💚

VAE: Auto-Encoding Variational Bayes

기간	@05/07/2025 → 05/13/2025
주차	10주차
논문	https://arxiv.org/pdf/1312.6114
상태	완료
예습/복습	예습과제
참고 자료	참고 블로그1 참고 블로그2

0. Abstract
1. Introduction
    논문이 다루는 분야
    해당 task에서 기존 연구 한계점
    논문의 contributions
2. Method
    2.1 Problem scenario(문제 시나리오)
    2.2 The variational bound
    2.3 The SGVB estimator and AEVB algorithm
    2.4 The reparameterization trick
3. Example : Variational Auto-Encoder
4. Related work
5. Experiments
    Dataset
    Baseline
    결과
6. Conclusion & Future work
    Conclusion (결론)
    Future work

0. Abstract

- 연속적인 잠재변수(latent variable) 포함하고 사후 분포가 계산 불가능한 경우와 대규모 데이터셋을 다루는 상황에서, 효율적인 추론 및 학습을 어떻게 수행할 수 있을지 질문함.

- 두 가지 주요 contribution
    1. **variational lower bound의 reparameterization** 통해, 일반적인 확률적 경사 하강법으로 간단히 최적화할 수 있는 하한 추정기(estimator) 도출
    2. **독립 동일 분포(i.i.d.)된 데이터셋**에서, 데이터마다 연속적인 잠재변수가 존재하는 경우, **근사 추론 모델(recognition model)** 학습하여 실제 사후 분포 근사하는 방식으로 추론을 매우 효율화할 수 있음을 보여줌.

1. Introduction

논문이 다루는 분야

- **연속적인 잠재 변수** 또는 파라미터를 포함하고, **사후 분포가 계산 불가능**한 확률적 생성 모델(directed probabilistic models)에서 어떻게 효율적인 근사 추론과 학습 수행할 수 있는가?
    - **변분 베이즈(VB) 접근법** → 이러한 비가역적인 사후 분포에 대한 근사를 최적화하는 방식.

해당 task에서 기존 연구 한계점

- 기존의 Mean-Field 방식
    - 근사 분포에 대한 기댓값을 **해석적으로 계산**해야 하므로 일반적인 경우 적용이 어려움.

논문의 contributions

- variational lower bound(변분 하한)의 reparameterization(재파라미터화)를 통해, 단순하고 미분 가능한 unbiased estimator 도출할 수 있음.
    - 이 **SGVB** 추정기(Stochastic Gradient Variational Bayes estimator)는 거의 모든 연속형 잠재 변수를 포함한 모델에서 표준 확률적 경사 상승법으로 최적화 가능함.
- i.i.d. 데이터셋 & 각 데이터에 대해 연속형 잠재 변수가 있는 경우, **Auto-Encoding Variational Bayes(AEVB)** 알고리즘 제안
    - SGVB 추정기 사용하면 recognition model 최적화함.
    - 단순한 샘플링 방식으로 매우 효율적인 사후 추론 가능하게 하며, 복잡한 반복적 추론 과정(MCMC 등) 대체할 수 있음.
- 이렇게 학습된 recognition model은 인식, 복원, 표현, 시각화 등 다양한 작업 활용 가능.
    - 이 recognition model로 신경망 사용 → **Variational Auto-Encoder(VAE)**

2. Method

- 연속형 잠재 변수를 가진 다양한 directed graphical models에 대해 변분 하한 추정기(확률적 목적 함수) 도출하는 데 사용되는 전략.
- **i.i.d. 데이터셋**을 전제로 하며, **최대우도(MLE) 또는 사후 최대 추정(MAP)**을 모델 파라미터에 대해 수행하고, **변분 추론**을 잠재 변수에 대해 수행하는 상황 다룸.

2.1 Problem scenario(문제 시나리오)

👉 데이터셋 $X = \{x^{(i)}\}_{i=1}^{N}$: N개의 i.i.d. 샘플로 이루어짐.

- 이 데이터가 관측되지 않은 연속형 잠재 변수 z에 의해 생성되었다고 가정

<생성 과정>

- $z^{(i)} \sim p_\theta(z)$: 랜덤 변수 z는 사전 분포에 의해 생성
- $x^{(i)} \sim p_\theta(x|z^{(i)})$: data x는 조건분포에 의해 생성

⇒ 실제로는 $z^{(i)}$와 θ 알 수 없으며, 사후 분포 $p_\theta(x|z^{(i)})$는 비가역적인 경우가 많음.

- 위 분포들을 계산하는 것은 기존의 방법으로 매우 다루기 어려움!
  - **Intractability**
    - marginal likelihood 수식 $p_\theta(x) = \int p_\theta(z)p_\theta(x|z)$
      - data x가 p_θ function으로부터 나와야 하지만, 해당 수식은 정의되지 않은 분포 z 이용하기 때문에 추정하기 매우 어려움. (MAP, ML(maximum likelihood) 사용하기 힘듦.)
  - **A large dataset**
    - sampling 기반으로 하는 Monte Carlo EM 등은 매우 속도가 느릴 수밖에 없음.
- 본 논문은 다음 3가지 문제를 해결하고자 함.
  1. **효율적인 MLE/MAP 추정**
     - 모델의 파라미터를 추정하여 실제 데이터를 잘 모사하는 생성 모델 학습
  2. **효율적인 사후 추론**
     - 관측값 x에 대해 잠재변수 z를 근사적으로 추정
  3. **효율적인 주변 추론**
     - x에 대한 사전 분포를 통해 이미지 복원, 초해상도 등의 작업 수행

  ⇒ 이를 위해 Recognitin Model $q_\phi(z|x)$ 도입
  - 비가역적인 사후 분포 $p_\theta(z|x)$의 근사치로서 작동함.
  - 확률적 인코더(encoder) 역할 하며, x가 주어졌을 때 가능한 z의 분포를 출력함.
  - 반대로 $p_\theta(x|z)$ → 확률적 디코더(decoder)로 작동하며, 주어진 z로부터 x의 분포를 생성함.

2.2 The variational bound

- **The marginal likelihood** → 각각의 data point마다 계산

$$logp_\theta(x^{(i)}) = D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z|x^{(i)})) + \mathcal{L}(\theta, \phi, x^{(i)})$$

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) \geq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[-\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) + \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})\right] \quad (2)$$

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})\right] \quad (3)$$

- ELBO(=second RHS term→ variational lower bound) → KL-divergence와 사후 분포에 대한 기댓값에 대한 식
  - 여기서 KL-divergence를 직접 구하는 것은 힘들기 때문에 사후 분포에 대한 값 maximize해서 latent variable z로 부터 x가 나올 기댓값을 크게 만듦.(최대화)

  ⇒ **variation lower bound 최대화하는 방향**으로 optimization 수행!

2.3 The SGVB estimator and AEVB algorithm

- Lower bound를 최대화하는 방식으로 학습 이루어지기 위해 → backpropagation 수행
  - lower bound 계산 시 z는 sampling되어 추출되므로 parameter ϕ는 backpropagation이 진행되지 않음. ⇒ 이를 해결하기 위한 것 = **Reparameterization trick!**
- **SGVB(Stochastic Gradient Variational Bayes) 추정기** 제안
  - θ, ϕ에 대한 미분이 가능한 추정기 구성할 수 있음.
- 핵심 아이디어
  - 재매변수화(reparameterization) 기법

$$\tilde{z} = g_\phi(\epsilon, x),\ \epsilon \sim p(\epsilon)$$

    - 미분가능한 변환함수 $g_\phi(\epsilon, x)$ 새로 정의
    - 새롭게 정의되는 latent variable z → $g_\phi(\epsilon, x)$

$$\mathbb{E}_{q_\phi(z|x^{(i)})}[f(z)] \approx \frac{1}{L}\sum_{l=1}^{L} f(g_\phi(\epsilon^{(l)}, x^{(i)})) \quad (5)$$

    - Monte Carlo estimates of expections 이용하여 g_ϕ에서 $f(z)$에 대한 기댓값 계산

$$\widetilde{\mathcal{L}}^B(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)}) || p_{\boldsymbol{\theta}}(\mathbf{z})) + \frac{1}{L}\sum_{l=1}^{L}(\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)}))$$
$$\text{where} \quad \mathbf{z}^{(i,l)} = g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}^{(i,l)}, \mathbf{x}^{(i)}) \quad \text{and} \quad \boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon}) \qquad (7)$$

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}) \simeq \widetilde{\mathcal{L}}^M(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}^M) = \frac{N}{M}\sum_{i=1}^{M}\widetilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) \qquad (8)$$

    - SGVB에서 lower bound 구하면 (7) 식 도출 (식 (5)를 대입한 것.)
    - backpropagation 가능해짐!!
      - 식 (3) : $z \sim q_\phi(z \mid x)$이기 때문에, 샘플링 연산이 **non-differentiable**하여 ϕ 로의 gradient 흐름이 차단됨.
      - 식 (7) : z를 $g_\phi(\epsilon, x)$로 치환함으로써, stochastic한 샘플링은 $\epsilon \sim p(\epsilon)$에서만 발생하고, z는 ϕ와 x에 대한 **미분 가능한 함수**가 되므로 backpropagation이 가능해짐.

> ⇒ 정리: 미분가능한 transformation function $g_\phi(\epsilon, x)$ 만들어서 $z = g_\phi(\epsilon, x)$ 수식 만들면 gradient 연산 가능해짐→ 찾고자 하는 $q_\phi(z \mid x)$ 이용하

> 지 않아도 ϕ에 대해 학습 가능!!

Algorithm 1 Minibatch version of the Auto-Encoding VB (AEVB) algorithm. Either of the two SGVB estimators in section 2.3 can be used. We use settings $M = 100$ and $L = 1$ in experiments.

$\boldsymbol{\theta}, \boldsymbol{\phi} \leftarrow$ Initialize parameters
repeat
  $\mathbf{X}^M \leftarrow$ Random minibatch of M datapoints (drawn from full dataset)
  $\boldsymbol{\epsilon} \leftarrow$ Random samples from noise distribution $p(\boldsymbol{\epsilon})$
  $\mathbf{g} \leftarrow \nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \widetilde{\mathcal{L}}^M(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}^M, \boldsymbol{\epsilon})$ (Gradients of minibatch estimator (8))
  $\boldsymbol{\theta}, \boldsymbol{\phi} \leftarrow$ Update parameters using gradients $\mathbf{g}$ (e.g. SGD or Adagrad [DHS10])
until convergence of parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$
return $\boldsymbol{\theta}, \boldsymbol{\phi}$

2.4 The reparameterization trick

- 앞서 설명한 핵심 아이디어와 동일.
- 어떤 $q_\phi(z \mid x)$에 대해 재매개변수화가 가능한가?
  1. 역 누적분포함수(inverse CDF)가 존재하는 경우: Exponential, Logistic, Gumbel 등
  2. location-scale family: 평균과 분산 조절만으로 샘플 생성 가능한 분포들(Gaussian, Laplace 등)
  3. 조합적 구조 가진 분포: Log-normal, Gamma, Bata 등

  ⇒ 위에 해당하지 않더라도, 근사적인 inverse CDF 사용하는 방법 존재

3. Example : Variational Auto-Encoder

- **VAE의 실제 예시 구현**
  - 핵심 아이디어 : 확률적 인코더 $q_\phi(z|x)$와 디코더 $p_\theta(x|z)$를 신경망(MLP)으로 구현하고, AEVB 알고리즘을 통해 최적화하는 것!
- **모델 구성(architecture)**
  - Prior $p_\theta(z)$: 평균 0, 단위분산 가지는 정규분포 $N(0, I)$
  - Decoder $p_\theta(x|z)$
    - 데이터가 실수형인 경우 → 다변량 Gaussian
    - 데이터가 이진(binary)인 경우 → Bernoulli 분포
    - 디코더는 신경망(MLP)으로 구성되어, z를 입력으로 받아 분포의 파라미터(평균, 분산 또는 확률) 출력함.
  - Encoder $q_\phi(z|x)$
    - 정규분포 $N(\mu(x), \sigma^2(x))$
    - $\mu(x), \sigma(x)$는 입력 x에 대해 인코딩된 값으로, 인코더 MLP의 출력
- **샘플링 방법**
  - 재매개변수화 기법 사용

$$z^{(i,l)} = \mu^{(i)} + \sigma^{(i)} \odot \epsilon^{(l)} \sim N(0, I)$$

        - 이 z를 디코더에 넣어 $p_\theta(x|z)$를 계산
- **최종 학습 목적 함수**
    - KL 발산은 해석적으로 계산 가능(정규분포 간 KL)
    - ELBO는 다음과 같이 표현

$$L(\theta, \phi; x(i)) \approx \frac{1}{2}\sum_{j=1}^{J}(1 + log((\sigma_j^{(i)})^2) - (\mu_j^{(i)})^2 - (\sigma_j^{(i)})^2) + \frac{1}{L}\sum_{l=1}^{L} logp_\theta(x^{(i)}|z^{(i,l)})$$

        - 첫 번째 항: KL 발산(정규화 효과)
        - 두 번째 항: 재구성(reconstruction) 오차(재구성 확률의 log)

⇒ VAE는 **인코더 → 추론 모델, 디코더 → 생성 모델의 역할**, 두 모델이 함께 학습됨.

4. Related work

- **Wake-Sleep 알고리즘**
    - 연속형 잠재변수를 가진 일반적인 클래스의 모델에 적용 가능한 유일한 온라인 학습 방법.
    - VAE와 유사하게 사후분포를 근사하는 recognition model 사용함.
        - But, 두 개의 목적 함수를 동시에 최적해야 함 → marginal likelihood 하한을 직접 최적화하는 것이 아니라는 점에서 이론적 문제!
    - 장점
        - 이산형 잠재변수 모델에도 적용 가능함.
        - per datapoint 기준, 계산 복잡도 AEVB와 동일함.
- **Stochastic Variational Inference(SVI)**
    - 단순한 gradient estimator의 고분산 문제를 제어 변수(control variate)를 통해 줄이는 기법 제안
        - exponential family의 근사 분포에 적용
    - 다양한 control variate 기반 일반 기법들 제안
        - 본 논문과 유사한 reparameterization 기법 → exponential family 근사 분포 학습에 효율적으로 적용함.
- **AEVB 알고리즘**
    - variational objective로 훈련되는 **directed probabilistic models와 auto-encoders 사이의 연결성**
        - 선형 auto-encoder와 특정 선형-가우시안 생성 모델 간 연결성은 오래 전부터 알려짐.
            - ex) PCA가 특정 조건의 선형-가우시안 모델의 최대우도 해.
- **Auto-encoder**

- unregularized autoencoders의 학습 기준이 mutual information의 하한을 최대화하는 것임을 보임.
    - autoencoding 모델 하에서 데이터의 log-likelihood를 최대화하는 것과 유사.
    - But, reconstruction error 자체는 유용한 표현 학습에 충분하지 않음.

      → denoising, contractive, sparse 등 여러 regularization 기법 제안.
- PSD (Predictive Sparse Decomposition), Generative Stochastic Networks, Deep Boltzmann Machines 학습에서 recognition model을 활용한 연구 등

  ⇒ 유사한 아이디어 가졌으나, 보통 **unnormalized model**(ex: Boltzmann Machines)에 적용되거나 **sparse coding**에 제한된 방식
- DARN
  - auto-encoding 구조 사용, directed model 학습
  - But, 이진 잠재변수에만 적용 가능함.

5. Experiments

Dataset

- **MNIST 이미지**
- **Frey Face 데이터셋**
- 다양한 알고리즘을 ELBO 및 margianl likelihood 측면에서 비교함.
  - 인코더와 디코더는 동일한 수의 은닉 노드 가짐.
  - Frey Face 데이터 → 연속형 데이터이므로 Gaussian 출력 디코더 사용, 출력을 (0,1)로 제한하기 위해 시그모이드 함수를 사용함.

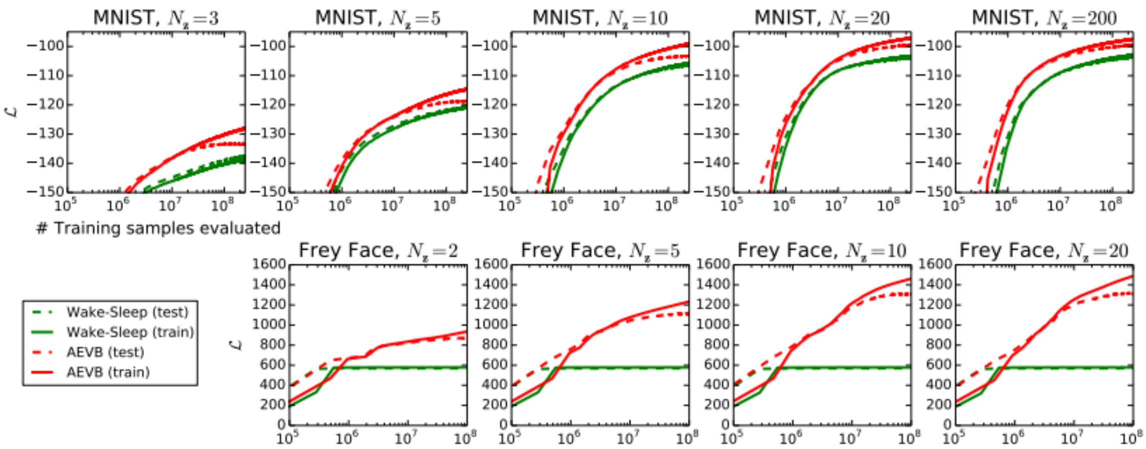
Baseline

- **Wake-Sleep과의 비교**
  - AEVB와 Wake-Sleep을 비교하기 위해 동일한 인코더 구조를 사용함.
  - 모든 파라미터는 $N(0, 0.01)$에서 초기화, Adagrad로 학습됨.

1. step size : {0.01, 0.02, 0.1} 중에서 학습 초반 성능 기준으로 선택
2. 미니배치 크기 : 100
3. 데이터당 샘플 수 $L = 1$

결과

- **Variational Lower Bound**

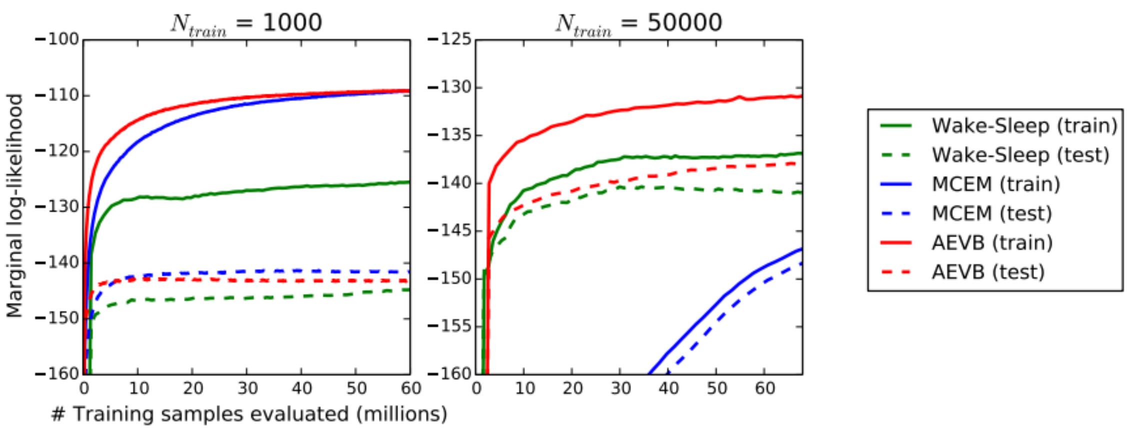


- MNIST : 은닉 노드 500개
- Frey Face : 은닉 노드 200개(과적합 방지)

⇒ 결과

- AEVB가 Wake-Sleep보다 **빠르게 수렴, 더 좋은 하한값 도달**
- 잠재공간 차원이 커져도 **overfitting 되지 않음** → ELBO 정규화 효과 때문.

- **Marginal Likelihood**



- 잠재 공간 차원이 작은 경우(3차원)에는 HMC 기반의 마진 우도 추정기 사용할 수 있음.
- MNIST에 대해 100 hidden unit, latent dim 3 사용
- 비교 대상 : AEVB, Wake-Sleep, Monte Carlo EM

⇒ 결과 : AEVB는 효율성과 정확성 측면 모두에서 우수함.

6. Conclusion & Future work

Conclusion (결론)

- 연속형 잠재변수 모델에 대한 효율적인 근사 추론을 위한 새로운 추정기 **'Stochastic Gradient Variational Bayes (SGVB)** 제안함.

- 이 추정기는 표준 확률적 경사법을 통해 간단하게 최적화할 수 있으며, i.i.d. 데이터셋 및 연속형 잠재 변수 갖는 경우, **Auto-Encoding Variational Bayes(AEVB)** 알고리즘 통해 recognition model 학습할 수 있음.

Future work

- SGVB 추정기와 AEVB 알고리즘은 **거의 모든 연속형 잠재변수 기반의 추론 및 학습 문제에 적용**될 수 있으므로, 다음과 같은 다양한 확장 방향 존재.
  1. 딥 신경망(ex: CNN) 사용하여 encoder/decoder 구조를 깊게 만들고 AEVB로 학습하는 계층적 생성 모델
  2. 시계열 모델(ex: 동적 베이즈 네트워크)에 적용
  3. SGVB를 모델 파라미터(global 변수) 추론에 적용
  4. supervised 학습 모델에서의 잠재변수 학습 → 복잡한 잡음 분포 다루는 데 유용