고급 컴퓨터 알고리즘 발표자료

G201949006 박건주

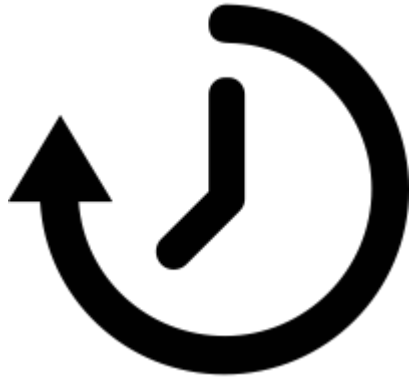
목차

1. 딥러닝 신경망에서의 전이 학습
2. 딥러닝 생성 모델이란?
3. 제한적 볼츠만 머신(RBM)

CNN에서의 딥러닝 신경망 전이 학습

1000개 이상의 이미지 클래스를 분류하는 CNN 네트워크 학습?

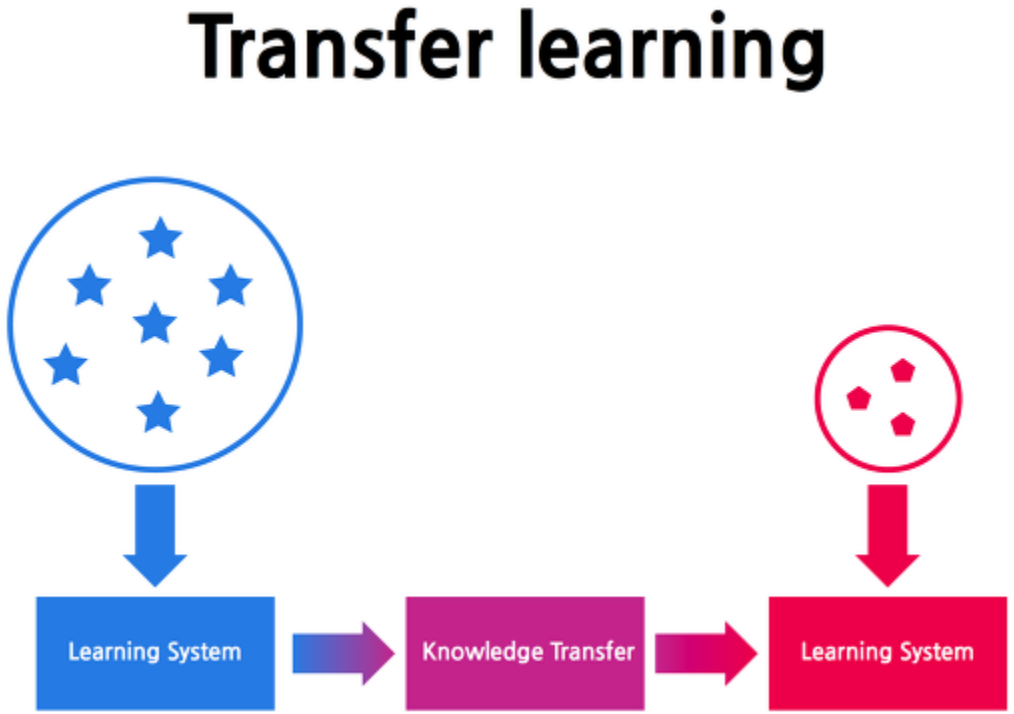
CNN에서의 딥러닝 신경망 전이 학습



너무 많은 양의 학습데이터 요구

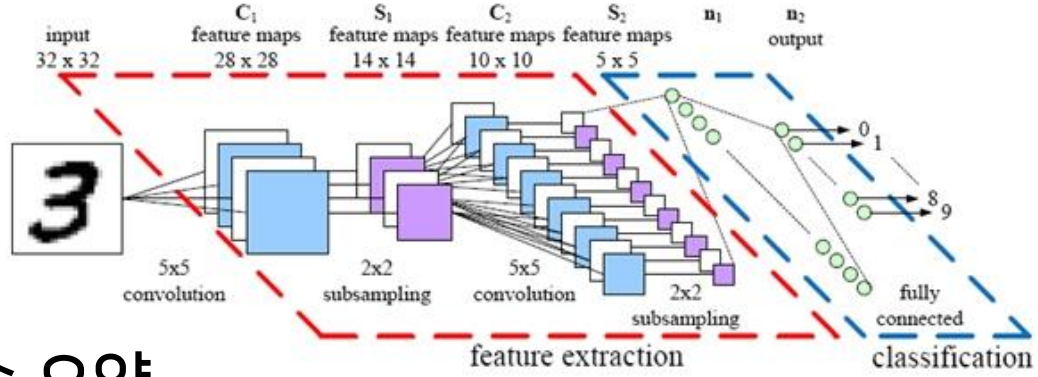
제한된 컴퓨터 자원으로 인한
상당한 학습 시간

CNN에서의 딥러닝 신경망 전이 학습



Pretrained 된 CNN 모델 중, Feature Extraction 레이어들이 가진 지식(Knowledge)를 활용하여 새로운 이미지 클래스 분류에 활용!

CNN에서의 딥러닝 신경망 전이 학습



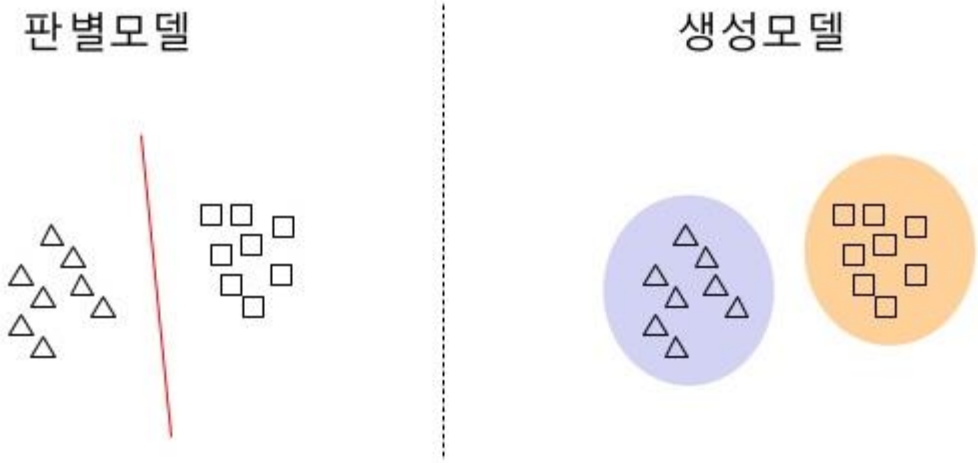
CNN 전이학습 요약

CNN은 크게 2개의 서브네트워크로 나뉩니다.

1. Feature Extractor (특징 추출)
   - 입력 이미지를 설명하기 위한 가장 핵심적인 특징이 어떤 것인지 뽑아내는 역할
   - 주로 Conv, ReLU, BatchNorm, Pooling 등
   - **이 부분은 이미 ImageNet 학습 데이터셋을 통해 훌륭하게 학습된 모델들이 많이 존재함 -> 그대로 사용**

2. Image Classifier (이미지 클래스 분류기)
   - 위에서 뽑아낸 특징들을 검토하여 해당 특징들이 어떤 클래스로 분류될 지 결정하는 서브네트워크
   - 주로 FC(Inner Product) 레이어로 구현
   - **새로운 이미지에 대해 분류기만 학습시키면 되기 때문에 시간이 적게 소요됨!**

딥러닝(머신러닝) 모델



판별모델 (discriminative model)
관심대상 = **입력에 대한 출력값 결정**
입력 x에 대해 출력 y의 **조건부 확률분포 형태**
$p(y|x)$, $y^* = argmax_y\, p(y|x)$

- SVM, Decision tree, CNN 등

생성모델 (generative model)
관심대상 = **학습(입력)데이터를 어떻게 표현할지에 대한 정보**
입력 x, 출력 y에 대해 **결합확률분포 형태**
$p(x,y)$ [분류,회귀] 또는 $p(x)$ [군집화]

- 확률그래프모델, RBM, DBM, GAN 등

RBM(제한적 볼츠만 머신) 개념

RBM의 배경

1. 기존 MLP 분류 문제 = **조건부 확률 p(y|x)**를 구하는 문제
   - 해당 입력의 Feature들을 분석하여
     어떤 출력이 가장 적합할지 여부에 집중
   - 은닉층의 가중치들은 출력 값을 결정하는데 사용되는 여러 가지의 Feature들로부터 클래스를 잘 판별하기 위한 방향으로 학습됨

2. 데이터의 확률을 모델링해보자! = **결합확률 p(x,y) 또는 p(x)**를 구하는 문제
   - 입력으로 준 데이터를 분석하여
     해당 입력과 가장 근접한 데이터를 출력함
   - 은닉층의 가중치는 입력 데이터를 설명하기 위한 특정한 Feature들을 뽑아내기위한 방향으로 학습된다.
     즉, 특정 Feature들을 입력으로 받을 경우 입력데이터와 유사한 데이터가 만들어지며, 입력데이터를 입력으로 받을 경우, Feature들을 출력한다.

     무슨 말이지?????

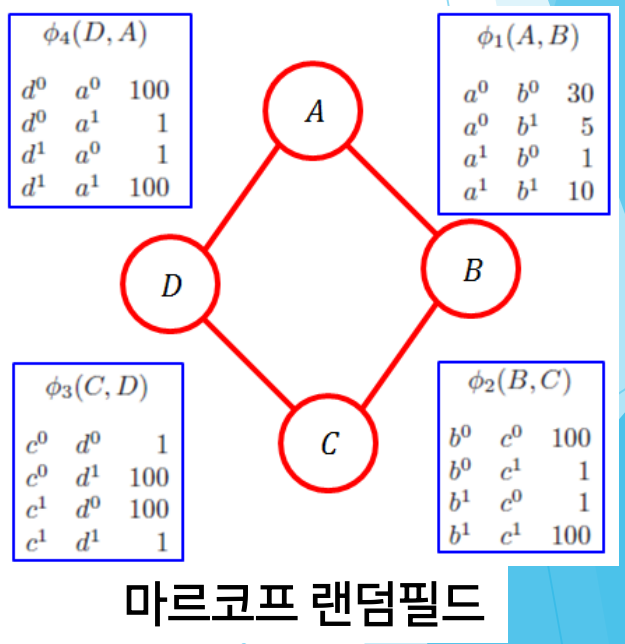
RBM(제한적 볼츠만 머신) 개념

RBM은 확률 그래프 모델을 기반으로 구현된다.

확률 그래프 모델이란 확률변수에 대응하는 **노드**(node)와 확률변수 간의 관계를 나타내는 **간선**(edge)으로 구성된 **그래프**를 사용하여 **결합 확률분포**(joint probability distribution)를 표현하는 모델이다.

확률 그래프 모델 중 **마르코프 랜덤 필드**는 확률 분포를 정의할 때, 각 확률 변수들간의 관계에 대해 **정의된 함수**를 사용한다. (이 때 정의된 함수는 Affinity 또는 Factor라고 칭함)

제한적 볼츠만 머신(RBM)은 이러한 Factor를 통계 물리학의 에너지 개념을 이용한 **에너지 함수(Energy Function)**으로 정의하여 에너지 함수를 구성하는 가중치를 학습시키는 방식이다.



마르코프 랜덤필드

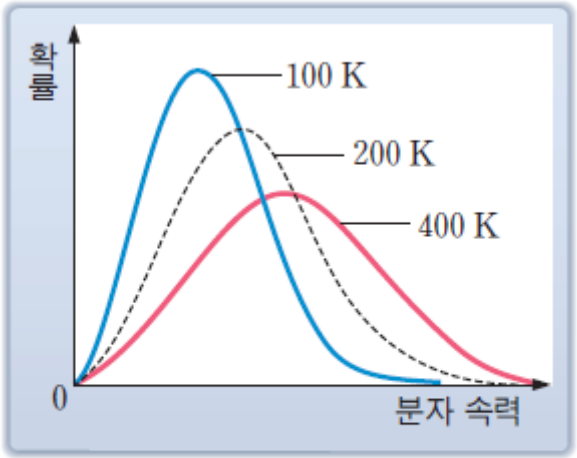
RBM(제한적 볼츠만 머신) 개념

열 역학, 통계 물리학에서의 에너지 함수란?

간단히 설명해서 에너지가 높아질 수록 특정 물질이
그 상태에 머무를 확률 값이 낮아지는 공식을 뜻한다.

예시로, 특정온도에 특정한 성질(분자속력)을 가지는 분자가 있을 확률을 측정할 때에 온도(에너지)가 낮을 수록 그 상태에 머무르는 분자가 존재할 확률이 높아진다는 열역학 법칙이 있다.

결국, 이러한 에너지 함수를 확률 그래프 모델 중 하나인 마르코프 랜덤필드에 적용한 것이 **볼츠만 머신**인 것이다.



맥스웰 · 볼츠만 분포 함수에 의한 공기 분자의 속력

RBM(제한적 볼츠만 머신) 개념

물질이 특정 형상(Configuration)의 상태에 있을 확률을 표현하기 위해 다음과 같이 정의

- 형상: 노드(확률변수)들에 허용되는 값의 조합
- 특정 형상에 해당하는 미시적 상태에 대한 에너지 정의
- 형상의 집합 : $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \dots, \boldsymbol{s}_N\}$
- 에너지의 집합 : $\{\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \dots, \boldsymbol{\epsilon}_N\}$

분할함수
= 각 확률의 합이 1이 되도록 정규화 한 형태

- 물질이 형상 $\boldsymbol{s}_i$에 있을 확률 $p(\boldsymbol{s}_i)$ 정의 **(볼츠만 분포)**

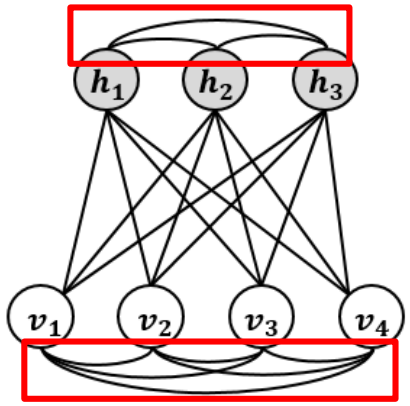
$$p(\boldsymbol{s}_i) = \frac{e^{-\beta\epsilon_i}}{\sum_j e^{-\beta\epsilon_j}} = \frac{e^{-\beta\epsilon_i}}{Z}$$

$$Z = \sum_j e^{-\beta\epsilon_j}$$

- 에너지 ϵ_i가 줄어들 수록 확률 값은 증가

볼츠만 머신과 제한적 볼츠만 머신의 차이

볼츠만 머신



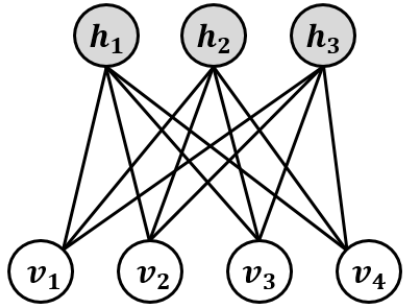
은닉층

가시층(입력층)

에너지 함수

$$\epsilon(\boldsymbol{v},\boldsymbol{h}) = -\left(\sum_{i<j} r_{ij}v_i v_j + \sum_{i<j} w_{ij}v_i h_j + \sum_{i<j} t_{ij}h_i h_j + \sum_i a_i v_i + \sum_j b_j h_j\right)$$

제한적 볼츠만 머신



에너지 함수

$$\epsilon(\boldsymbol{v},\boldsymbol{h}) = -\left(\sum_{i<j} w_{ij}v_i h_j + \sum_i a_i v_i + \sum_j b_j h_j\right)$$

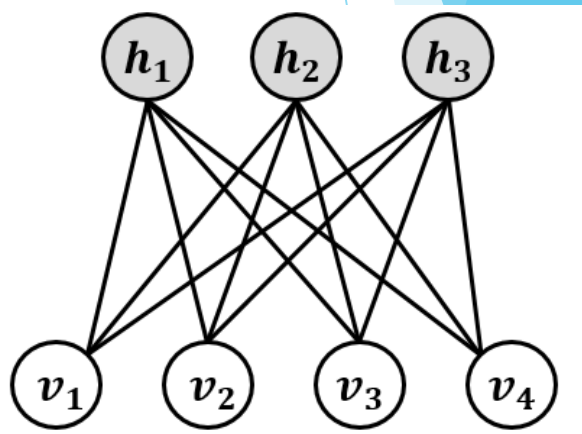
볼츠만 머신과 제한적 볼츠만 머신의 차이

제한적 볼츠만 머신의 특징

1. 볼츠만 머신의 특징 중 그래프 노드 간의 연결이 **무방향 간선**으로 연결되어 있어 **가중치가 대칭적 ($w_{ij} = w_{ji}$)**

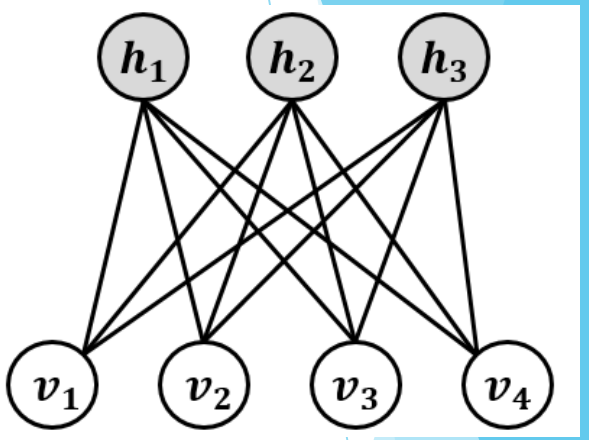
2. MLP와 같은 NN에 적용하기 위해서는 기존의 볼츠만 머신은 각 레이어 내부의 노드 간에 서로 영향을 주고 받기 때문에 적합하지 않으므로, 이러한 노드 간 연결을 모두 제거하여 **NN에 적합한 형태**로 만들어지기 때문에 RBM을 누적해서 Deep한 네트워크를 구성하는 것도 가능

3. 또한 레이어 내부의 노드간 서로 영향을 주고 받지 않는다는 것은 해당 레이어 내부의 모든 노드들은 서로 **조건부 독립**이라는 뜻 즉 특정 입력 노드(v)와 연결된 모든 은닉 노드(h)간의 조건부 확률을 계산할 때에 편리해짐



제한적 볼츠만 머신의 추론

관심 대상인 노드들의 확률 분포를 계산하는 것이 볼츠만 머신의 추론이다.
v값(입력, 가시 노드)과 **h값(은닉 노드)** 사이의
결합 확률 분포 및 조건부 확률 분포를 계산해 내는 것이다.



$$p(\boldsymbol{v},\boldsymbol{h}) = \frac{e^{-\epsilon(\boldsymbol{v},\boldsymbol{h})}}{Z} \quad Z = \sum_{\boldsymbol{v},\boldsymbol{h}} e^{-\epsilon(\boldsymbol{v},\boldsymbol{h})} \quad \Rightarrow \quad p(\boldsymbol{v}) = \frac{1}{Z}\sum_{\boldsymbol{h}} e^{-\epsilon(\boldsymbol{v},\boldsymbol{h})}$$

p(v)는 모든 은닉 노드 h와의 **결합 확률분포의 합**

$$p(\boldsymbol{h}|\boldsymbol{v}) = \prod_{j=1}^{m} p(h_j|\boldsymbol{v})$$

$$p(\boldsymbol{v}|\boldsymbol{h}) = \prod_{i=1}^{n} p(v_i|\boldsymbol{h})$$

각 레이어 내부의 모든 노드들은 **조건부 독립**이기 때문

제한적 볼츠만 머신의 추론

입력(가시) 노드와 은닉 노드의 값이 0 또는 1로 이진값을 가질 때를 가정한 노드 확률 분포

입력 노드 v에 대해 특정 은닉 노드 h_j 가 1일 확률

$$p(h_j = 1|\boldsymbol{v}) = \frac{1}{1+\exp\left(-b_j - \sum_i v_i w_{ij}\right)} = sigm\left(b_j + \sum_i v_i w_{ij}\right)$$

에너지 함수 대입
h 또는 v값에 1 대입

히든 노드 h에 대해 특정 입력 노드 v_i 가 1일 확률

$$p(v_i = 1|\boldsymbol{h}) = \frac{1}{1+\exp\left(-a_i - \sum_j h_j w_{ij}\right)} = sigm\left(a_i + \sum_j h_j w_{ij}\right)$$

제한적 볼츠만 머신의 학습 (개념)

1. RBM에서의 학습이란 에너지 함수의 파라미터를 학습시키는 것(w-가중치, a, b –편차항)

2. 분류기에서의 cost function의 gradient를 구하는 것과 같이
   RBM에서는 **log likelihood = log p(V)** 의 gradient를 구해서 학습에 활용

3. RBM의 학습 목표는 **p(V)값, 입력 데이터의 확률이 최대가 되도록 하는 것**이기 때문에
   gradient 값이 커지는 방향으로 w, a, b를 조금씩 update하는 방식으로 이루어진다.
   (경사하강법이 아닌 **경사상승법**)

4. 경사상승법을 통해 파라미터를 업데이트하기 위해서는 입력층(학습데이터) 뿐만 아니라,
   은닉층의 확률분포도 동시에 고려해야 하는 문제가 있음. **(negative gradient 값)**
   이를 계산하기 위해 **깁스 표본추출과 마르코프 체인 몬테카를로 방법**을 사용하여 계산할 수
   있지만 본래의 계산을 모두 수행하는 것은 **계산량이 너무 방대하여 시간이 오래걸림**

5. **k-단계 대조분기(K-step Contrastive divergence)**은 CD-k라고도 불리며,
   일정 확률분포에 수렴할 때까지 반복하는 것이 아닌, k번 깁스 표본추출을 통해
   해당 negative gradient 값을 구하는 방식을 이용

제한적 볼츠만 머신의 학습

모든 학습 데이터에 대해 높은 확률값을 갖도록 가중치 w_{ij}와 편차항 a_i, b_j를 결정

학습 데이터: 가시층에 입력으로 제공되는 데이터

$V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_M\}$

가능도(likelihood)

RBM와 같은 확률 모델을 사용하여 계산한 데이터에 대한 확률

전체 학습 데이터에 대한 가능도

각 학습 데이터의 가능도를 곱한 것

$$p(V) = \prod_{i=1}^{M} p(\boldsymbol{v}_i)$$

로그 가능도(log likelihood) ⇒ 주로 사용하는 **목적함수** (cost function의 개념과 유사)

$$l = \sum_{i=1}^{M} \log p(\boldsymbol{v}_i)$$

제한적 볼츠만 머신의 학습

학습데이터(입력)에 대한 확률분포(가능도)를 최대로 만드는 파라미터 선택
= Maximum likelihood estimation (ML)

learning rate

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} + \eta \frac{\partial l}{\partial w_{ij}}$$

← 편미분을 이용해 게산

$$a_i^{(t+1)} = a_i^{(t)} + \eta \frac{\partial l}{\partial a_i}$$

$$b_j^{(t+1)} = b_j^{(t)} + \eta \frac{\partial l}{\partial b_j}$$

$$\frac{\partial l_v}{\partial w_{ij}} = \frac{\partial \log p(\boldsymbol{v})}{\partial w_{ij}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}$$

$$\frac{\partial l_v}{\partial a_i} = \langle v_i \rangle_{data} - \langle v_i \rangle_{model}$$

$$\frac{\partial l_v}{\partial b_j} = \langle h_j \rangle_{data} - \langle h_j \rangle_{model}$$

제한적 볼츠만 머신의 학습

$$\frac{\partial l_v}{\partial w_{ij}} = \frac{\partial \log p(\boldsymbol{v})}{\partial w_{ij}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}$$

이진값을 가지는 RBM에서

$\langle v_i h_j \rangle_{data} = p(h_j = 1|\boldsymbol{v}) v_i$ ➡ v값(입력 데이터)만 알면 쉽게 계산가능

$\langle v_i h_j \rangle_{model} = \sum_{\boldsymbol{v}} p(\boldsymbol{v}) p(h_j = 1|\boldsymbol{v}) v_i$ ➡ v값과 h값을 모두 알아야하며,
$\sum_v p(v)$ 에 의해 Z값도 구해야 하는
복잡한 연산
=> n차원 데이터의 경우 2^n개의 연산이 필요

⬇

마르코프 체인 몬테카를로 방법을 사용해 추정이 가능

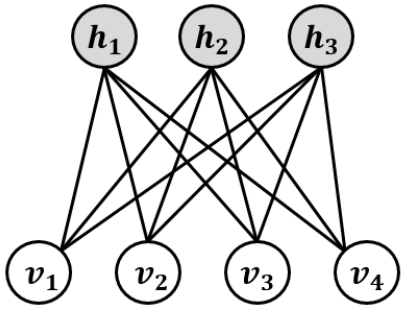
- 몬테카를로 기법 : 무작위로 많은 표본을 추출해 기대값 추정
- 마르코프 체인 기법 : i+1번째 상태의 값은 i번째 확률값에만 영향을 받음

제한적 볼츠만 머신의 학습

몬테카를로 방법 중 **깁스 표본추출**(Gibbs Sampling) 방법 이용

깁스 샘플링 : 다른 레이어의 확률 값은 모두 고정되어 있는 것으로 간주

ex)



구간[0,1]에서 무작위 난수를 하나 샘플링한 뒤,
조건부 확률 값보다 작으면 1, 크면 0으로 값을 결정한다.

v값들이 주어지면 h값들을 샘플링 할 수 있으며,
h값들이 주어지면 v값들을 샘플링 할 수 있다.

$$h_1 \sim p(h_1 \mid v_1, v_2, v_3, v_4)$$
$$h_2 \sim p(h_2 \mid v_1, v_2, v_3, v_4)$$
$$h_3 \sim p(h_3 \mid v_1, v_2, v_3, v_4)$$

$$v_1 \sim p(v_1 \mid h_1, h_2, h_3)$$
$$v_2 \sim p(v_2 \mid h_1, h_2, h_3)$$
$$v_3 \sim p(v_3 \mid h_1, h_2, h_3)$$
$$v_4 \sim p(v_4 \mid h_1, h_2, h_3)$$

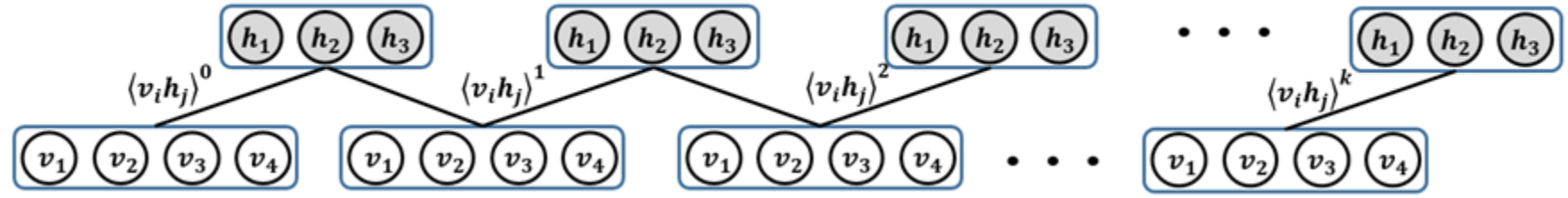
제한적 볼츠만 머신의 학습

k-단계 대조분기(CD-k) 학습법

$$\frac{\partial l_v}{\partial w_{ij}} = \frac{\partial \log p(\boldsymbol{v})}{\partial w_{ij}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}$$

$$\langle v_i h_j \rangle_{model} = \sum_{\boldsymbol{v}} p(\boldsymbol{v}) p(h_j = 1 | \boldsymbol{v}) v_i$$

위의 식에서 $< v_i h_j >_{model}$을 계산할 때에 기존 몬테카를로 방식에서
데이터 확률이 특정 분포에 수렴할 때까지 진행하는 것이 아닌,
k번 대조 분기 시킴으로써, k 단계에서의 v값과 h값을 대입하여 계산한다.



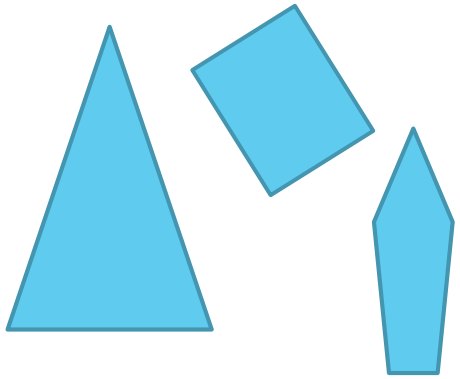
제한적 볼츠만 머신의 사전학습(Pre training)

RBM의 학습은 비지도 학습(unsupervised learning)이며, 학습데이터와 유사한 데이터분포를 생성하는 생성모델이라고 할 수 있다.

RBM에서 학습된 가중치(w)와, 편차항(a,b)를 통해 계산된 은닉노드들의 값(h)들은 입력 데이터 값(v)들을 구성하는 특징을 나타낸다고 볼 수 있다.

즉 CNN에서 Feature extraction으로 검출된 Feature map과 유사한 개념이지만, 이때의 Feature map은 분류를 위한 feature 였다면,

RBM에서 **은닉노드 h 값**들은 **입력데이터를 표현하기 위한 feature**라고 생각할 수 있다.

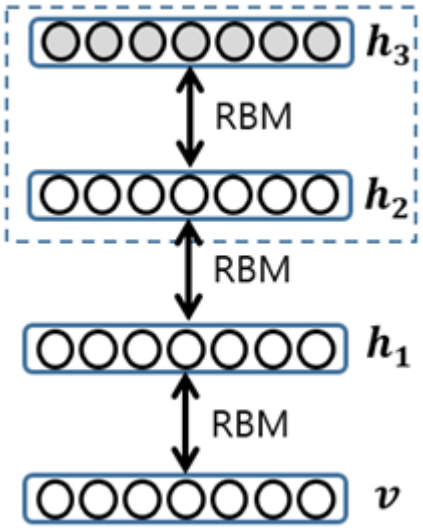


→ CNN
분류를 위한 Feature → 꼭지점의 개수
간선의 개수 등

→ RBM
입력데이터 표현을 위한
Feature → 너비, 각 변의 길이
각 꼭지점의 위치 등

제한적 볼츠만 머신의 사전학습(Pre training)

이러한 RBM의 특성을 이용하여, RBM을 누적해서 네트워크를 구성하게 되면



… 추상화된 입력데이터의 Feature들

v를 표현하기위한 Feature들을 표현하기위한 Feature들

v를 표현하기위한 Feature들

위와같은 형태의 네트워크가 학습이 되는데,
이 때 각 층의 가중치값과 편차항 값들을 MLP의 초기값으로 설정하게 되면
더 높은 성능의 네트워크를 구성 할 수 있다고 한다.

Thanks you

and **Q & A**