta(\boldsymbol{x},\boldsymbol{h}) = \frac{1}{2}\left(\boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{h}^T\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{h}^T\boldsymbol{h}\right) \tag{3}$$

여기서 $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$는 입력층에서 은닉층으로 변환하는 파라메터 θ을 갖는 심층신경망으로, 입력 형태에 따라 FFNN(Feed-Forward Neural Networks) 또는 CNN(Convolutional Neural Network)을 나타낸다. 위의 에너지 함수하에서 $\boldsymbol{x}$가 주어진 경우 $\boldsymbol{h}$에 대한 조건부 확률분포는 다변수(multivariate) Gaussian 형태, $p_\theta(\boldsymbol{h}|\boldsymbol{x}) = N(\boldsymbol{h}; \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{I})$를 갖는다. 여기서 I는 단위행렬을 나타낸다. 주목할 점은 신경망 $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$는 입력 데이터 $\boldsymbol{x}$가 주어졌을 때 은닉변수 $\boldsymbol{h}$의 평균 벡터를 나타낸다는 것이다. 위의 결합확률분포를 기반으로 $\boldsymbol{x}$의 한계확률분포와 에너지 함수는 아래와 같이 구할 수 있다[8].

$$p(\boldsymbol{x}) = \frac{\exp\{-E_\theta(\boldsymbol{x})\}}{Z_\theta} \tag{4}$$

$$E_\theta(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x}^T\boldsymbol{x} - \boldsymbol{f}_{\boldsymbol{\theta}}^T(\boldsymbol{x})\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) = \frac{1}{2}(\|\boldsymbol{x}\|^2 - \|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})\|^2) \tag{5}$$

여기서 $E_\theta(\boldsymbol{x})$는 $\boldsymbol{x}$와 $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$로만 구성된 $\boldsymbol{x}$에 대한 에너지 함수이다.

DELVM 학습방법은 앞에서 언급한 KLD을 최소화하기 위한 CD 기법을 사용한다. DELVM에 대한 CD 기법은 아래와 같은 모델 분포로부터 MCMC 샘플링을 통해 얻어진 샘플값을 통해 목적함수의 경사도가 구해진다.

$$\begin{aligned} \nabla_\theta D_{KL}(p_d \| p_\theta) = &-\mathbf{E}_{\boldsymbol{x} \sim p_d(\boldsymbol{x}), \boldsymbol{h} \sim p_\theta(\boldsymbol{h}|\boldsymbol{x})}[\nabla_\theta E_\theta(\boldsymbol{x},\boldsymbol{h})] \\ &+ \mathbf{E}_{\boldsymbol{x}' \sim p_\theta(\boldsymbol{x}'), \boldsymbol{h}' \sim p_\theta(\boldsymbol{h}'|\boldsymbol{x}')}[\nabla_\theta E_\theta(\boldsymbol{x}',\boldsymbol{h}')] \end{aligned} \tag{6}$$

위 식의 두 번째 항의 계산을 위해 모델로부터 새로운 샘플이 필요하게 되는데, 간단한 과정을 통해 샘플은 $\boldsymbol{x}' = \sum_i h_i \nabla_{\boldsymbol{x}} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})_i$와 같이 얻을 수 있다[8]. 여기서 $\boldsymbol{x}'$는 주어진 $\boldsymbol{h}$로부터의 샘플이며, $\nabla_{\boldsymbol{x}} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})_i$는 심층신경망 $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$의 i번째 은닉변수 출력에 대한 경사도이다. 따라서 CD 기반 DELVM의 학습은 식 (6)을 사용한 경사하강법을 통해 반복적으로 갱신하게 된다. 위와 같은 방법은 학습 과정 중에 반복적으로 샘플링 과정이 필요하여 많은 계산량을 요구하는 단점을 갖는다.

2.2 SM 기반 DELVM 학습

이 단원에서는 SM 기법을 소개하고, 이를 사용한 새로운 SM 기반 DELVM 학습기법을 제시한다. SM 학습기법은 FD라 불리는 두 분포 사이의 목적함수를 최소화하는 방식이다. 즉 연속 값을 갖는 데이터 공간에서 데이터 분포 $p_d(\boldsymbol{x})$와 모델분포 $p_\theta(\boldsymbol{x})$의 로그 미분값 사이의 유클리드 거리를 다음과 같이 최소화하는 것이다.

$$D_F(p_d \parallel p_\theta) = \mathbf{E}_{p_d(\boldsymbol{x})}\left[\frac{1}{2} \parallel \nabla_{\boldsymbol{x}} \log p_d(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log p_\theta(\boldsymbol{x}) \parallel_2^2\right] \tag{7}$$

여기서 로그-분포의 일차 미분값을 score 함수라 한다. 따라서 SM 학습 방식은 데이터의 score 함수와 분포의 score 함수가 일치하도록 학습하는 방식이다. 여기서 주목할 점은 모델분포가 EBM 형태로 주어진 경우, score 함수는 계산하기 힘든 파티션 함수 Z_θ 을 포함하지 않아 쉽게 훈련이 가능하다는 것이다. 식 (7)을 최적화 하는데 $p_d(\boldsymbol{x})$에 대한 의존성 때문에 위 목적함수를 계산하기가 쉽지 않다. 그러나 어떤 정규적 조건하에서 아래와 같은 식으로 유도할 수 있다[2,13,14].

$$\begin{aligned} D_F(p_d \| p_\theta) &= \mathbf{E}_{p_d(\boldsymbol{x})}\left[\frac{1}{2} \parallel \nabla_{\boldsymbol{x}} E_\theta(\boldsymbol{x}) \parallel_2^2 + tr(\nabla_{\boldsymbol{x}}^2 E_\theta(\boldsymbol{x}))\right] \\ &= \mathbf{E}_{p_d(\boldsymbol{x})}\left[\frac{1}{2}\sum_{i=1}^{d}\left(\frac{\partial E_\theta(\boldsymbol{x})}{\partial x_i}\right)^2 + \frac{\partial^2 E_\theta(\boldsymbol{x})}{\partial^2 x_i}\right] \end{aligned} \tag{8}$$

여기서 d는 입력벡터의 차원이며, 오른쪽 두 번째 항, $tr()$은 $\boldsymbol{x}$에 대한 2차 도함수(Hessian)의 trace을 나타낸다. SM의 단점은 2차 도함수에 대한 계산량이 많이 들기 때문에 2차 도함수 계산이 가능한 간단한 에너지 함수에 적용되고 있다. 또한 이러한 단점을 극복하기 위한 denoising SM기법이 제안되었다[2].

본 논문에서는 SM 목적함수를 DELVM에 적용하여 학습하는 새로운 기법을 제안한다. 먼저 DELVM의 특수한 경우인 $\boldsymbol{f_\theta}(\boldsymbol{x})$가 하나의 은닉층만을 갖는 신경망에 대한 SM 학습기법을 유도하고, 그다음으로 심층신경망 $\boldsymbol{f_\theta}(\boldsymbol{x})$로 일반화하는 식을 제시한다. $\boldsymbol{f_\theta}(\boldsymbol{x})$의 입력은 $\boldsymbol{x} = (x_1,...,x_d)^T$로 d차원 데이터공간에서 정의된 벡터이고 하나의 은닉층이 m개의 유닛을 갖는다고 가정한다. 그러면 입력층과 은닉층 사이는 m개의 d차원의 가중치 벡터와 바이어스, $(\boldsymbol{w}_j, b_j)_{j=1,...,m}$로 구성된다. 그리고 은닉층에서 ReLU(Rectified Linear Unit)[16] 활성화 함수를 갖는 $\boldsymbol{f}_\theta(\boldsymbol{x})$라 가정한다. 식 (5)의 DELVM의 $\boldsymbol{x}$에 대한 에너지 함수는 다시 쓰면 아래와 같다.

$$E_\theta(\boldsymbol{x}) = \frac{1}{2}\left(\|\boldsymbol{x}\|^2 - \|\boldsymbol{f}(\boldsymbol{x})\|^2\right) = \frac{1}{2}\left(\sum_{i=1}^{d} x_i^2 - \sum_{j=1}^{m} \hat{h}_j^2\right) \tag{9}$$

여기서 $\hat{h}_j = f(\boldsymbol{w}_j^T\boldsymbol{x} + b_j) = f(\sum_{i=1}^{d} w_{ji}x_i + b_j)$은 j번째 은닉층의 값을 나타내며, $f()$는 ReLU 함수를 나타낸다. 이런 가정하에 $E_\theta(\boldsymbol{x})$을 식 (8)에 대입하면 SM 목적함수를 얻을 수 있다. 이때 $E_\theta(\boldsymbol{x})$에 대한 x_i의 1차 미분값과 2차 미분값은 각각 아래와 같다.

$$\frac{\partial E_\theta(\boldsymbol{x})}{\partial x_i} = x_i - \sum_{j=1}^{m}(\hat{h}_j \nabla f_j) w_{ji} \tag{10}$$

$$\frac{\partial^2 E_\theta(\boldsymbol{x})}{\partial^2 x_i} = 1 - \sum_{j=1}^{m}(\nabla f_j)^2 w_{ji}^2 \tag{11}$$

여기서 ∇f_j는 j번째 함수 f_j에 대한 도함수를 의미하며, w_{ji}는 가중치 벡터 $\boldsymbol{w}_j$의 i번째 성분을 나타낸다. 위 두 식을 식 (8)에 대입하여 정리하면 하나의 ReLU 은닉층 갖는 DELVM에 대한 SM 목적함수는 다음식과 같다.

$$\begin{aligned} J(\theta) = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{d}\Bigg[&\frac{1}{2}\left(x_i^{(n)} - \sum_{j=1}^{m}(\hat{h}_j^{(n)} \nabla f_j^{(n)}) w_{ji}\right)^2 \\ &+ \left(1 - \sum_{j=1}^{m}(\nabla f_j^{(n)})^2 w_{ji}^2\right)\Bigg] \end{aligned} \tag{12}$$

여기서 $\hat{h}_j^{(n)}$와 $\nabla f_j^{(n)}$은 n번째 입력 $\boldsymbol{x}^{(n)}$이 주어진 경우, 은닉층 j번째 성분과 그때의 도함수를 각각 나타낸다. 위 식의 첫번째 항은 입력과 재생된 입력 사이의 재생오차(reconstruction error, RE)의 제곱항에 해당되며, 두 번째 항은 신경망에 대한 가중치에 대한 규제화 항으로 해석할 수 있다.

이제 DELVM이 일반적인 심층신경망 $\boldsymbol{f_\theta}(\boldsymbol{x})$을 갖는 경우는 위와 비슷한 과정을 통해 SM 학습에 필요한 목적함수를 유도할 수 있다. 심층신경망 $\boldsymbol{f_\theta}(\boldsymbol{x})$을 갖는 식 (9)에 대한 1차 미분값과 2차 미분값의 trace는 아래와 같다.

$$\nabla_{\boldsymbol{x}} E_\theta(\boldsymbol{x}) = \boldsymbol{x} - \nabla_{\boldsymbol{x}}^T \boldsymbol{f}_\theta(\boldsymbol{x}) \boldsymbol{f}_\theta(\boldsymbol{x}) \tag{13}$$

$$tr(\nabla_{\boldsymbol{x}}^2 E_\theta(\boldsymbol{x})) = tr\left(I_d - \nabla_{\boldsymbol{x}}^2 \|\boldsymbol{f}_\theta(\boldsymbol{x})\|^2\right) \tag{14}$$

여기서 $\nabla_{\boldsymbol{x}} \boldsymbol{f}_\theta(\boldsymbol{x})$은 $m \times d$ Jacobian 행렬이며, I_d는 $d \times d$ 단위 행렬 그리고 $tr(\nabla_{\boldsymbol{x}}^2 \|\boldsymbol{f}_\theta(\boldsymbol{x})\|^2) = \sum_{i=1}^{d} \frac{\partial^2 \|\boldsymbol{f}_\theta(\boldsymbol{x})\|^2}{(\partial x_i)^2}$을 나타낸다. 위 식에 근거하여 일반적인 심층신경망을 갖는 DELVM에 대한 SM 목적함수는 다음식과 같다.

$$J(\theta) = \boldsymbol{E}_{p_{data}(\boldsymbol{x})}\left[\frac{1}{2}\| \boldsymbol{x} - \nabla_{\boldsymbol{x}}^{T}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})\|^{2} + tr\left(I_d - \nabla_{\boldsymbol{x}}^{2}\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})\|^2\right)\right] \quad (15)$$

위 식에서 일반적인 심층신경망에 대한 Jacobian과 Hessian에 대한 계산은 TensorFlow 또는 PyTorch와 같은 딥러닝 소프트에서 자동 미분 패키지를 사용함으로써 쉽게 가능하다. 식 (15)와 같은 DELVM에 대한 SM 목적함수가 정의되면, 심층신경망의 최적의 파라메터는 stochastic GD(SGD) 알고리즘을 사용하여 반복적으로 파라메터를 갱신함으로 추정할 수 있다.

Ⅲ. 실 험

SM 기법을 이용한 DELVM의 학습 성능을 평가하기 위해 비지도 특징학습 기반 이미지 인식과 데이터 이상탐지 실험을 수행하여 성능을 비교한다.

3.1 비지도 특징학습

이 장에서는 비지도 특징학습을 통한 이미지 인식을 위해 Fashion MNIST[17]와 CIFAR10[18] 데이터 셋을 사용하였다.

3.1.1 Fashion MNIST

Fashion MNIST 데이터 셋은 패션 상품에 대한 10개 클래스로, 28×28 gray 이미지로 구성되어 있다. 학습 데이터는 60,000개, 테스트 데이터는 10,000개가 사용되었다. Fashion MNIST를 이용하여 CD와 SM 학습기법을 새로운 DELVM과 기존의 DEM에 적용하여 파라메터를 훈련하였고, 이로부터 추출된 특징들을 이용하여 인식실험을 진행하였다. 여기서 입력은 이미지 형태의 행렬이기 때문에 $\boldsymbol{f}_{\theta}(\boldsymbol{x})$을 위해 총 4층의 CNN 구조를 사용하였다. 학습을 위한 batch size는 100이며, epoch은 CD은 100, SM은 10을 사용하였다. 학습률은 각 층에 따라 1e-3에서 1e-5 사이의 값을 사용하였다. 각 모델의 구현은 TensorFlow로 작성하였고, Adam optimizer을 사용하였다. 각 모델을 학습한 후 CNN으로부터 각 층의 feature map을 추출하고 크기가 2인 필터를 사용하여 max pooling 과정을 수행하였다. 그리고 각 층마다 추출한 새로운 특징들을 합쳐서 sofatmax 분류기를 통해 학습 및 인식 실험을 진행하였다. 더 상세한 성능 비교를 위해 최근 제안된 CEBM을 이용한 실험을 수행하였다. CEBM을 위한 구조는 [6]과 같이 사용하였다. 합성곱 층은 총 4층이고, 필터의 수와 크기는 각각 (64, 64, 32, 32) (3, 4, 4, 4)이다.

CNN을 학습할 때 필터 수와 크기가 성능에 미치는 영향을 알아보기 위해 단층 구조에 대해 각각 10-80개의 필터 수와 (3,5,7,9)의 필터 크기를 사용하였다. 여러 층에 대한 필터 수와 크기에 대한 매개변수 설정이 어렵기 때문에 단층 성능이 가장 높을 때의 필터 수와 크기를 선택하여 모든 층에 사용하였다. 그림 1은 SM와 CD 기법을 DELVM의 학습에 적용할 때 필터 수와 크기에 따른 인식실험 결과이다. SM인 경우 필터 사이즈가 5와 7일 때, CD은 7와 9일 때 성능이 더 좋았다. SM과 CD의 최고의 성능은 필터 수와 크기가 (70, 5,) 와 (80, 9)일 때로 나타났다. 그림 1에서 DELVM을 학습할 때 SM이 CD보다 성능이 같은 필터 수와 크기에 대해 더 높게 나타났다. 그림 2은 SM 기법을 DELVM와 DEM의 학습에 적용할 때 필터 수와 크기에 따른 인식실험 결과이다. DELVM의 경우 필터 수가 증가함에 따라 성능이 향상되는 것에 비해 DEM은 필터 수 증가에 따라 성능이 저하되었다. DEM의 경우 필터 수와 크기가 (20, 5)일 때 가장 높은 성능을 나타내었다. 그림 2에서 제안된 SM 기반 DELVM이 기존의

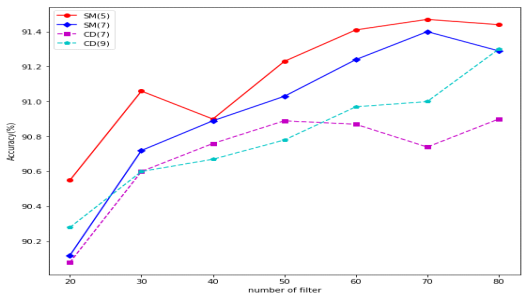

그림 1. Fashion-MNIST에서 SM와 CD 기반 DELVM의 필터 수와 크기에 따라 인식 정확도.
Fig. 1. The recognition accuracy (%) of SM and CD-based DELVMs in Fashion-MNIST according to the number and size of filters.

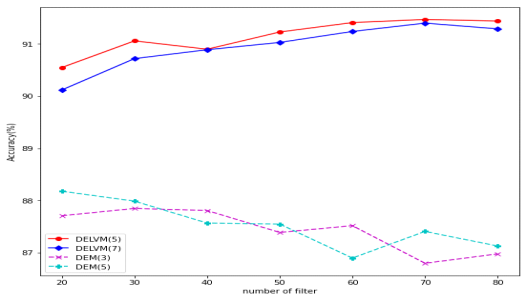

그림 2. Fashion-MNIST에서 SM 기반 DELVM와 DEM의 필터 수와 크기에 따라 인식 정확도.
Fig. 2. The recognition accuracy (%) of SM-based DELVM and DEM in Fashion-MNIST according to the number and size of filters.

DEM보다 더 높은 성능을 나타냄을 알 수 있다.

그림 3은 SM와 CD 기반 DELVM와 DEM 그리고 CEBM의 은닉층 수에 따른 인식 결과를 나타낸다. SM 기법을 이용한 두 모델을 학습하는 경우, 2개의 은닉층일 때 가장 좋은 성능을 나타내었고, 3과 4층 경우에 약간의 성능이 하락하였다. 이는 은닉층이 많아질수록 파라메터수가 증가하고, 따라서 모델에서 과적합(overfitting) 문제가 발생하여 성능의 저하가 발생하는 것으로 해석할 수 있다. CEBM은 모든 층에 대해 비슷한 성능을 나타내며, DEM에 비해 성능이 좋지만, SM와 CD 기반의 DELVM보다 낮은 결과를 나타내었다.

그림 1-3의 결과를 통해 Fashion MNIST 데이터 셋에 대해 제안된 SM 기반 DELVM 학습기법이 특징학습에서 CD 학습기법 또는 기존의 DEM, CEBM과 비교하여 더 나은 성능을 나타내었다.

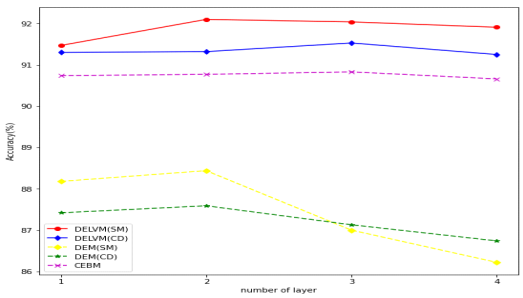

그림 3. Fashion-MNIST에서 SM과 CD 기반 DELVM와 DEM 그리고 CEBM의 은닉층 수에 따라 인식 정확도 (%).
Fig. 3. Recognition accuracy (%) of DELVM and DEM-based on SM and CD, and CEBM according to the number of hidden layers in Fashion-MNIST.

3.1.2 CIFAR10

CIFAR10 데이터 셋은 32x32 크기의 color 이미지이며, 10개의 클래스를 포함하고 있다. 학습 시 50,000개, 테스트 시에 10,000개의 이미지가 각각 사용되었다. 데이터 전처리 과정으로 standard normalization을 수행하였다. CIFAR10 인식실험은 Fashion MNIST와 동일하게 진행되었다. 그림 4은 CIFAR10에 대해 SM와 CD 기법을 단층의 DELVM의 학습에 적용할 때 필터 수와 크기에 따른 특징학습에 따른 인식실험 결과이다. SM을 사용하는 경우 필터 크기가 3과 5일 때 성능이 더 좋았으며, CD의 경우 7와 9일 때 더 높은 성능을 나타내었다. 두 경우 모두 필터 수가 증가함에 따라 성능이 향상되었다. 최고의 성능은 필터의 수와 크기가 SM 경우 (80, 3)일 때, CD의 경우 (60, 7)일 때 나타났다. SM와 CD의 학습기법을 비교할 때 단층의 DELVM의 경우 거의 비슷한 인식성능을 나타내었다.

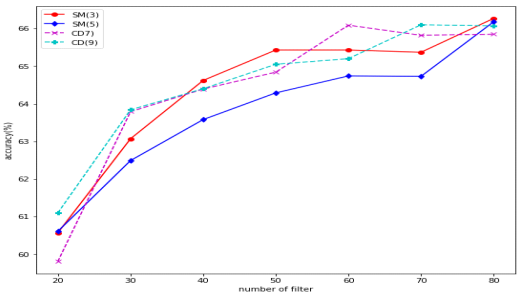

그림 4. CIFAR10에서 SM와 CD 기반 DELVM의 필터 수와 크기에 따라 인식 정확도 (%).
Fig. 4. The recognition accuracy (%) of SM and CD-based DELVMs in Fashion-MNIST according to the number and size of filters.

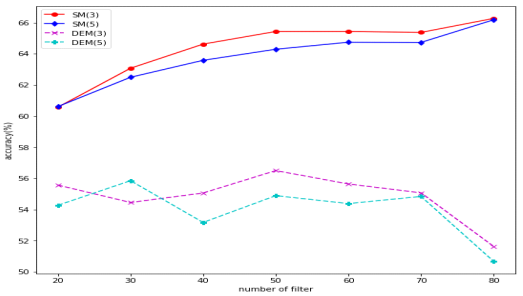

그림. 5. CIFAR10에서 SM 기반 DELVM와 DEM의 필터 수와 크기에 따라 인식 정확도 (%).
Fig. 5. The recognition accuracy (%) of SM-based DELVM and DEM in Fashion-MNIST according to the number and size of filters.

그림 5은 SM 기법을 DELVM와 DEM의 학습에 적용할 때 필터 수와 크기에 따른 인식실험 결과이다. 두 모델 모두 필터 크기는 3와 5가 7과 9보다 더 높은 성능을 나타내었다. 모든 필터 수에 대해 제안된 SM 학습하에서 DELVM이 DEM에 비해 더 높은 인식성능을 나타내었다.

그림 6은 SM와 CD 기반 DELVM와 DEM 그리고 CEBM의 은닉층 수에 따른 인식 결과를 나타낸다. CEBM은 Fashion-MNIST와 같은 구조를 사용하였다. 단층의 DELVM의 경우 SM과 CD는 비슷한 성능을 나타내지만, 2-4층의 경우 SM이 CD보다 약간 더 향상된 인식성능을 나타내었다. 또한 DELVM의 두 학습기법이 기존의 DEM보다 훨씬 더 높은 인식결과를 나타내었다[1]. CEBM은 Fashion MNIST와 같이 DEM보다 좋지만, SM와 CD 기반 DELVM에 비해 낮은 성능을

1) 최근 [19]논문에서는 비지도 clustering을 위한 robust learning을 사용하여 CIFAR10의 경우 특징학습 인식성능 90.3%로 본 논문보다 훨씬 뛰어난 성능을 보이고 있다.

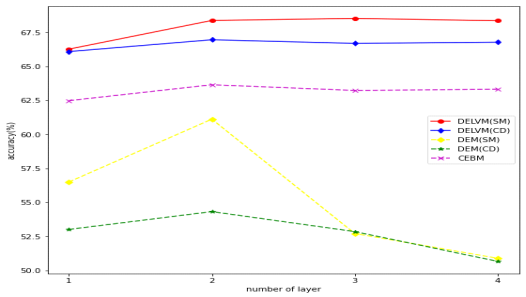

그림 6. CIFAR10에서 SM과 CD 기반 DELVM와 DEM 그리고 CEBM의 은닉층 수에 따라 인식 정확도 (%).
Fig. 6. Recognition accuracy (%) of DELVM and DEM-based on SM and CD, and CEBM according to the number of hidden layers in CIFAR10.

나타내었다. 따라서 CIFAR10와 같은 color 이미지에 대해서도 제안된 SM 기반 DELVM이 비지도 특징학습에 있어서 다른 에너지 모델과 비교해서 매우 효과적임을 알 수 있다.

3.2 이상탐지(Anomaly Detection)

이 장에서는 제안된 알고리즘의 성능을 이상탐지 실험에 대해 평가하기 위해 ECG[20]와 CIFAR10 데이터셋을 사용하였다.

3.2.1 ECG

ECG5000[20]은 Physionet에서 다운로드한 20시간 길이의 심전도 데이터 셋이다. 데이터는 각 하트비트를 추출하고 보간법을 사용하여 각 하트비트의 길이를 동일하게 만드는 두 단계로 전처리되었다. 전체 데이터는 5,000개의 심전도로, 각 심전도는 140개의 데이터 포인트를 포함한다. 데이터는 학습전에 [0,1] 사이로 minmax normalization을 수행하였다. 각 심전도는 1 (이상 리듬) 또는 0(정상 리듬)로 표시된다. 이때 학습에 사용되는 정상 데이터는 1639개, 테스트는 1000개를 사용하였다. 제안한 SM 기반 DELVM의 성능을 알아보기 위해 CD 기반 DELVM와 이상탐지 분야에서 가장 많이 사용되는 AE(AutoEncoder)와 CEBM을 비교하였다. CEBM 은닉층의 뉴런수는 2048로 사용하였다. Epoch은 CD DELVM과 AE에 대해 100을, CEBM은 50을, SM DELVM은 10을 사용하였다. DELVM을 위한 학습률은 각 층에 따라 1e-2 또는 1e-3을 사용하였다. 모두 모델에 대해 batch size는 100이고 Adam optimizer을 사용하였다.

본 실험에서는 제안된 모델의 성능을 평가하기 위해 준지도(semi-supervised) 학습 기반 ECG의 이상탐지 실험을 수행하였다. 즉 정상 데이터만을 사용하여 모델을 훈련하고 정상/비정상 데이터를 사용하여 RE을 임계값과 비교하여 테스트를 수행하였다. 이러한 RE가 임계값보다 크면 비정상, 작으면 정상으로 분류하였다. 이때 임계값은 정상 훈련 데이터의 RE의 평균과 표준편차를 더한 값으로 설정하였다. 탐지성능을 평가하기 위해 precision, recall, F1-score 3가지 지표를 선택하였고, 최종적으로 F1-score에 따라 성능을 비교하였다. 입력이 벡터이기 때문에 심층신경망으로 FFNN을 사용하였다. 모든 모델에 대해 은닉층의 뉴런 수에 따른 성능을 살펴보기 위해 다양한 유닛 수에 대한 성능 실험을 수행하였다. 심층신경망의 각 은닉층의 유닛 수는 단층의 성능이 가장 높을 때의 수를 선택하여 구성하였다.

그림 7은 단층 구조를 갖는 각 모델에 은닉층의 유닛 수에 따른 F1-score를 보여주고 있다. 그림 7에 나타나듯이 SM 기반 DELVM이 다른 모델에 비해 모든 은닉층의 유닛 수에 대해 더 높은 F1-score를 보이고 있다. 이 때 SM DELVM, CD DELVM 그리고 AE에 대해 최고 성능의 뉴런 수는 각각 2048, 512, 256로 나타났다. 그리고 이때는 precision, recall는 각 모델에 대해

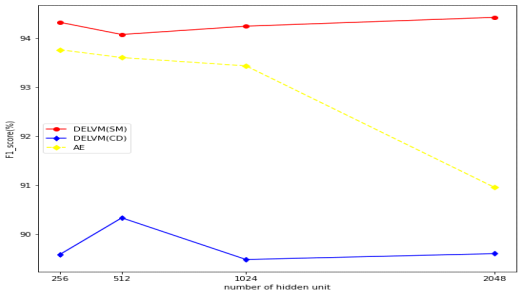

그림 7. ECG에서 단층 구조를 갖는 3가지 모델에 대해 은닉층 유닛 수에 따른 F1 score (%)
Fig. 7. The F1-score (%) of three models with a single-layer in ECG according to the number of hidden unit.

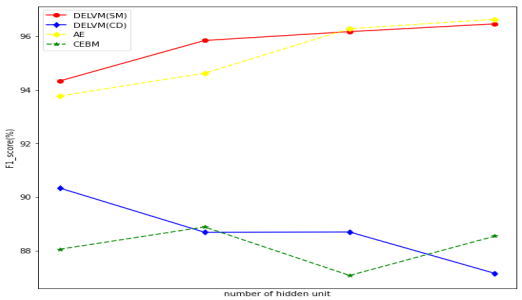

그림 8. ECG에서 은닉층 수에 따른 4가지 모델에 대한 F1 score (%).
Fig. 8. The F1-score (%) of four models in ECG according to the number of hidden layer.

(90.57%, 99.46%), (90.5%, 90.18%), (96.24%, 91.43%)로 나타났다. 심층신경망의 은닉층 수에 따른 각 모델의 결과는 그림 8에 제시되었다. SM DELVM은 4개의 층일 때, F1-score는 96.46%로 최고성능을 보였다. 다른 모델과 비교하였을 때, 제안 모델은 CD DELVM과 CEBM보다 성능이 훨씬 좋았고, AE와는 유사한 성능을 보였다. 따라서 심층 구조도 제안된 기법이 ECG 이상탐지에 대해 다른 기법에 비교하여 매우 효과적임을 나타내었다.

3.2.2 CIFAR10

CIFAR10에 대한 이상탐지 실험을 수행하기 위해 0-9번 클래스 중에서 클래스 0번 만을 정상으로 '0'으로 라벨링하였으며, 나머지 중에서 무작위로 30%를 비정상으로 '1'로 라벨링하였다. ECG와 마찬가지로 준지도 학습기반 이상탐지를 위해 정상 학습 데이터는 4223개를, 테스트 시에는 정상은 1777개와 비정상 4883개를 사용하였다. 학습하기 전에 standard normalization을 수행하여 데이터를 전처리하였다. 성능 비교를 위해 SM와 CD 기반 DELVM, Convoutional AE (CAE) 그리고 CEBM을 사용하였다. 입력이 이미지이므로 모두 CNN 구조를 갖도록 구성하였다. 학습에 필요한 epoch, batch size, 학습률은 특징학습을 위한 CIFAR10과 비슷하게 설정하였다. 정상 데이터만을 사용하여 모델을 훈련한 후 전체 테스트 데이터를 사용하여 평가하였다. 이상탐지를 위해 RE을 임계값과 비교하여 크면 비정상, 작으면 정상으로 판별하였다. 여기에서 임계값은 정상/비정상 훈련 데이터의 RE의 히스토그램에 근거하여 RE의 평균을 중심으로 최적의 값을 선택하였다. CIFAR10의 특징학습와 유사하게, 먼저 단일 구조를 사용하여 가장 좋은 성능을 갖는 필터의 수와 크기를 선택하였다.

표 1은 단층 구조를 갖는 최적의 필터의 수와 크기에 대해 3가지 모델에 대한 탐지성능을 나타낸다.

표 1. CIFAR10에서 단층 구조의 각 모델에 대해 최적의 필터 수와 크기에 따라 성능 (%)
Table 1. The performance (%) of the single-layer models in CIFAR10 according to the optimal number and size of filters.

모델	DELVM		CAE
	SM	CD	
Precision	74.39	74.44	74.15
Recall	99.67	99.52	98.91
F1-score	**85.23**	**85.17**	**84.76**
최적의 필터 (수, 크기)	(70, 7)	(70, 3)	(50, 5)

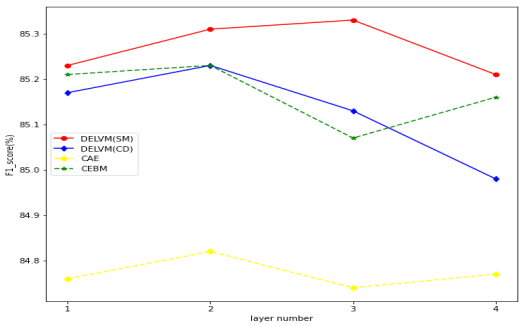

그림 9. CIFAR10에서 은닉층 수에 따른 4가지 모델에 대한 F1 score (%).
Fig. 9. The F1-score (%) of four models in CIFAR10 according to the number of hidden layer.

DELVM가 CAE에 비해 더 높은 F1-score를 가지며, SM DELVM이 CD보다 약간 향상된 성능을 나타내었다. 각 모델에 대한 최적의 필터 수와 크기는 각 모델에 따라 다르게 나타났다.

그림 9는 은닉층 수에 따라 4가지 모델에 대한 F1-score를 나타낸다. DELVM이 CAE에 비해 모든 층에 대해 더 높은 성능을 나타내며, SM이 CD보다 더 우수한 성능을 보였다. SM DELVM은 3층일 때 최고의 성능을 나타냈으며, CD 기법은 2층일 때 가장 좋은 결과를 나타내었다. CEBM이 CD DELVM보다 비슷하거나 조금 더 나은 성능을 나타내었다. 그림에서 보듯이 심층구조를 갖는 제안된 SM 기반 DELVM이 CIFAR10 이미지의 이상 탐지 분야에 매우 효과적임을 알 수 있다.

Ⅳ. 결 론

본 논문에서는 EBM중에 하나인 DELVM의 파라메터를 추정하기 위해 SM을 이용한 학습 기법을 제안하였다. DELVM의 확률분포의 에너지 함수를 이용하여 SM 기반의 목적함수를 유도하였고 이를 경사하강법을 통해 파라메터를 학습하였다. 제안된 SM 기반 DELVEM의 학습의 성능을 위해 특징학습을 통한 이미지인식와 이상탐지 실험을 진행하였다. Fashion MNIST과 CIFAR10 데이터을 이용한 특징학습에 있어서 제안된 SM기반 DELVM이 심층 구조하에서 기존의 CD 기반 DELVM와 DEM 그리고 최근에 제안한 CEBM보다 더 향상된 성능을 나타내었다. 또한 ECG와 CIFAR10을 이용한 이상탐지에서도 기존의 학습 기법, AE 그리고 CEBM에 비해 비슷하거나 더 뛰어난 F1-score을 보여주었다. 향후 연구는 SM 기반

DELVM을 이용하여 이미지 생성과 같은 새로운 분야에 적용하는 것이 필요하다.

References

[1] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. “A tutorial on energy-based learning,” *Predicting structured data,* vol. 1, 2006.

[2] Y. Song and D. P. Kingma, “How to train your energy-based models,” *arXiv preprint arXiv:2101.03288,*2021. (https://doi.org/10.48550/arXiv.2101.03288)

[3] F. K. Gustafsson, M. Danelljan, R. Timofte, and T. B. Schön, “How to train your energy-based model for regression,” *arXiv preprint arXiv:2005.01698,* 2020. (https://doi.org/10.48550/arXiv.2005.01698)

[4] M. M. Al Rahhal, Y. Bazi, R. Al-Dayil, B. M. Alwadei, N. Ammour, and N. Alajlan, “Energy-based learning for open-set classification in remote sensing imagery,” *IJRS*, vol. 43, no. 15-16, pp. 6027-6037, 2022. (https://doi.org/10.1080/01431161.2022.2044539)

[5] K. Swersky, M. A. Ranzato, D. Buchman, N. D. Freitas, and B. M. Marlin, “On autoencoders and score matching for energy based models,” in *Proc. 28th ICML-11*, pp. 1201-1208, 2011.

[6] H. Wu, B. Esmaeili, M. Wick, J. B. Tristan, and J. W. Van De Meent, “Conjugate energy-based models,” *ICML*, pp. 11228-11239, 2021. (https://doi.org/10.48550/arXiv.2106.13798)

[7] J. Ngiam, Z. Chen, P. W. Koh, and A. Y. Ng, “Learning deep energy models,” in *Proc. 28th ICML*, pp. 1105-1112, 2011.

[8] G. Peng and D. Kim, “A new energy-based latent-variable model for unsupervised feature learning,” *J. KICS*, vol. 48, no. 05, 2023. (https://doi.org/10.7840/kics.2023.48.5.509)

[9] G. E. Hinton, “A practical guide to training restricted Boltzmann machines,” *Tech. Rep. UTML TR2010-003*, University of Toronto, 2010. (https://doi.org/10.1007/978-3-642-35289-8_32)

[10] G. E. Hinton, S. Osindero, and Y. W. Teh, “A fast learning algorithm for deep belief nets,” *Neural computation*, vol. 18, no. 7, pp. 1527-1554, 2006. (https://doi.org/10.1162/neco.2006.18.7.1527)

[11] R. Salakhutdinov and H. Larochelle, “Efficient learning of deep Boltzmann machines,” in *Proc. Thirteenth ICARS*, PMLR 9, pp. 693-700, 2010.

[12] M. Norouzi, M. Ranjbar, and G. Mori, “Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning,” *IEEE CVPR*, pp. 2735-2742, 2009. (https://doi.org/10.1109/CVPR.2009.5206577)

[13] Y. Song and S. Ermon, “Generative modeling by estimating gradients of the data distribution,” *Advances in NIPS*, 32, 2019. (https://doi.org/10.48550/arXiv.1907.05600)

[14] D. P. Kingma, “Improving score matching for learning statistical models of natural images,” PhD. dissertation, New York University, 2020.

[15] D. J. MacKay and D. J. Mac Kay, “*Information theory, inference and learning algorithms*,” Cambridge university press, 2003.

[16] V. Nair and G. E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” in *Proc. 27th ICML,* pp. 807-814, 2010.

[17] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms,” *arXiv preprint arXiv:1708.07747*, 2017. (https://doi.org/10.48550/arXiv.1708.07747)

[18] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” 2009.

[19] S. Park, et al., “Improving unsupervised image clustering with robust learning,” *arXiv preprint arXiv:2012.11150*, 2021. (https://doi.org/10.48550/arXiv.2012.11150)

[20] S. K. Berkaya, A. K. Uysal, E. S. Gunal, S. Ergin, S. Gunal, & M. B. Gulmezoglu, “A survey on ECG analysis,” *Biomed. Signal Process. and Control*, vol. 43, pp. 216-235,

2018.
(https://doi.org/10.1016/j.bspc.2018.03.003)

곽 봉 (Guo Peng)

2014년 9월 : 전남대학교 전자공학과 학사
2018년 9월 : 전남대학교 전자공학과 석사
2021년 3월~현재 : 전남대학교 전자 공학과 박사과정

<관심분야> 영상처리, 기계학습, 딥러닝
[ORCID:0009-0009-0123-9957]

김 동 국 (Dong Kook Kim)

1989년 2월 : 전남대학교 전자공학과 학사
1991년 2월 : 포항공과대학 전자전기공학과 석사
2003년 2월 : 서울대학교 전기컴퓨터공학부 박사
1991년 2월~1999년 2월 : 삼성전자 전문연구원
2003년 4월~2004년 2월 : 한국전자통신연구원 선임연구원
2004년 2월~현재 : 전남대학교 전자공학과 교수
<관심분야> 딥러닝, 기계학습, 인공지능신호처리
[ORCID:0000-0001-9316-7069]