DBpia

논문 19-44-09-01 The Journal of Korean Institute of Communications and Information Sciences '19-09 Vol.44 No.09
https://doi.org/10.7840/kics.2019.44.9.1633

비지도 특징학습을 위한 커널 기반 제한된 볼츠만 머신

김 동 국◆, 신 종 원°

Kernel-Based Restricted Boltzmann Machine for Unsupervised Feature Learning

Dong Kook Kim◆, Jong Won Shin°

요 약

본 논문은 비지도 특징학습을 위한 커널 기법에 근거한 제한된 볼츠만 머신 (RBM)을 제시한다. 커널 기반 RBM의 핵심 아이디어는 입력 데이터를 비선형 함수를 이용하여 고차원의 특징공간으로 맵핑하고, 이 공간에서 가시유닛과 은닉유닛에 대한 에너지 함수를 정의한다. 딥러닝에서 많이 쓰이고 있는 rectified linear unit을 커널 RBM을 위한 커널 함수로 사용한다. 커널 RBM을 학습하기 위해 경사기반 contrastive divergence 알고리즘을 유도하고, 학습에 필요한 파라메터의 갱신법을 제시한다. MNIST와 STL-10 데이터를 사용한 실험에서 제안된 기법은 유용한 특징들을 학습하며, 기존의 RBM보다 더 향상된 인식결과를 나타낸다.

Key Words : feature learning, restricted Boltzmann machine, kernel method, image classification

ABSTRACT

This paper presents a restricted Boltzmann machine (RBM) based on kernel method for unsupervised feature learning. The key idea of kernel RBM is that the input data is first implicitly mapped into a high-dimensional feature space, and the energy function is defined with the visible units and hidden units in that space. We propose the use of rectified linear unit for the kernel RBM as a kernel function, which is widely used in deep learning. The gradient-based contrastive divergence algorithm is used for the training of the kernel RBM, and the parameter update rules are derived for the learning. Experimental results on MNIST and STL-10 dataset show that the proposed approach can learn the useful representations, and it outperforms the conventional RBMs on the classification task.

Ⅰ. 서 론

특징학습(feature learning) 또는 표현학습(representation learning)은 많은 데이터로부터 검출 또는 인식을 위해 필요한 적절한 특징 또는 유용한 표현들을 자동적으로 추출하기 위한 기계학습의 한 분야이다[1-4]. 특징학습 기법은 이미지, 음성 그리고 기타 데이터 등에서 좋은 특징들을 뽑아 인식 시스템의 입력으로 사용되어 기존의 특징들에 비해 인식성능 향상을 목적으로 하고 있다. 현재 특징학습은 음성인식, 신호처리, 물체인식, 자연언어처리 등의 기계학습 및 딥러닝 분야에서 사용되고 있으며, 관련된 연구가 활발하게 진행되고 있다[1,10,17].

특징학습 기술은 지도와 비지도 기법으로 분류된다

※ 이 연구는 한국연구재단 논문연구과제(NRF-2016R1D1A1B03932813) 지원으로 수행되었습니다.
◆ First Author : Chonnam National University, School of Electronic and Computer Engineering, dkim@jnu.ac.kr, 정회원
° Corresponding Author : Gwangju Institute of Science and Technology, School of Electrical Engineering and Computer Science, jwshin@gist.ac.kr, 정회원
논문번호 : 201905-095-A-RE, Received May 30, 2019; Revised July 17, 2019; Accepted July 18, 2019

www.dbpia.co.kr

[1,3]. 지도 특징학습은 입력 데이터와 관련된 레이블을 사용하여 특징을 학습하는 기법이다. 학습 시스템이 나타내는 출력과 데이터 레이블 사이의 오차를 계산하여 학습과정 중에 이를 반영하여 유용한 특징이 발생하도록 한다. 지도 특징학습 기법으로는 지도 사전학습(dictionary learning)[5], 다층 퍼셉트론(multilayer perceptron, MLP)[2] 또는 CNN (convolutional neural network)[2]을 이용한 지도 신경망이 주로 사용된다. 한편 비지도 특징학습은 데이터의 레이블이 없이 단지 입력 데이터만을 사용하여 특징들을 학습하는 기법이다. 지도 특징학습에 비해 레이블이 필요 없기 때문에 많은 데이터를 이용할 수 있어 현재 특징학습을 위해 널리 사용되고 있다. 비지도 특징학습의 주요 기법으로는 K-평균 군집화(K-mean clustering)[6], 자기부호화기(autoencoder)[2] 그리고 제한된 볼츠만머신(restricted Boltzmann machine, RBM)[7-11] 등이 있다.

RBM 기법은 입력 데이터에 대한 확률분포(probability distribution)을 학습할 수 있는 단층의 생성모델 (generative model) 신경망이다[7,8]. RBM은 일반적인 볼츠만 머신의 제한된 형태로 두 개의 유닛 그룹으로 구성된 이진 그래프(bipartite grape) 형태를 갖고 있다. 두 그룹의 유닛은 가시유닛과 은닉유닛으로 두 그룹사이의 대칭 연결을 갖고 있지만, 그룹 내의 유닛사이에는 연결이 없는 제한된 형태의 모델이다. 제한된 형태 때문에 다른 모델에 비해 효율적인 학습 알고리즘이 사용되는데, 이를 경사기반 CD(contrastive divergence) 알고리즘이라 한다[7,8]. 학습된 RBM에서 입력이 가시유닛에 주어질 때, 이때 은닉유닛에 나타나는 값들을 입력 데이터의 새로운 특징으로 사용된다. RBM은 딥러닝에서 기본 형태로 사용되는데, 여러 개의 RBM을 계층적으로 쌓아서 깊은 계층을 갖는 신경망을 형성하는데 이용된다[12,17].

본 논문에서 비지도 특징학습을 위한 커널 기반의 새로운 RBM을 제안한다. 커널 기반 RBM은 기존의 RBM의 개념과 기계학습에서 널리 사용되는 커널 개념[13,14]을 결합한 형태이다. 커널 RBM의 핵심 아이디어는 입력 데이터를 비선형 함수를 이용하여 고차원의 특징공간 (high-dimensional feature space)으로 암시적으로 (implicitly) 맵핑하고, 그 공간에서 실수값을 갖는 가시유닛과 은닉유닛으로 구성된 RBM을 형성하는 것이다. 그리고 이 공간에서 RBM에 필요한 에너지 함수 및 결합 확률분포를 정의한다. 입력 데이터가 맵핑될 때 커널 함수가 이용되는데, 본 논문에서는 딥러닝에서 많이 사용되는 ReLU(rectified linear unit)[2,11]을 커널 RBM을 위한 커널 함수로 제안한다. 이 기법은 기존 RBM과 다르게 가시유닛과 은닉유닛에 대해 실수값을 갖는 가우시안 확률분포를 사용하는 특징을 갖고 있다. 이러한 특징에 근거하여 커널 RBM을 학습하기 위해 특징공간에서 CD 알고리즘을 유도하고, 경사상승 기반 학습 알고리즘에 필요한 파라메터 갱신법을 제시한다. MNIST[16]와 STL-10[4] 데이터 베이스를 사용하여 제안된 기법의 학습과정의 특징들을 살펴보고, 비지도 학습을 통해 얻어진 특징들이 이미지 인식실험에서 기존의 RBM 기법에 비해 더 높은 인식 성능을 나타낸다.

본 논문의 본론 II장에서는 기존의 RBM을 간단히 소개하고, 새로운 커널 기반 RBM 기법을 제시한다. III장에서는 실험과 결과에 대해 나타내고, IV에서는 결론을 맺는다.

Ⅱ. 본 론

2.1. RBM

이 단원에서 기존의 RBM[7-11]에 대한 구조와 학습 알고리즘을 간단히 살펴본다. RBM는 비방향성 그래프 모델(undirected graphical model)로 관측 데이터를 나타내는 n개의 가시유닛(visible units), $\boldsymbol{v}=(v_1,...,v_n)$ 와 관측 데이터의 새로운 특징을 표현할 수 있는 m개의 은닉유닛(hidden units), $\boldsymbol{h}=(h_1,...,h_m)$ 으로 구성된다. 가시유닛과 은닉유닛의 형태에 따라 다양한 구조의 RBM이 제안되었다. 가장 전형적인 RBM의 구조는 입력 데이터의 형태에 따라 일반적으로 이진 또는 실수 값을 표현하는 가시유닛과, 이진 값을 갖는 은닉유닛으로 구성된다. 이진 값을 갖는 가시유닛과 은닉유닛의 경우, 보통 Bernoulli-Bernoulli RBM (BBRBM)이라 부르는데, 이 경우 가시유닛과 은닉유닛의 모든 조합에 대해 에너지 함수를 정의할 수 있으며 이는 다음과 같이 주어진다[7,8].

$$E(\boldsymbol{v},\boldsymbol{h}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}v_ih_j - \sum_{i=1}^{n} a_iv_i - \sum_{j=1}^{m} b_jh_j \quad (1)$$

여기서 v_i, h_j는 i번째 가시유닛과 j번째 은닉유닛의 이진 상태를 나타내며, w_{ij}는 두 유닛사이의 가중치이고, a_i, b_j는 각각의 바이어스를 나타낸다. 실수값을 갖는 입력 데이터의 경우에 i번째 가시유닛에

www.dbpia.co.kr

대해 평균과 분산이 a_i와 σ_i^2인 가우시안 분포로 모델링하여 표현하며, 이를 Gaussian-Bernoulli RBM(GBRBM)이라 부른다[9,10]. 이때의 에너지 함수는 다음과 같다[9].

$$E(\boldsymbol{v},\boldsymbol{h}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}\frac{v_i}{\sigma_i^2}h_j - \sum_{i=1}^{n}\frac{(v_i-a_i)^2}{2\sigma_i^2} - \sum_{j=1}^{m} b_j h_j \quad (2)$$

이러한 두 가지의 모델 가정 하에 두 유닛에 대한 결합 확률분포는 에너지 함수에 의해, $p(\boldsymbol{v},\boldsymbol{h}) = \frac{1}{Z}e^{-E(\boldsymbol{v},\boldsymbol{h})}$ 형태로 정의된다. 여기서 Z는 파티션(partition) 함수를 나타내며, $Z = \sum_{\boldsymbol{v},\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h})}$에 의해 표현된다. 그리고 가시유닛의 한계(marginal) 확률분포는 모든 가능한 은닉유닛에 대해 합하므로 확률적인 공식에 의해 $p(\boldsymbol{v}) = \frac{1}{Z}\sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h})}$의 형태를 갖는다. RBM 구조는 가시유닛 층과 은닉유닛 층 사이에 가중치 연결만이 존재하고, 같은 층 유닛사이에는 연결이 존재하지 않기 때문에 가시변수가 주어진 경우 은닉변수의 조건부 확률이 독립적이다. 마찬가지로 은닉변수가 주어진 조건하에서 가시유닛의 조건부 확률도 서로 독립적이다. 따라서 BBRBM의 경우 각 유닛에 대한 조건부 확률은 다음과 같이 간단한 형태를 갖는다[7,8].

$$p(h_j = 1|\boldsymbol{v}) = \sigma(\sum_i w_{ij}v_i + b_j) \quad (3)$$

$$p(v_i = 1|\boldsymbol{h}) = \sigma(\sum_j w_{ij}h_j + a_i) \quad (4)$$

여기서 $\sigma(x) = 1/(1+\exp(-x))$는 sigmoid 함수이다. GBRBM의 경우 가시유닛의 조건부 확률은 다음과 같이 가우시안 형태이다[9].

$$p(v_i = v|\boldsymbol{h}) = N(v|\sum_j w_{ij}h_j + a_i, \sigma_i^2) \quad (5)$$

여기서 $N(|\mu,\sigma_i^2)$은 평균이 μ이고 분산이 σ_i^2인 가우시안 확률분포를 나타낸다. 이러한 성질은 $\boldsymbol{v}$와 $\boldsymbol{h}$사이에 깁스 샘플링(Gibbs sampling)을 효율적으로 수행하여 학습과정이 빠르게 할 수 있다.

RBM을 학습하기 위해 가장 많이 사용되는 기법은 경사기반 CD 알고리즘이다[7,8]. 이는 RBM의 유사도 함수(likelihood function)을 최대화 하도록 하는 근사적인 기법으로, 내부적으로 깁스샘플링 과정을 수행한다. 1단계(single-step) CD (CD-1) 알고리즘에 의해 가중치 파라메터에 대한 유사도 함수의 경사치(gradient)은 다음과 같이 간단한 형태로 주어진다[7,8].

$$\nabla w_{ij} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \quad (6)$$

첫 번째 부분은 양수 경사치(positive gradient)라 부르며 확률분포에서 쉽게 구할 수 있으며, 두 번째 부분은 음수 경사치(negative gradient)로 깁스샘플링을 통해 얻어진 데이터를 사용해 구하게 된다. 바이어스 파라메터에 대한 경시치도 유사하게 구할 수 있다. 본 논문에서는 위의 CD-1 알고리즘을 이용하여 기존의 RBM을 학습한다.

2.2. 커널 RBM

2.2.1 커널 RBM 구조

이 단원에서는 새로운 커널 RBM에 대해 제안한다. 커널 RBM은 기계학습에서 많이 사용되는 커널 기법[13,14]을 RBM에 적용하는 모델이다. 커널 RBM에 대한 핵심 아이디어는 가시변수를 비선형 함수에 의해 고차원의 특징공간(feature space)으로 암시적으로 맵핑하여 그 공간에서 커널함수에 의해 RBM의 에너지 함수를 정의하는 것이다. 먼저 $\boldsymbol{v} = (v_1,...,v_n)^T$을 원래의 n차원 데이터 공간에서 정의된 가시변수 벡터라 하자. 그리고 비선형함수 $\phi: R^n \rightarrow R^f$ 는 n차원의 입력공간에서 f차원의 특징벡터 공간으로 맵핑하는 비선형 함수라 가정하자. 이때 경우에 따라 f는 무한 차원도 가능하다. $\phi(\boldsymbol{v})$는 가시변수 $\boldsymbol{v}$에 해당되는 비선형적으로 맵핑되는 f차원의 특징벡터 공간에서 가시변수이다. 비록 $\phi(\boldsymbol{v})$의 형태는 분명하게 규정되지 않지만 커널트릭(kernel trick)[13,14]을 사용하여 특징벡터 공간에서 두 벡터의 내적을 계산할 수 있다. 예를 들면, 특징벡터 공간 R^f에서 두 가시벡터를 $\phi(\boldsymbol{v}_i)$와 $\phi(\boldsymbol{v}_j)$라 하면, 이 두 벡터 사이의 내적은 $k(\boldsymbol{v}_i,\boldsymbol{v}_j) = \phi(\boldsymbol{v}_i)^T\phi(\boldsymbol{v}_j)$의 형태의 커널함수(kernel function), $k(,)$에 의해 정의된다.

커널 RBM 구조는 특징벡터 공간 R^f에서 정의된

www.dbpia.co.kr

가시변수 $\phi(\boldsymbol{v})$에 의한 층과 R^m공간에서 정의된 실수 값을 갖는 m차원 은닉변수 벡터 $\boldsymbol{h}=(h_1,...,h_m)^T$에 의한 층으로 구성된다. 두 층 사이의 연결 가중치를 위해 m개의 n차원의 가중치 벡터 $(\boldsymbol{w}_1,\dots,\boldsymbol{w}_m)$들을 정의하고, 이들 각각을 가시변수와 같이 특징벡터 공간으로 비선형 함수에 의해 맵핑한다. 이때 $\Phi(W)=(\phi(\boldsymbol{w}_1),\dots,\phi(\boldsymbol{w}_m))$을 가중치 벡터들에 대응되는 특징벡터 공간에서의 가중치 행렬이라고 가정하자. 그러면 $\phi(\boldsymbol{w}_j)$는 j번째 은닉유닛 h_j와 특징공간에서 가시벡터 $\phi(\boldsymbol{v})$을 연결하는 가중치 벡터가 된다. 이러한 커널 RBM 구조 하에서 다음과 같은 에너지 함수를 제안한다.

$$\begin{aligned}&E(\phi(\boldsymbol{v}),\boldsymbol{h})\\&=\frac{1}{2}(\phi(\boldsymbol{v})^T\phi(\boldsymbol{v})-2\boldsymbol{h}^T\Phi(W)^T\phi(\boldsymbol{v})+\boldsymbol{h}^T\boldsymbol{h})\\&=\frac{1}{2}(k(\boldsymbol{v},\boldsymbol{v})-2\sum_{j=1}^{m}h_jk(\boldsymbol{w}_j,\boldsymbol{v})+\sum_{j=1}^{m}h_j^2)\end{aligned}\tag{7}$$

여기서 커널트릭에 의해 $k(\boldsymbol{v},\boldsymbol{v})=\phi(\boldsymbol{v})^T\phi(\boldsymbol{v})$이고, $k(\boldsymbol{w}_j,\boldsymbol{v})=\phi(\boldsymbol{w}_j)^T\phi(\boldsymbol{v})$이다. 이러한 에너지 함수의 형태는 R^f와 R^m공간에서 $\phi(\boldsymbol{v})$와 $\boldsymbol{h}$가 실수 값을 갖는다는 가정 하에서 결합 확률분포가 가우시안 분포를 갖도록 이차항들의 형태로 구성되었다. 이에 근거하여 RBM과 같은 형태의 결합 확률분포 $p(\phi(\boldsymbol{v}),\boldsymbol{h})=e^{-E(\phi(\boldsymbol{v}),\boldsymbol{h})}/Z$을 정의할 수 있다. 위와 같은 커널 함수에 의해 비선형 특징공간에서 정의된 에너지 함수와 결합 확률분포를 갖는 RBM을 커널(kernel) RBM(KRBM)이라 한다. 이때 파티션 함수는 다음과 같이 주어진다.

$$Z=\int\int e^{-E(\phi(\boldsymbol{v}),\boldsymbol{h})}d\boldsymbol{h}d\phi(\boldsymbol{v})\tag{8}$$

KRBM에 대한 결합 확률분포가 주어진 경우 특징공간에서 $\phi(\boldsymbol{v})$에 대한 한계 확률분포는 다음과 같다.

$$p(\phi(\boldsymbol{v}))=\frac{1}{Z}\int e^{-E(\phi(\boldsymbol{v}),\boldsymbol{h})}d\boldsymbol{h}\tag{9}$$

그리고 식 (7)의 에너지 함수가 $\phi(\boldsymbol{v})$와 $\boldsymbol{h}$에 이차함수 형태를 갖고 있기 때문에 KRBM에 대한 조건부 확률분포들을 다음과 같이 각각 다변수(multivariate) 가우시안 형태를 갖는다.

$$p(\boldsymbol{h}|\phi(\boldsymbol{v}))=N(\boldsymbol{h}|\Phi(W)^T\phi(\boldsymbol{v}),I_m)\tag{10}$$

$$p(\phi(\boldsymbol{v})|\boldsymbol{h})=N(\phi(\boldsymbol{v})|\Phi(W)\boldsymbol{h},I_f)\tag{11}$$

여기서 I_m와 I_f은 각각 m차원과 f차원에서 단위행렬을 나타낸다. 위의 두 가지 조건부 확률분포가 각 성분에 대해 독립적이기 때문에 CD 알고리즘의 깁스샘플링 과정을 기존 RBM과 같이 효율적으로 수행할 수 있다.

KRBM에서는 가시벡터가 커널함수를 통해 고차원의 특징공간으로 맵핑되므로 커널 함수는 매우 중요한 역할을 수행한다. 커널함수의 선택은 주어진 데이터와 수행하는 일에 따라 다르게 선택된다. 기계학습에서 가우시안 커널과 같이 매우 다양한 커널 함수들이 사용되고 있다[13,14]. 최근에는 ReLU 함수는 딥러닝 분야에서 가장 많이 사용되는 activation 함수이다[2,11]. 또한 SVM을 위한 커널함수로 사용되고 있다[15]. ReLU 함수는 딥러닝뿐아니라 커널 기법에서 학습에 매우 효과적이라고 알려졌다[11,15]. 따라서 본 논문에서는 KRBM을 위한 커널함수로 ReLU 함수를 사용한다. n차원의 두 벡터 $\boldsymbol{w}$와 $\boldsymbol{v}$에 대해 ReLU 커널함수의 정의는 다음과 같다[2,11].

$$k(\boldsymbol{w},\boldsymbol{v})=\max(\boldsymbol{w}^T\boldsymbol{v}+b,0)\tag{12}$$

여기서 일반화를 위해 bias 성분 b가 포함되었다. 위와 같은 ReLU 함수는 0에서 미분가능하지 않고 커널함수로서 성질이 증명되지 않았지만 실험적으로 좋은 결과를 나타내고 있다[11,15].

2.2.2 커널 RBM 학습

이 장에서는 ReLU 커널 함수를 사용하여 위에서 제시된 KRBM을 학습시키기 위한 CD 알고리즘을 제시한다. KRBM을 학습하기 위한 방법은 로그 유사도(log-likelihood) 함수를 최대화하는 경사 승강법(gradient ascent)을 이용한다. 고차원의 특징공간에서 하나의 학습 데이터 $\phi(\boldsymbol{v})$가 주어진 경우, 로그 유사도 함수는 식 (8)과 (9)로부터 다음과 같이 주어진다[8].

$$\begin{aligned}l&=\ln p(\phi(\boldsymbol{v}))=\ln\frac{1}{Z}\int e^{-E(\phi(\boldsymbol{v}),\boldsymbol{h})}d\boldsymbol{h}\\&=\ln\int e^{-E(\phi(\boldsymbol{v}),\boldsymbol{h})}d\boldsymbol{h}\\&\quad-\ln\int\int e^{-E(\phi(\boldsymbol{v}),\boldsymbol{h})}d\boldsymbol{h}d\phi(\boldsymbol{v})\end{aligned}\tag{13}$$

www.dbpia.co.kr

이 때 KRBM의 파라메터 $\theta = (w_{ij}, b_j)_{i=1\ldots n, j=1\ldots m}$에 대한 로그 유사도에 대한 경사값은 다음과 같이 주어진다.

$$\frac{\partial l}{\partial \theta} = -\int p(\boldsymbol{h}|\phi(\boldsymbol{v}))\frac{\partial E(\phi(\boldsymbol{v}),\boldsymbol{h})}{\partial \theta}d\boldsymbol{h} + \int\int p(\phi(\boldsymbol{v}),\boldsymbol{h})\frac{\partial E(\phi(\boldsymbol{v}),\boldsymbol{h})}{\partial \theta}d\boldsymbol{h}d\phi(\boldsymbol{v}) \quad (14)$$

위의 KRBM의 로그 유사도의 경사값은 두 기댓값의 합으로 구성된다. 첫 번째 항은 쉽게 계산이 가능하지만, 두 번째 항은 결합 확률분포하에서 기댓값을 효율적으로 계산할 수 없다. 이러한 문제를 극복하기 위해 RBM에 적용되었던 CD-1 알고리즘을 사용한다[8]. CD 알고리즘은 두 번째 항의 기댓값을 모델 분포로부터 샘플값으로 근사화한다. 이러한 샘플들은 깁스 샘플링을 통해 얻어진다. 깁스 샘플링은 데이터 샘플 $\boldsymbol{v}^{(0)}$으로 먼저 초기화하고 이를 통해 $\phi(\boldsymbol{v}^{(0)})$으로 맵핑한다. 그리고 $p(\boldsymbol{h}|\phi(\boldsymbol{v}^{(0)}))$로부터 효율적으로 $\boldsymbol{h}^{(0)}$ 샘플을 얻을 수 있다. 순차적으로 $p(\phi(\boldsymbol{v})|\boldsymbol{h}^{(0)})$으로부터 $\boldsymbol{v}^{(1)}$의 샘플을 얻는 것이 가능하다. 이러한 CD알고리즘에 기초하여 위의 파라메터에 대한 로그 유사도 경사값은 다음과 같이 근사화 된다.

$$\frac{\partial l}{\partial \theta} \approx -\int p(\boldsymbol{h}|\phi(\boldsymbol{v}^{(0)}))\frac{\partial E(\phi(\boldsymbol{v}^{(0)}),\boldsymbol{h})}{\partial \theta}d\boldsymbol{h} + \int p(\boldsymbol{h}|\phi(\boldsymbol{v}^{(1)}))\frac{\partial E(\phi(\boldsymbol{v}^{(1)}),\boldsymbol{h})}{\partial \theta}d\boldsymbol{h} \quad (15)$$

만약 ReLU 커널함수가 사용되는 경우, 가중치 w_{ij}와 바이어스 b_j에 대한 경사값은 식 (15)로부터 다음과 같이 쉽게 유도할 수 있다.

$$\nabla w_{ij} = v_i^{(0)}\hat{h}_j^{(0)} - v_i^{(1)}\hat{h}_j^{(1)} \quad (16)$$

$$\nabla b_j = \hat{h}_j^{(0)} - \hat{h}_j^{(1)} \quad (17)$$

여기서 $\hat{h}_j^{(k)} = \max(\boldsymbol{w}_j^T\boldsymbol{v}^{(k)} + b_j, 0), k = 0, 1$이다. 깁스 샘플링 과정중에 $p(\phi(\boldsymbol{v})|\boldsymbol{h})$로부터 $\boldsymbol{v}$의 샘플을 얻는 과정이 필요하다. $\boldsymbol{v}$의 샘플 값은 식 (11)의 조건부 확률이 최대가 되는 값을 다음과 같이 선택한다.

$$\begin{aligned}\boldsymbol{v}^* &= \arg\max_{\boldsymbol{v}} \ln p(\phi(\boldsymbol{v})|\boldsymbol{h}) \\ &= \arg\min_{\boldsymbol{v}} \|\phi(\boldsymbol{v}) - \sum_{j=1}^{m} h_j\phi(\boldsymbol{w}_j)\|^2\end{aligned} \quad (18)$$

위 식을 커널 트릭을 사용하여 전개하면, $\boldsymbol{v}^*$는 다음과 같은 목적함수 $C(\boldsymbol{v})$를 최소화하는 값으로 구해진다.

$$C(\boldsymbol{v}) = k(\boldsymbol{v},\boldsymbol{v}) - 2\sum_{j=1}^{m} h_j k(\boldsymbol{v},\boldsymbol{w}_j) \quad (19)$$

만약 ReLU 커널함수가 사용되는 경우, ReLU 함수를 위 목적함수에 대입한 후, $\boldsymbol{v}$에 대해 미분을 취하고 영으로 놓으면 고정점 반복법(fixed-point iteration)에 의해 다음과 같은 반복적인 방법에 의해 t번째 단계에서 샘플 값을 다음과 같이 추정할 수 있다.

$$\boldsymbol{v}_{t+1}^* = \sum_{j=1}^{m} h_j \nabla k(\boldsymbol{v}_t^*, \boldsymbol{w}_j)\boldsymbol{w}_j \quad (20)$$

여기서 $\nabla k(\boldsymbol{v}_t^*, \boldsymbol{w}_j)$는 ReLU 함수의 미분값을 나타낸다. 본 논문에서는 초기값으로 $\boldsymbol{v}_0^* = \boldsymbol{v}^{(0)}$로 놓고, 다음 단계 추정값인 $\boldsymbol{v}_1^*$을 샘플값 $\boldsymbol{v}^{(1)}$로 사용하였다.

2.2.3 기존 RBM과 비교

본 논문에서 제안된 KRBM은 은닉유닛을 ReLU 함수를 사용하는 Noisy-ReLU RBM(NReLU RBM)[11]와 비슷한 구조를 갖는다. 입력 데이터가 주어진 경우 은닉유닛에서 ReLU 함수를 사용하여 특징을 추출하는 과정은 두 구조가 같다. 하지만 제안된 KRBM과 기존의 RBM는 다음과 같은 차이점을 갖는다. 첫째 제안된 구조는 기존 RBM과는 다르게 커널 기법에 근거하여 RBM을 구성함으로써 RBM을 새로운 관점에서 해석할 수 있다. 두 번째는 제안된 구조에서는 가시유닛과 은닉유닛의 분포가 모두 가우시안으로 모델링함으로 학습과정에서 필요한 깁스샘플링 과정이 기존과 다르다. 예를 들면, KRBM의 은닉유닛의 샘플링은 식 (10)을 통해 수행하지만, Noisy-ReLU RBM의 경우 샘플링을 위해 근사적으로 $\max(0, x + N(0, \sigma(x)))$ 형태의 함수를 사용한다[11]. 마지막으로 기존 RBM에서도 다양한 형태의 가시유닛과 은닉유닛의 사용하지만, KRBM에서는 커널함수를 사용함으로 기계학습에서 널리 사용되고 있는

www.dbpia.co.kr

많은 커널함수를 통해 앞으로 새로운 KRBM 구조가 가능하다는 것이다.

Ⅲ. 실험 및 결과

제안된 KRBM의 성능을 평가하기 위해 두 가지 다른 데이터 베이스 MNIST[16]와 STL10[4]에 대한 비지도 특징 학습 실험을 수행하였다.

3.1 MNIST

먼저 MNIST 데이터 셋에 대한 학습 및 인식실험을 수행한다. MNIST 데이터 셋은 28×28 크기의 grayscale 형식의 손으로 쓴 0-9 사이의 숫자 영상이다[16]. 60,000개의 학습 샘플과 10,000개의 테스트 샘플로 구성된다. 제안된 KRBM과 비교하기 위해 BBRBM과 이진 가시유닛을 갖는 NReLU RBM을 학습하였다. 학습을 위해 CD-1 알고리즘을 사용하였고, 경사 승강법의 학습율은 KRBM에 대해 0.001을 사용하고 나머지 기법에 대해서는 0.01을 사용하였다. 그리고 0.9의 모멘텀을 사용하였다. 가중치의 초기값은 영평균과 0.01의 표준편차를 갖는 가우시안으로부터 발생된 랜덤값을 사용하였다. 학습을 위한 배치 크기는 100이며, 1000 epoch 만큼 학습하였다. RBM의 입력은 $28 \times 28 = 784$ 크기의 벡터이며, 은닉층의 수는 가변적으로 사용하여 성능을 비교하였다.

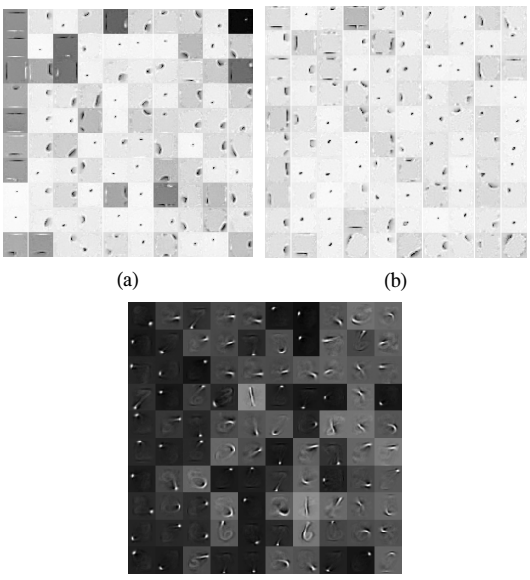

그림 1. MNIST에서 학습된 1024개 은닉유닛을 가지는 RBM에서의 가중치(필터)의 시각화. (a) BBRBM, (b) NReLU RBM, (c) KRBM
Fig. 1. Visualization of the weights of RBMs with 1024 hidden units trained on MNIST.

각각 학습된 RBM에 대한 특징 표현에 대한 시각화를 살펴보기 위해 학습된 필터(가중치)에 대한 값을 그림으로 표현하였다. 그림 1은 BBRBM, NReLU RBM 그리고 제안된 KRBM에 대한 필터로서 MNIST에 대해 1024개의 은닉유닛중에 분산이 가장 큰 100개에 대한 그림이다. 일반적으로 RBM은 Gabor 필터와 닮은 국부적인 형태의 필터를 나타낸 것으로 잘 알려졌다[4,7,8]. 또한 NReLU에 의해 학습된 필터는 조금 더 sparse한 형태를 갖고 있음을 알 수 있다. 한편 KRBM은 기존 RBM과는 다른 형태의 필터 그림이 나타남을 알 수 있다. Sparse한 형태의 필터뿐아니라 다양한 형태를 갖는 필터가 학습됨을 알 수 있다.

위에서 학습된 RBM을 사용하여 입력이 주어진 경우, 은닉유닛의 수만큼 은닉유닛에 나타나는 값들을 새로운 특징벡터로 추출하고, 이를 인식하기 위해 softmax 인식기를 사용해 3가지 RBM에 대해 인식실험을 수행하였다. 그림 2는 가변적인 은닉유닛의 수에 따른 테스트 데이터에 대한 인식 정확도를 나타낸다. 전체적으로 은닉유닛의 수가 증가할수록 인식 정확도가 향상됨을 알 수 있다. NReLU RBM은 적은 수의 은닉수에서 RBM보다 더 낮은 성능을 보이나 2048개 에서는 RBM이 약간 더 나은 성능을 보였다. 제안된 KRBM은 적은 수에 대해서 기존의 BBRBM과 NReLU RBM에 비해 훨씬 향상된 정확도를 나타내며, 2048개에서 거의 비슷한 성능을 보임을 알 수 있다. 따라서 제안된 기법의 기존의 기법에 비해 비지도 특징학습에 있어 효과적임을 알 수 있다.

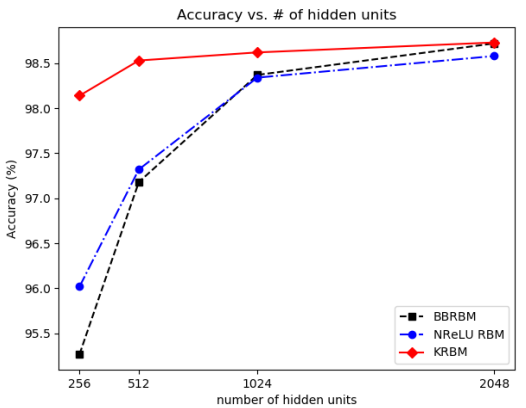

그림 2. MNIST에서 가변적인 은닉유닛 수에 따른 테스트 인식 정확도
Fig. 2. Test classification accuracy with various number of hidden units on MNIST

www.dbpia.co.kr

3.2 STL-10

STL-10은 비교사 특징학습 또는 딥러닝을 위한 영상 인식용 데이터 셋이다[4]. 이는 기존의 CIFAR-10 데이터 셋과 비슷하지만 약간의 수정을 통해 만들어졌다. 각 클래스는 CIFAR-10보다 적은 레이블을 갖는 학습 샘플과, 훨씬 많은 양의 레이블이 없는 샘플 영상으로 구성되어 교사 학습 전에 비교사 학습을 위해 제작되었다. 레이블이 없는 샘플들은 모델 학습을 위해 이용되며, 특히 여러 가지 비교사 학습 방법들을 개발하기 위해 사용된다. 영상 샘플은 96×96 픽셀 크기의 컬러 영상이며, 10개의 클래스 (airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck)으로 구성된다. 비교사 학습을 위한 데이터 수는 100,000개이며, 교사학습을 위해 클래스당 500개의 학습 영상과 800개 테스트 영상으로 구성된다[4]. 비교사용 데이터 셋은 레이블을 갖는 영상뿐아니라 다양한 형태의 영상들이 추가로 포함되어 있다.

STL-10 데이터를 비지도 학습과 인식에 사용하기 위해 MNIST와 다르게 전처리 과정을 수행한다. 본 논문에서는 비지도 특징학습에 널리 쓰이는 방법을 따른다[4]. 먼저 모든 96×96 크기의 영상을 계산량과 메모리 사용을 줄이기 위해 32×32 크기로 줄인다. 그리고 RBM의 학습을 위해 6×6 크기의 컬러 패치(patch)을 클래스가 없는 100,000개의 영상의 임의의 위치에서 총 500,000개를 추출하여 사용한다. 따라서 RBM의 입력의 개수는 $6 \times 6 \times 3 = 108$이다. 모든 데이터는 whitening을 과정을 수행하고, KRBM에 대해서는 영평균을 그리고 나머지 RBM에 대해서는 영평균과 단위분산을 갖도록 정규화한다. 정규화된 입력 데이터가 실수값을 가지므로 가시유닛에 대해 가우시안 모델을 사용하였다. 학습 데이터로 RBM을 학습한 후 인식 실험을 수행하기 위해 입력 영상에 대해 같은 간격으로 패치들을 생성하고 RBM을 통해 특징들을 추출한다. 최종적으로 한 입력 영상당 가변적인 은닉유닛의 수에 따라 특징 벡터를 발생하였다[4]. 이를 MNIST와 마찬가지로 softmax 인식기를 사용해 인식실험을 수행하였다. 그림 3은 STL-10에서 가변적인 은닉유닛의 수에 따른 테스트 데이터에 대한 인식 정확도를 나타낸다. 특징학습이 없는 경우 31.8%의 매우 낮은 인식율을 나타내었다. RBM에 의해 특징학습을 수행하고 그 특징들을 이용하는 경우 그림 3.과 같이 매우 높은 인식율의 향상이 보임을 알 수 있다. STL-10의 경우 많은 수의 은닉유닛의 경우에 NReLU보다 GBRBM이 더 높은 인식율을 나타내었다. 제안된 KRBM은 두 기법에 비해 전체적으로 같거나 약간 높은 인식율을 나타내었다. 따라서 STL-10 데이터에서도 제안된 커널 기반 RBM이 효과적임을 알 수 있다.

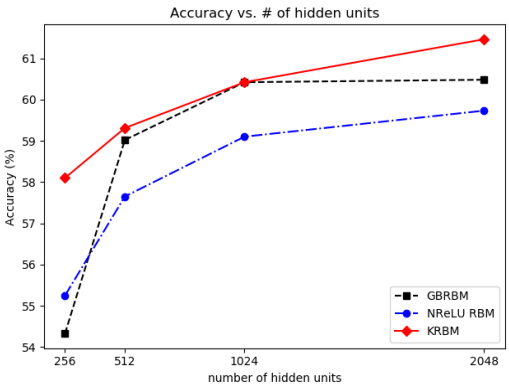

그림 3. STL-10에서 가변적인 은닉유닛 수에 따른 테스트 인식 정확도
Fig. 3. Test classification accuracy with various number of hidden units on STL-10

Ⅳ. 결 론

본 논문에서는 기계학습에서 많이 사용되는 커널 기법을 RBM에 적용하여 커널 RBM을 제안하였고, 이를 비교사 특징학습에 이용하였다. 입력 데이터를 비선형 함수를 이용하여 고차원의 특징공간으로 맵핑하여 이 공간에서 에너지 함수와 결합 확률분포를 제안하였다. 커널함수로 ReLU 함수를 이용하고, 커널 RBM 학습을 위한 CD 알고리즘 기반 파라메터 갱신법을 제시하였다. MNIST와 STL-10 데이터에서의 실험결과 제안한 커널 RBM은 인식에 유용한 특징들을 추출하고 인식실험에서 기존 알고리즘 보다 우수한 성능을 나타내었다.

앞으로의 연구 방향으로 제안된 기법을 음성, 텍스트 등 다양한 형태의 데이터에 적용이 필요하며, 커널기반 convolutional RBM와 다양한 커널 함수에 대해서도 연구가 필요하다. 그리고 커널 RBM을 계층적으로 적용하여 깊은 구조를 갖는 신경망을 구성하고 이를 학습시키는 연구도 필요하다.

References

[1] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Trans. PAMI*, special issue Learning Deep Architectures, vol. 35,

www.dbpia.co.kr

no. 8, pp. 1798-1828. Aug. 2013.
[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
[3] Wikipedia, *Feature learning*, Retrieved May 30, 2019, from https://en.wikipedia.org/wiki/Feature_learning
[4] A. Coates, H. Lee, and A. Ng, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. Fourteenth Int. Conf. Artificial Intell. and Statistics,* PMLR 15, pp. 215-223, 2011.
[5] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," *Advances in NIPS*, 2008.
[6] A. Coates and A. Ng, "Learning feature representations with k-means," *Neural Networks: Tricks of the trade*, pp. 561-580, Springer, 2012.
[7] G. Hinton, "A practical guide to training restricted boltzmann machines," *Tech. Rep. UTML TR 2010‐003*, University of Toronto, 2010.
[8] A. Fischer and C. Igel, "Training restricted Boltzmann machines: An introduction," *Pattern Recognition*, vol. 47, pp. 25-39, 2014.
[9] K. Cho, A. Ilin, and T. Raiko, "Improved learning of Gaussian-Bernoulli restricted Boltzmann machines," in *Proc. 21st ICANN*, pp. 10-17, 2011.
[10] A. Krizhevsky, "Learning multiple layers of features from tiny images," Master's thesis, Dept. of Compt. Science, Univ. of Toronto, 2009.
[11] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," *ICML,* pp. 807-814, 2010.
[12] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504-507, 2006.
[13] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis,* Cambridge University Press, 2004.
[14] B. Schölkopf and A. Smola, *Learning with Kernels*, MIT Press, 2002.
[15] B. Pradhan and M. I. Sameen "Manifestation of SVM-based rectified linear unit (ReLU) kernel function in landslide modelling," *Space Sci. and Commun. for Sustainability*, Springer, pp. 185-195, 2018.
[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proc. IEEE,* vol. 86, no. 11, pp. 2278-2324, Nov. 1998.
[17] B.-K. Lee and J.-H. Chang, "A study on the speech bandwidth extension using deep learning in speech communication systems," in *KICS Summer Conf.*, pp. 944-945, Jun. 2017.

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 전자컴퓨터공학부 교수
<관심분야> 딥러닝, 음성처리, 기계학습
[ORCID:0000-0001-9316-7069]

신 종 원 (Jong Won Shin)

2002년 2월 : 서울대학교 전기공학부 학사
2008년 8월 : 서울대학교 전기컴퓨터공학부 박사
2008년 12월~2012년 8월 : Qualcomm Inc., Senior Engineer
2012년 9월~현재 : 광주과학기술원 전기전자컴퓨터공학부 부교수
<관심분야> 음성/음향/오디오처리
[ORCID:0000-0002-8910-0264]

www.dbpia.co.kr