커널기반 제한된 볼츠만머신을 위한 Score Matching 학습

낭철성, 곽봉, 김동국

전남대학교

langtiecheng@gmail.com, bongkua0379@naver.com, dkim@jnu.ac.kr

Score Matching Training for Kernel-based Restricted Boltzmann Machine

Lang Tie Cheng, Guo Peng and Dong Kook Kim

Chonnam National University

요 약

본 논문에서는 커널기반 제한된볼츠만머신(KRBM)에 대해 score matching(SM)을 사용한 학습 기법을 제안한다. KRBM에 대한 확률분포함수를 이용하여 SM 목적함수를 정의하고 이를 KRBM의 파라메터를 추정하는데 이용하였다. 제안된 KRBM에 대한 SM 학습 기법을 평가하기 위해 CIFAR10 데이터 셋에 대해 특징학습을 수행하였다. 실험결과 제안된 기법이 기존 Gaussian-Bernoulli RBM에 비해 더 나은 인식성능을 나타내었다.

Ⅰ. 서론

최근 기계학습 분야에서 에너지 기반 모델(energy-based models, EBMs)에 대한 연구가 활발하게 진행되고 있다[1]-[4]. EBM은 에너지 함수를 통해 확률밀도함수를 정의하기 위해 사용되는 확률적인 모델이다. EBM중에 가장 대표적인 것은 제한된볼츠만머신(restricted Boltzmann machine, RBM)으로 특징학습, 밀도추정, 영상 생성 등 다양한 분야에 사용되고 있다[5]. EBM을 학습하기 위한 방법으로는 최적 이론적 성질 때문에 최대우도비 (maximum-likelihood, ML)기법이 가장 많이 사용되고 있다[5]. 그러나 ML기법은 EBM에 존재하는 파티션 함수(partition function) 때문에 학습시 계산적으로 많이 요구되는 단점을 가지고 있다. 따라서 파티션 함수를 계산하지 않고 EBM을 학습하기 위한 다양한 기법들이 제안되었다[1]-[4]. Score matching(SM)은 EBM에서 ML 학습의 단점을 극복하기 위해 제한된 기법이다[1]. SM의 목적함수는 입력에 대해 모델과 데이터 분포의 로그 미분값의 차이에 L2 loss을 적용하여 정의된다. 이 기법은 특히 심층 EBM을 학습하는데 ML에 비해 계산적으로 효과적임을 나타내었다.

본 논문에서는 EBM중에 하나인 커널기반 RBM(kernel-based RBM, KRBM)[6]을 위한 SM 학습을 제안한다. KRBM은 기존 RBM과 다르게 입력 벡터를 비선형 함수에 의해 고차원의 특징공간으로 사상한 다음, 이 특징 공간에서 가시층과 은닉층을 통해 데이터를 모델링하는 기법이다. 이 논문에서는 이러한 KRBM을 학습하기 위한 SM을 통한 파라메터 갱신식을 유도하고 이를 통한 학습을 제안한다. 제안된 학습 기법의 성능을 검증하기 위해 기존 RBM의 SM 학습과 비교하였다. 실험 결과 CIFAR10 데이터 셋에 대해 특징학습을 진행한 결과 기존 RBM에 비해 제안된 KRBM의 SM 학습이 더 높은 인식성능을 나타내었다.

본 논문의 Ⅱ장에서는 KRBM을 간단히 소개하고, KRBM에 대한 SM 기법을 제시한다. Ⅲ장에서는 실험과 결과에 대해 나타내고, Ⅳ에서는 결론을 맺는다.

Ⅱ. 본론

2.1 KRBM

KRBM은 최근 제안된 모델로 커널기법을 이용한 EBM중의 하나의 모델이다[6]. 먼저 $\boldsymbol{v}$은 n차원의 데이터 공간에서 정의된 가시변수 벡터이다. KRBM은 비선형함수 $\phi : R^n \rightarrow R^f$을 사용해 n차원의 입력공간에서 f차원의 특징벡터 공간으로 사상한다. $\phi(\boldsymbol{v})$는 비선형적으로 사상되는 f차원의 특징벡터 공간에서 가시변수이다. KRBM 구조는 $\phi(\boldsymbol{v})$에 의한 가시유닛 층과 m차원 실수 벡터 $\boldsymbol{h}$에 의한 은닉유닛 층으로 구성된다. 두 층사이의 연결 가중치를 위해 m개의 n차원의 가중치 벡터 $(\boldsymbol{w}_1, \dots, \boldsymbol{w}_m)$들을 정의하고, 이를 비선형 함수에 의해 사상된 가중치 행렬, $\Phi(W) = (\phi(\boldsymbol{w}_1), \dots, \phi(\boldsymbol{w}_m))$을 구성한다. 그러면 $\phi(\boldsymbol{w}_j)$는 j번째 은닉유닛 h_j와 특징공간에서 가시벡터 $\phi(\boldsymbol{v})$을 연결하는 가중치 벡터가 된다. 이 때기존의 EBM과 비슷하게 KRBM은 특징공간에서 다음과 같은 결합확률분포를 갖는다[6].

$$p(\phi(\boldsymbol{v}), \boldsymbol{h}) = \frac{1}{Z} e^{-\frac{1}{2} E(\phi(\boldsymbol{v}), \boldsymbol{h})} \tag{1}$$

위 식의 지수항은 모델의 에너지 함수이며 아래와 같다.

$$E(\phi(\boldsymbol{v}), \boldsymbol{h}) = k(\boldsymbol{v}, \boldsymbol{v}) - 2\sum_{j=1}^{m} h_j k(\boldsymbol{w}_j, \boldsymbol{v}) + \sum_{j=1}^{m} h_j^2 \tag{2}$$

여기서 $k(,)$은 커널 함수이며, 커널 트릭[7]에 의해 $k(v,v) = \phi(v)^T \phi(v)$이고, $k(w_j, v) = \phi(w_j)^T \phi(v)$,이다. 그리고 Z는 파티션 함수이다. 커널함수로 ReLU 함수가 보통 사용된다. 위 결합확률 분포로부터 식을 유도하면, 한계 확률분포는 $p(\phi(\boldsymbol{v})) = \frac{1}{Z} \exp(-\frac{1}{2} F(\phi(\boldsymbol{v})))$을 얻을 수 있다. 이 때 한계 확률분포의 에너지 함수 $F(\phi(\boldsymbol{v}))$는 다음과 같다.

$$F(\phi(\boldsymbol{v})) = (k(\boldsymbol{v}, \boldsymbol{v}) - \sum_{j=1}^{m} k(\boldsymbol{w}_j, \boldsymbol{v})^2) \tag{3}$$

기존의 KRBM을 학습하기 위해 $p(\phi(\boldsymbol{v}))$의 로그 유사도 (log-likelihood) 함수를 최대화하는 경사 승강법(gradient ascent) 기반 알고리즘을 이용한다. 이러한 알고리즘은 파티션 함수를 계산하기 위해

근사적인 방법과 많은 계산량을 요구하는 샘플링 기법을 요구하는 단점을 갖고 있다.

2.2 KRBM을 위한 SM 학습

이 단원에서 KRBM을 학습하기 위한 SM 기법을 제시한다. SM은 EBM에서 학습을 수행할 때 파티션 함수 Z의 계산의 어려움을 극복하기 위해 제시되었다[1]. $p_{data}(\boldsymbol{v})$은 N개의 한정된 iid 데이터 샘플 $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N\}$을 생성하는 알려지지 않은 데이터 분포라 하자. 그러면 SM의 목표 함수는 데이터 분포 $p_{data}(\boldsymbol{v})$와 모델 분포 $p_\theta(\boldsymbol{v})$의 사이에 다음과 같은 목적함수를 파라메터 θ에 대해 최소화하여 파라메터를 추정하는 기법이다[1]-[4].

$$J(\theta) = E_{p_{data}(\boldsymbol{v})}\left[\frac{1}{2}\| \nabla_{\boldsymbol{v}} \log p_{data}(\boldsymbol{v}) - \nabla_{\boldsymbol{v}} \log p_\theta(\boldsymbol{v})\|^2\right]$$

하지만 위식을 최적화 하는데는 $p_{data}(\boldsymbol{v})$에 대한 의존성 때문에 $J(\theta)$을 실제로 처리하기 어렵다. 그러나 어떤 정규적 조건하에서 데이터 분포의 기대치를 훈련 샘플의 경험 평균치로 대체함으로써 다음과 같이 쓸 수 있다.

$$J = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}\|\nabla_{\boldsymbol{v}_i} \log p_\theta(\boldsymbol{v}_i)\|_2^2 + tr(\nabla_{\boldsymbol{v}_i}^2 \log p_\theta(\boldsymbol{v}_i))\right] \quad (4)$$

위 식은 SM을 이용한 EBM을 학습하기 위해 사용된 목적함수로 다양한 EBM에 적용이 가능하다. 본 논문에서는 위의 SM 목적함수를 KRBM에 적용하여 학습하는 기법을 제안한다. 이를 위해 식 (4)의 모델 분포 $p_\theta(\boldsymbol{v})$ 대신에 KRBM의 한계 확률분포 $p(\phi(\boldsymbol{v}))$로 대체하여 목적함수를 구성할 수 있다. 그러면 최종적인 KRBM을 위한 목적함수는 식(3)에 기반하여 다음과 같다.

$$J = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}\|\nabla_{\boldsymbol{v}_i} F(\phi(\boldsymbol{v}_i))\|_2^2 + tr(\nabla_{\boldsymbol{v}_i}^2 F(\phi(\boldsymbol{v}_i)))\right] \quad (5)$$

KRBM에 대한 SM에 의한 학습은 위 목적함수를 경사하강법을 이용하여 파라메터를 반복적으로 갱신함으로 추정할 수 있다.

Ⅲ. 실험 및 결과

본 논문에서 제안된 기법을 평가하기 위해 CIFAR10 데이터 셋을 이용한 특징학습을 수행하였다. CIFAR10 데이터는 32*32*3 크기의 컬러 영상으로 10가지 클래스를 포함한다. 학습과 테스트를 수행하기 전에 CIFAR10 데이터를 전처리하여 사용하였다. 각 이미지로부터 6*6*3 patch을 임의적으로 추출하여 입력으로 사용하였고, whitening과 standard normalization을 수행하였다[8]. SM을 사용해 KRBM을 학습할 때 학습률은 0.00001, momentum은 0.9을 사용하였다. 테스트를 위해서는 입력 영상에 대해 같은 간격으로 patch을 생성하여 KRBM을 통해 특징을 추출하였다[8]. SM을 이용한 다른 EBM 모델을 비교하기 위해 Gaussian-Bernoulli RBM (GBRBM)[5]을 사용하였다. 각각의 모델에 대해 은닉수에 따른 성능 변화를 알아보기 위해 다양한 크기의 은닉 노드수를 사용해 실험을 수행하였다. 각 특징들을 추출한 후 인식을 위해 softmax 분류기를 사용하였다. 그림1. 은 가변적인 은닉유닛의 수에 따른 테스트 데이터에 대한 인식 정확도를 나타낸다. 그림에 나타나듯이 은닉노드수에 따라 점진적으로 KRBM은 성능향상이 이루어진 것에 비해 GBRBM은 2048개의 은닉 노드수에서 성능이 하락하였다. 그리고 모든 은닉수에 대해서 KRBM이 GBRBM에 비해 더 뛰어난 성능을 나타남을 알 수 있었다. 위 실험으로부터 SM기법을 이용한 GBRBM에 비해 더 KRBM이 효과적임을 알 수 있다. 하지만 본 실험 결과에 나타나지 않았지만 기존의 KRBM을 학습하기 위한 근사적인 ML 방법은 SM 기법에 비해 더 높은 성능을 나타내었다. 이러한 결과에 대한 분석은 앞으로 연구가 더 필요하다.

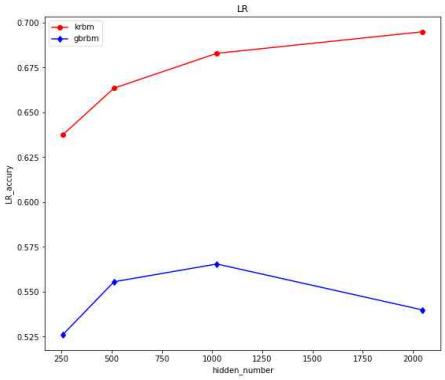

그림 1. CIFAR10에서 가변적인 은닉 노드수에 따른 테스트 인식 정확도

Ⅳ. 결론

본 논문에서는 EBM중에 하나인 KRBM의 파라메터를 추정하기 위해 SM을 이용한 학습 기법을 제안하였다. KRBM의 확률분포함수을 이용하여 SM기반의 목적함수를 유도하였고 이를 경사하강법을 통해 파라메터를 학습하였다. CIFAR10 데이터을 이용한 특징학습에 있어서 제안된 SM기반 KRBM이 기존의 GBRBM보다 더 높은 인식성능을 나타내었다.

현재 연구는 SM 기법을 한층의 EBM에 적용한 결과로 향후 연구는 심층 구조를 갖는 다양한 EBM에 SM 기법을 적용하는 것이 필요하다.

참 고 문 헌

[1] Swersky, K., et al., "On autoencoders and score matching for energy based models." Proceedings of the 28th international conference on machine learning (ICML-11). 2011.

[2] Song Y., and Ermon S., "Generative modeling by estimating gradients of the data distribution." Advances in Neural Information Processing Systems 32, 2019.

[3] Kingma, D. P., "Improving Score Matching for learning statistical models of natural images." PhD Diss.. New York University, 2010.

[4] Song Y., and Kingma D. P.. "How to train your energy-based models." arXiv preprint arXiv:2101.03288, 2021.

[5] Hinton, G. E., "A practical guide to training restricted Boltzmann machines." Neural networks: Tricks of the trade. Springer, Berlin, Heidelberg, 599-619, 2012.

[6] 김동국, and 신종원. "비지도 특징학습을 위한 커널 기반 제한된 볼츠만 머신." 한국통신학회논문지 44(9), 1633-1640, 2019.

[7] Schölkopf B. and Smola A.. *Learning with kernels*. MIT press, 2002.

[8] Coates A., Ng A. and Lee H.. "An analysis of single-layer networks in unsupervised feature learning." Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011.