**베타 변분 오토인코더의**

**변형 모델에 대한 연구**

****

중앙대학교 컴퓨터공학부

20131362 정기민

**요약**

베타 변분 오토인코더 (-VAE)의 추론 모델은 분리 표현 학습과 차원 축소 학습을 동시에 담당합니다. 하나의 네트워크가 이를 동시에 감당하는 것은 다소 부담이 될 수 있습니다. 또 한 기존의 -VAE는 생성 모델과 추론 모델이 서로 비율의 제약을 받는 잠재채널을 통해 상호간의 전파를 주고받습니다. 이 과정에서 생성 모델은 학습 과정에서 양질의 입력을 입력으로 받아 양질의 복원 생성물을 생성할 기회를 갖지 못하고 추론 모델 또한 복원 손실항에 대한 양질의 역전파를 전달받지 못합니다. 본 논문에서는 잠재변수와 관측변수 사이에 중간변수를 도입하고 각 확률변수들의 관계를 바탕으로 -VAE의 목적함수를 변형하여 차원 축소 학습을 담당하는 네트워크와 분리 표현 학습을 담당하는 네트워크로 분리된 추론 모델을 갖으며 학습 과정에서 생성 모델과 추론 모델이 잠재채널을 거치기 전의 정보를 공유하는 모델을 소개합니다.

**서론**

변분 오토 인코더(Variational autoencoder)는 일반적인 오토 인코더와 달리 비지도 분리 표현(Unsupervised Disentangled Representation) 능력이 동반된 차원 축소 학습을 할 수 있다는 점에서 큰 주목을 받고 있습니다. 비지도 분리 표현 능력 이란 데이터 분포나 데이터 집합이 주어졌을 때 이를 구성하는 데이터들이 가상의 생성 인자(Generative factor) 들에 의해 생성되며 이들이 서로 독립된(분리된) 확률분포의 확률변수들의 형태로 존재할 것이라는 가정 하에 이를 비지도적으로 추론해내고 데이터를 추론해낸 생성 인자의 분포에 대한 확률변수의 형태로 표현하는 능력을 말합니다. 손 글씨 숫자 데이터 집합(MNIST)을 예로 든다면 숫자의 종류나 기울기, 너비, 선의 굵기와 같은 것들을 각각의 독립된 생성 인자라고 추론할 수 있고 하나의 손 글씨 이미지가 주어졌을 때 이를 숫자의 종류, 기울기의 정도, 너비의 정도, 선의 굵기의 정도 등의 확률변수로 표현하는 것이 분리 표현 능력이라고 말할 수 있습니다.

변분 오토인코더(VAE) 모델의 목적함수 는 근거 하한(Evidence Lower Bound)이라고 불리는 실제(True)분포(생성 모델)의 로그 가능도()의 하한의 형태로 처음 소개되었습니다 [7]. 그로 인하여 항은 통상적으로 음수가 곱해진 형태로 생성 모델의 관점에서 해석되어 복원 손실(Reconstruction loss) 항으로 불리고 있습니다. 한편 를 구성하는 나머지 항은 변분 분포의 사후 확률분포 와 실제 사전 확률분포 간의 쿨백-라이블러 발산(Kullback-Leibler divergence) 항으로 정규화(regularization) 항이라고 불립니다. VAE의 분리 표현 능력은 복원 손실 항에 더불어 이 정규화 항이 공동으로 감소되는 과정에서 얻어지는 것으로 알려져 있습니다. 정규화 항이 분리 표현 학습에 밀접한 연관이 있다는 점은 의 정규화 항에 강도를 조절하는 가중치 를 도입하는 시도를 한 -VAE 모델의 연구를 시작으로 본격적으로 주목받게 되었습니다 [3].

**(1)**

그 후의 연구는 -VAE의 목적함수 가 정보 병목 목적함수(Information bottleneck objective) 의 하한으로 해석될 수 있다는 관점을 제시하였습니다 [1]. (식 **(1)**에 대한 내용은 부록 A에서 다루어 집니다.) 이는 생성 모델의 관점에서 탄생한 VAE모델을 비율-왜곡 문제 (Rate-Distortion problem)를 해결하는 인코더(추론 모델)의 관점에서 바라보는 기회를 제공함과 동시에 -VAE에서 도입된 의 존재를 채널의 비율(Rate)을 제약하는 가중치로 바라볼 수 있다는 이론적인 배경을 제시하였습니다 [4]. 논문 [2]에서는 왜 의 강도를 높일수록 -VAE의 추론 모델이 인간의 직관에 들어맞으면서 일반화된 생성 요인을 표현하는 능력을 갖는가에 대한 분석을 잠재채널에서 형성되는 공유 코딩 공간(shared coding space)을 중심으로 분석하였습니다. 논문 [5]에서는 실제 사전 분포 의 결합분포를 독립으로 설정하였을 때 정규화 항으로부터 잠재벡터를 구성하는 잠재변수들 간의 종속성을 나타내는 종합 상관관계(Total correlation) 항을 분리해 낼 수 있다는 것을 보였습니다. 이는 정규화 항이 사후 집계 확률분포 또한 독립이 되도록 유도한다는 것을 의미하는데, 해당 논문에서는 이러한 작업이 분리 표현 능력의 향상에 상당한 도움을 준다는 것을 실험적으로 입증하였습니다. 이를 통해 추론 모델은 분리 표현 능력을 향상시키기 위해서는 추가적으로 추론해낸 잠재변수들 간의 종속성을 최소화하는 것이 요구된다는 것을 알 수 있습니다.

최근 -VAE 및 변형모델을 이용한 비지도 분리 표현 학습은 많은 주목을 받고 있으며 이미지 생성분야 뿐만 아니라 강화학습, 자연어 처리 등 다양한 분야에서 활용되고 있습니다. 앞서 언급한 연구와 활용사례를 통해 -VAE의 추론 모델에게 요구되는 분리 표현 학습은 상당히 부담되면서도 중요한 작업임을 알 수 있습니다. 비단 추론 모델이 해결해야 하는 문제는 분리 표현 학습만이 아닙니다. 일반적으로 VAE의 잠재벡터 Z는 관측벡터 X에 비해 상대적으로 현저히 작은 크기를 갖도록 설정됩니다. 이러한 통상적인 VAE모델의 구조적 설정은 목적함수 가 추론 모델에게 분리 표현과 더불어 차원 축소(Dimensionality Reduction) 표현 능력 또한 요구하도록 만듭니다. 여기서 차원 축소 표현 능력이란 작은 차원(크기)를 갖는 잠재벡터 Z에 이보다 현저히 큰 크기의 관측벡터 X에 대한 상호 정보량 을 보존하는 것을 의미합니다. 더불어 기존의 VAE 모델은 그 구조적으로 생성 모델과 추론 모델이 서로 정규화 항의 제약을 받는 잠재채널을 통해 전파와 역전파를 주고받습니다. 따라서 학습 과정에서 서로는 큰 정보적 손실을 입게 됩니다.

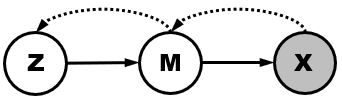
본 논문에서는 기존의 -VAE와 세 가지 차이점을 갖는 -VAE의 변형 모델을 소개하고자 합니다.

1. 추론 모델이 분리 표현에 전념하는 네트워크와 효율적으로 차원 축소 표현 학습을 진행하는 네트워크로 분리됩니다.
2. 추론 모델과 생성 모델이 학습 과정에서 잠재채널을 지나기 전의 정보들을 공유합니다.
3. 관측벡터를 생성하는 네트워크가 잠재채널을 거치지 않은 양질의 입력 정보를 받아 양질의 복원을 생성하는 기회와 능력을 갖습니다.

앞으로의 논의의 편의를 위해 해당 모델을 모델 A라고 부르도록 하겠습니다.

**배경**

본 논문에서는 VAE와 마찬가지로 생성 모델이 실제 분포인 관점에서 변분 추론 기법을 이용합니다. 추론 모델의 차원축소 학습과 분리 표현 학습을 분리하기 위해서 본 논문에서는 관측벡터 X와 잠재벡터 Z 사이에 이를 매개하는 벡터 M을 도입합니다. 편의를 위해 이를 중간벡터라고 부르도록 하겠습니다. 각각의 확률변수(벡터)들은 **[그림 1]**의 관계를 같는 것으로 가정됩니다.



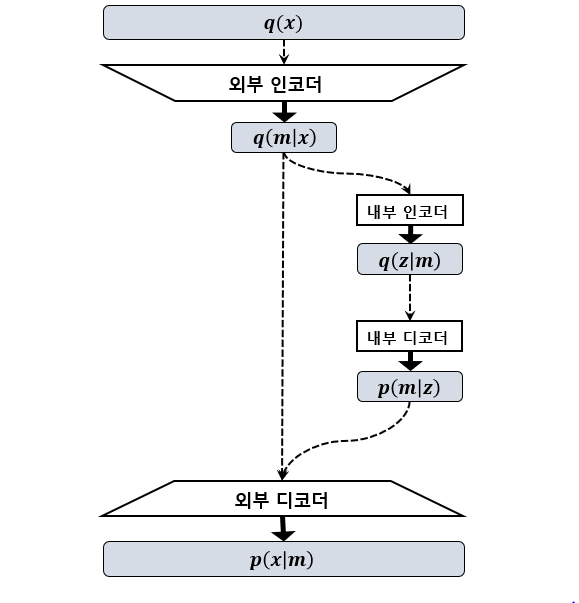
**[그림 1]**

**(2)**

**(3)**

실선 화살표는 실제 분포를, 점선으로 이루어진 화살표는 변분 분포가 갖는 관계를 표현합니다. 이러한 관계로부터 만족되는 식 **(2)**에 근거하여 추론 모델 을 각각 와 로 분리합니다. 편의를 위해 추론하는 네트워크를 외부 인코더, 추론하는 네트워크를 내부 인코더라고 부르도록 하겠습니다. 본 논문에서는 중간채널 M의 크기를 잠재채널 Z의 크기와 같도록 설정합니다. 따라서 내부 인코더는 차원축소의 부담 없이 M을 매개하여 관측변수 X에 대한 분리 표현 학습을 진행합니다. 외부 인코더는 차원축소 학습에 전념함과 동시에 타 네트워크들의 일부 요구사항들과 타협합니다.

추론 모델과 마찬가지로, 생성 모델 또한 식 **(3)**에 근거하여 을 추론하는 내부 디코더, 을 추론하는 외부 디코더로 분리합니다. 내부 디코더는 잠재채널로부터 변분 분포의 중간채널 M을 복원하고 외부 디코더의 복원 손실을 줄이는 양질의 입력 M을 생성합니다. 외부 디코더는 각각 외부 인코더와 내부 디코더로부터 추출된 표본 M을 입력으로 하여 관측벡터 X를 복원합니다.



**[그림 2]**

**[그림 2]**는 본 논문에서 소개하고자 하는 모델 A의 전체적인 구조를 나타냅니다. 실선 화살표는 네트워크가 확률 분포의 모수를 출력한다는 것을 표현합니다. 점선 화살표는 분포로부터 추출되는 표본이 네트워크의 입력으로 들어가는 것을 표현합니다.

**목적함수**

**(4)**

**(5)**

모델A의 목적함수 는 로부터 유도됩니다. -VAE 모델은 다루기 힘든 항을 증가시키기 위해서 계산하기 쉬운 하한인 음의 복원 손실 항을 계산하여 증가시킴으로써 를 간접적으로 증가시킵니다. 본 모델은 -VAE에서 이용한 음의 복원 손실 항과 더불어 또 다른 방식으로 유도될 수 있는의 하한을 동시에 증가시키는 전략을 이용합니다.

**(6)**

**(7)**

식 **(6)**은 -VAE모델에서 이용된 하한입니다. 식 **(7)**은 모델 A의 목적함수를 유도하기 위해 도입되는 의 하한입니다. (증명은 부록 B에서 다루어 집니다) 식 **(6)**, **(7)**은 식 **(5)**에서 나뉘어진 에 대한 각각의 하한으로 이용됩니다.

식 **(7)**에서 먼저 주목할만한 항은 항입니다. 이는 외부 인코더의 결과 ​에서 추출된 m​을 외부 디코더의 입력으로 하였을 때 계산되는 음의 복원 손실 항입니다. 이를 통해 모델 A의 외부 디코더는 잠재채널을 거치지 않은 양질의 정보를 입력으로 받아 양질의 복원을 생성하는 기회를 갖습니다. 또 한 외부 인코더는 잠재채널을 거치지 않은 복원 손실항의 역전파를 취합니다. 한 편, 항은 의 하한으로 해석될 수 있습니다. 이 해석을 바탕으로 이를 증가시키는 것은 외부 인코더에 의해 결정되는 분포 와 가 을 증가시키는 방향으로 구성되도록 유도할 것이라고 기대할 수 있습니다. 이 말은 잠재채널 Z 보다 앞선 중간채널 M에서 차원 축소 학습이 선행되는 것을 의미합니다. 즉 외부 인코더에 의해 항이 감소된다면 내부 인코더는 차원 축소의 부담을 덜고 분리 표현 학습에 집중할 수 있게 될 것입니다.

항은 외부 인코더의 결과 ​와 내부 디코더의 결과 ​간의 쿨백-라이블러 발산 항입니다. 이는 간단명료하게 와 를 서로 닮도록 유도하는 효과를 불러일으킵니다. 이를 통해 잠재채널을 지난 내부 디코더의 결과는 잠재채널을 지나기 전의 외부 인코더의 결과로부터 지도학습을 받는 효과를 얻습니다. 내부 디코더의 관점에서 이는 외부 인코더에 대한 순(forward) 쿨백-라이블러 발산 함수로, ​인 공간에 대하여 의 밀도(질량) 또한 0 보다 커지도록 요구합니다. 즉 ​는 ​의 공간을 포함하도록 강제 받게 됩니다. 한편, ​의 관점에서 이는 ​에 대한 역(reverse) 쿨백-라이블러 발산 함수입니다. 이는 ​의 밀도(질량)가 큰 공간을 ​의 밀도가 큰 공간에 위치시키려는 성질을 가지고 있습니다. 결과적으로 는 ​의 봉우리(Mode) 내에 머무르면서 ​을 보존하는 적합한 ​M의 공간을 탐색하는 기회를 노립니다. 한편 ​​는 관측변수에 대한 상호 정보량을 많이 가지고 있는 ​의 공간을 포함하도록 강제 받는 과정에서 관측변수의 질을 향상시키는 M​의 공간 주변에 산재하게 되며 자체적으로 그 주변에서 관측변수의 왜곡(Distortion)을 줄이는 m​의 공간을 탐색하게 됩니다.

항은 항과 같습니다. 이는 **[그림 1]**의 관계에서 보았을 때 Z에 담긴 M에 대한 정보량 중, X와 관련 없는 정보량이라고 해석할 수 있습니다. 우리의 목적이 잠재채널 Z이 M이 아니라 X에 대한 정보를 담도록 요구한다는 점에서 이를 감소시키는 것이 타당할 것입니다. 순, 역 쿨백-라이블러 관점에서 보았을 때 이는 관측벡터 x에 의해 추론될 수 있는 들이 서로 겹치는 공간을 형성하며 매개되는 M과 관계없이 공유되는 공간을 가리도록 유도할 것입니다.

항은 항으로 대체됩니다.

**(8)**

**실질적 목적함수**

식 **(8)**의 항과 항은 **[그림 2]**의 구조를 갖고 있는 모델 A에 의해서는 직접적으로 계산될 수 없습니다. 본 논문에서는 이 두 항에 대한 각각의 상한의 추정치를 구하여 이를 감소시키는 방식으로 이를 간접적으로 감소시키는 것을 시도하였습니다.

**(9)**

**(10)**

이는 외부 인코더의 결과 로부터 M개의 표본을 추가적으로 추출하고 이를 내부 인코더에 통과시켜 계산된 M개의 를 이용해 계산할 수 있습니다. 이는 독립 추출이므로 한꺼번에 빠르고 효율적으로 계산할 수 있습니다.

실험 결과, 식 **(9)**을 감소시키는 것은 모델의 학습을 더디고 불안정하게 만드는 것으로 나타났습니다. 본 논문에서는 학습 시 해당 항의 강도를 조절하는 가중치 를 도입하여 이를 해결하였으며, 0에서 0.1 사이의 값을 사용하였습니다.

실질적 목적함수는 과 같습니다. 항과 항은 M개의 표본 의 평균을 이용하여 추정치를 구합니다. 이를 위해 내부 인코더는 추가적으로 M번의 계산이 요구됩니다. 그 외의 항들은 각각 한 번씩 계산합니다. 정리를 하자면, 하나의 데이터 포인트를 학습할 때, 총 외부 인코더는 1번, 내부 인코더는 M+1번, 내부 디코더는 1번, 외부 디코더는 2번의 연산을 하게 됩니다. 항은 데이터 x에 대한 손실함수를 나타냅니다. 이는 에서 직접적으로 유도됩니다.

**훈련**

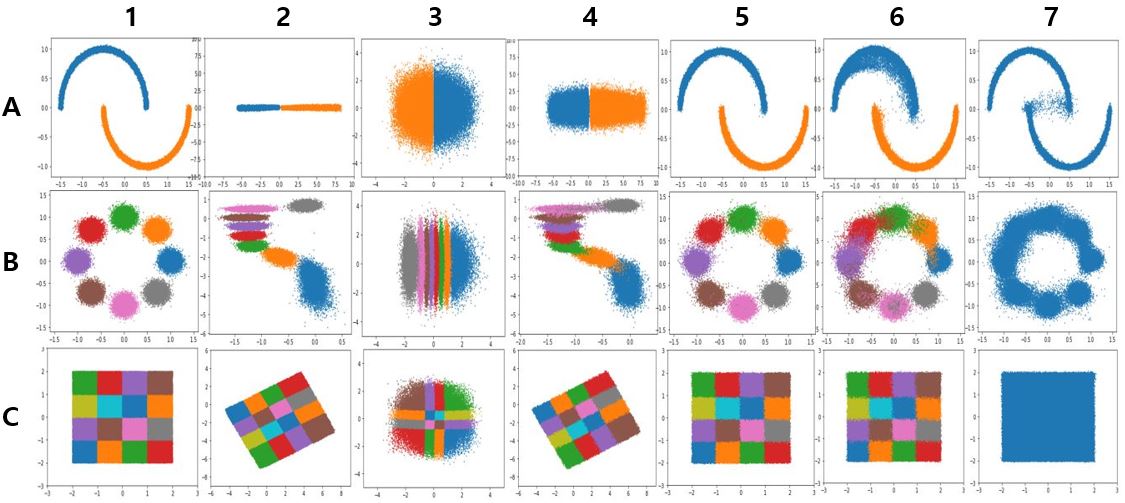
모델A는 다른 VAE 변형 모델들과 마찬가지로 큰 값이 설정될 때 잠재 변수 붕괴(Latent variable Collapse) 현상을 보였습니다. 이는 생성 모델이 특정 잠재변수를 무시하는 현상으로, 복원 손실에 영향을 미치지 않게 되어 결국 정규화 항이 0에 수렴하며 비활성화되는 현상을 말합니다.

본 논문에서는 이를 해결하기 위해 두 가지 논문을 참고하였습니다. 논문 [9]에서는 잠재벡터를 디코더의 출력층에 가까운 층으로 스킵 연결(Skip connection)되는 것이 잠재 변수 붕괴를 완화한다고 주장하였습니다. 이를 참고하여 외부 디코더의 입력 M에 잠재벡터 Z을 연결하였더니 잠재 붕괴 문제가 완화되는 것을 확인하였습니다. 이는 **[그림 1]**의 관계에 근거하여 이 만족한다는 점에서 확률 변수 관계의 측면에서도 문제되지 않습니다. 논문 [10]에서는 잠재 변수 붕괴를 방지하는 방법으로, 주기적으로 를 0으로 초기화하고 이를 점진적으로 증가시키는 것을 반복하는 주기적 어닐링 스케줄링(cyclical annealing schedule)을 제안하였고, 해당 어닐링 기법이 본 모델을 효과적으로 학습시키면서 잠재 변수 붕괴를 방지하는 것을 확인하였습니다.

**실험**

**실험 1**

실험 1은 관측벡터 X, 중간벡터 M, 잠재벡터 Z의 변분 분포와 실제 분포를 시각화 하는 실험입니다. 이 실험에서는 서로 다른 세개의 데이터셋이 사용되었습니다. 세 데이터셋 모두 2차원의 크기를 가졌으며, 각각의 데이터집합에 서로 다른 모델이 사용되었습니다. 각각의 모델은 시각화를 위해 중간벡터 M, 잠재벡터 Z 모두 2차원의 크기를 갖도록 설정하였습니다.



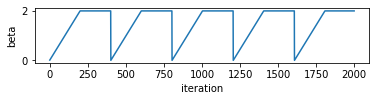
**[그림 3]**

[그림 3]은 실험 1에 대한 그림입니다. 각각의 행은 서로 다른 데이터 집합의 결과를 나타냅니다. 데이터들의 색상은 위치를 식별하기 위해 시각화 과정에서 부여한 것일 뿐 모델의 학습에 이용된 정보는 아닙니다. 1열은 데이터 셋, 2열은 외부 인코더의 결과의 표본 , 3열은 내부 인코더의 결과의 표본 , 4열은 내부 인코더로부터 입력을 받았을 때의 외부 인코더의 결과의 표본 를 나타냅니다. 5, 6, 7열은 모두 외부 디코더의 결과물입니다. 5열은 외부 인코더로 입력을 받았을 때의 외부 디코더의 결과의 표본 , 6열은 내부 인코더의 으로부터 내부 디코더를 거친 외부 디코더의 결과의 표본 , 7열은 실제 사전분포 로부터 추출된 모델이 학습한 실제 분포의 표본 를 나타냅니다.

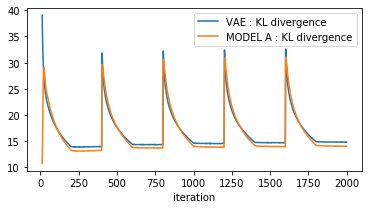
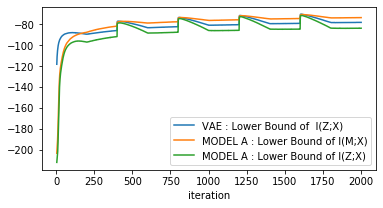
3열을 통해 본 모델이 분리 표현 학습을 잘 수행하는 것을 확인할 수 있습니다. 2열과 4열을 보면 와 ​가 서로 매칭되며, 특히 A행에서는 앞서 언급한 쿨백-라이블러 발산항의 특징이 잘 나타나는 것을 관찰할 수 있습니다.

**실험 2**

실험 2는 학습 과정에서 모델 A의 음의 복원 손실항과 잠재채널의 실제 사전분포와의 쿨백-라이블러 발산 항의 변화 양상을 보입니다. 비교를 위해 같은 조건에서 학습된 β-VAE의 모델의 결과를 같이 나타냈습니다. 실험은 MNIST 데이터셋을 이용하였고, 각각의 잠재벡터의 크기는 8, 모델A의 중간벡터의 크기는 8을 설정하였습니다.



**[그림 4]**



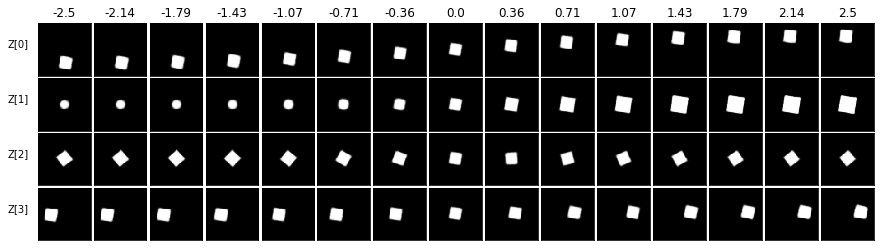
**[그림 5]**

**[그림 4]**는 두 모델의 학습에 사용된 β 주기적 어닐링 스케줄입니다. **[그림 5]**의 왼쪽 그림은 두 모델의 복원 손실항을 나타냅니다. 연두색 선은 모델 A의 학습 과정에서 계산된 항, 주황색 선은 항의 평균을 나타냅니다. 하늘색 선은 비교를 위해 따로 학습된 β-VAE모델의 학습 과정에서 계산된 항의 평균을 나타냅니다. 주황색 선은 외부 인코더의 차원 축소 학습의 성능을 나타낸다고 할 수 있습니다. 실험 결과 이는 실제로 기존의 β-VAE의 복원성능보다 조금 더 좋은 성능을 보였습니다. 반면 모델 A의 실제 복원 능력에 해당한다고 할 수 있는 연두색 선은 기존의 β-VAE보다 조금 뒤떨어지는 것으로 확인되었습니다. 그러나 복원 손실항이 높다는 것이 분리 표현 능력이 뒤떨어진다는 것을 의미하는 것은 아니며 오른쪽 그림에서 확인할 수 있듯 잠재채널과 실제 사전분포와의 쿨백-라이블러 발산 항이 모델 A가 근소하게 더 작다는 점에서 모델 A가 기존의 β-VAE에 비하여 뒤떨어진다고 볼 수는 없을 것입니다.

**실험 3**

실험 3은 두 가지 데이터셋(dSprites, 3D chairs)을 학습한 모델의 생성 모델의 결과를 보이는 실험입니다.

**실험 3-1:**

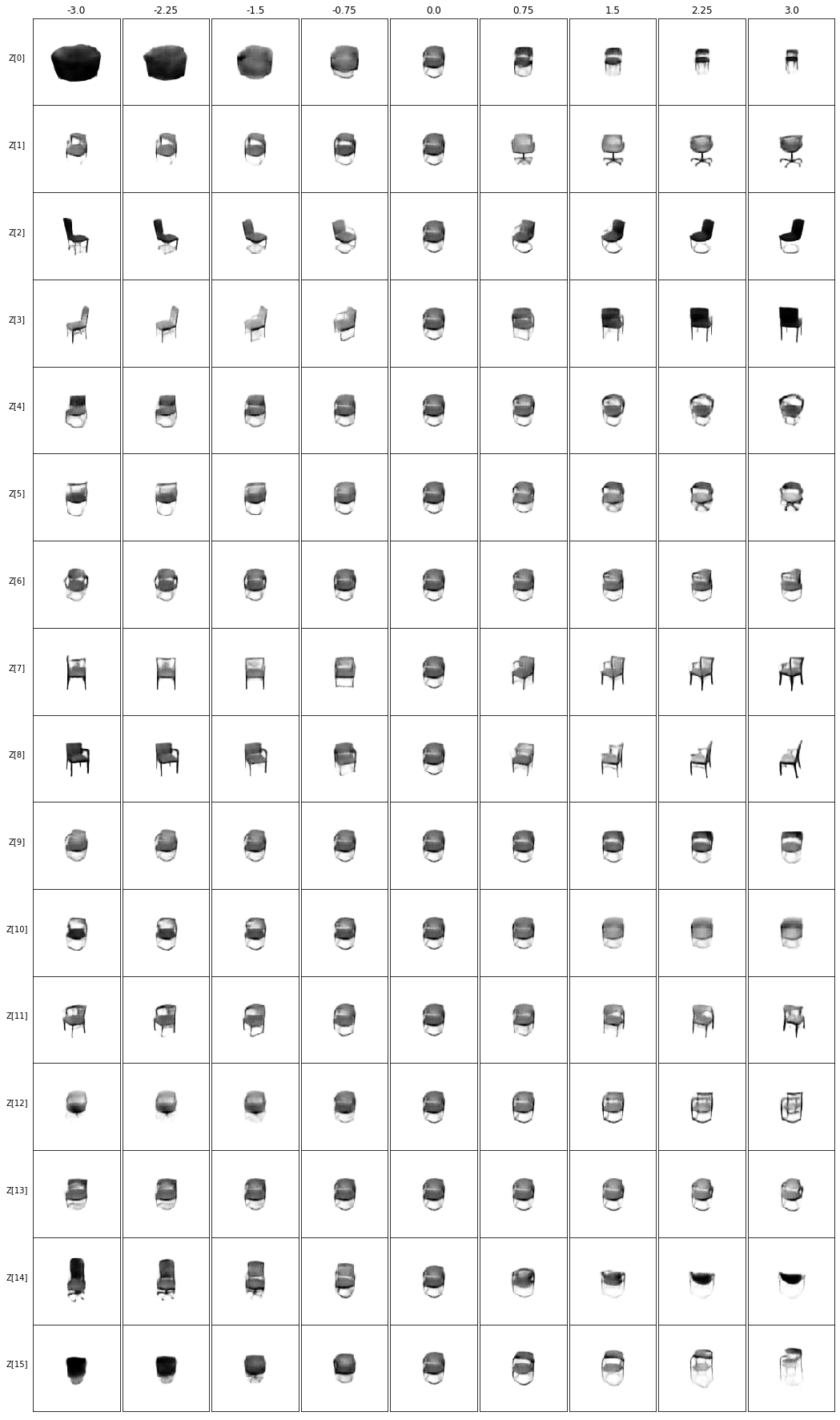


**[그림 5]**

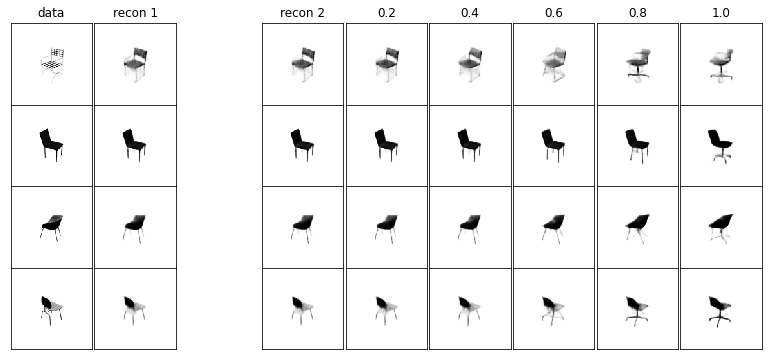
실험 3-1은 dSprites 데이터 셋에 대한 실험입니다. dSprites 데이터 셋은 x, y축 위치, 물체의 크기, 물체의 회전 각도, 물체의 종류(사각형, 타원, 하트) 총 다섯 가지의 생성 요인이 있습니다. 본 실험에서는 사각형 데이터만을 이용하였습니다. 중간벡터와 잠재벡터의 크기는 4, β는 0부터 10까지 주기적 어닐링 스케줄링을 사용하였습니다. **[그림 5]**는 모델 A의 생성 결과를 나타냅니다. 각각의 행은 나머지 잠재변수들이 고정된 채 해당 행의 왼쪽에 표시된 잠재변수의 변화에 따른 생성물의 결과를 나타냅니다. 각 열의 위쪽 값은 변화되는 잠재변수의 값을 나타냅니다. 3번째, 0번째 잠재변수는 각각 x, y축의 위치를 표현하는 것으로 보이며, 1번째 잠재변수는 크기, 2번째 잠재변수는 회전 각도를 표현한 것으로 보입니다.

**실험 3-2:**

실험 3-2는 3D chairs 데이터 셋에 대한 실험입니다. 3D chairs 데이터 셋은 약 1000개 종류의 3D 의자 모델을 다양한 각도로 렌더링한 이미지로 이루어진 데이터 셋입니다. 학습을 위해 가로 세로 128크기의 흑백 이미지로 변환하였으며, 모델 A의 잠재벡터, 중간벡터의 크기는 16으로 설정하였습니다. β는 0부터 8까지 주기적 어닐링 스케줄링을 사용하였습니다. **[그림 6]**은 **[그림 5]**와 같은 방식으로 16개의 잠재변수들의 변화에 대한 생성 모델의 생성 결과를 나타냅니다. 1열의 잠재변수는 의자의 부피, 2열은 다리 스타일, 14열의 잠재벡터는 등받이의 길이 등을 표현한 것으로 보입니다. 그러나 방위(의자의 방향)에 대한 표현은 3, 4, 7, 8, 11열의 잠재변수들에 얽혀진 형태로 표현되는 것으로 보입니다.

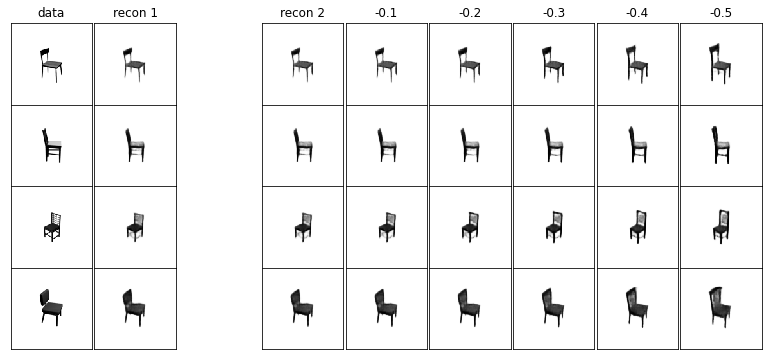


**[그림 6]**



**[그림 7]**

**[그림 7]**은 2열의 잠재변수(Z[1])의 변화(latent traversal)에 대한 결과를 나타냅니다. 첫 번째 열은 데이터, 두 번째 열(recon 1)은 외부 인코터의 결과를 이용한 생성 결과, 세 번째 열은 내부 디코더의 결과를 이용한 생성 결과를 나타냅니다. 네 번째 열부터는 해당 잠재변수를 인위적으로 변화시켜가며 얻은 결과를 나타냅니다.



**[그림 8]**

**[그림 8]**은 잠재변수 Z[14]에 대한 결과입니다. 등받이의 길이를 표현한 것으로 보입니다.

**결론**

본 논문에서는 두 가지 네트워크로 분리된 추론 모델과 두 가지 네트워크로 분리된 생성 모델을 갖는 -VAE의 변형 모델(모델 A)에 대하여 다루었습니다. 본론에서는 모델 A가 가정하는 확률변수들의 관계와 네 가지 네트워크로 구성된 모델의 구조에 대해서 소개하였습니다. 그 후 -VAE의 목적함수 로부터 본 모델의 실질적 목적함수를 유도하고 이를 구성하는 각각의 항에 대한 분석을 시도하였습니다. 그 후 본 논문의 주장을 뒷받침하고 모델 A의 성능을 보이기 위해 세 가지 실험을 진행하였습니다. 첫 번째 실험을 통해 모델 A가 추론하는 변분 분포 및 실제 분포에 대한 시각화를 시도하였습니다. 두 번째 실험에서는 모델A와 기존의 -VAE모델의 학습 과정에서 나타난 복원 손실항과 정규화 항의 변화 양상을 비교하였습니다. 세 번째 실험에서는 분리 표현 학습에 자주 사용되는 두 가지 데이터 셋에 대한 모델 A의 생성 모델의 결과물을 보였습니다.

**참고 문헌**

[1] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. International Conference on Learning Representations, 2016.

[2] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins and Alexander Lerchner. Understanding disentangling in -VAE, 2017.

[3] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed and Alexander Lerchner. -VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017.

[4] Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous and Kevin Murphy. Fixing a Broken ELBO, 2017.

[5] Ricky T. Q. Chen, Xuechen Li, Roger Grosses and David Duvenaud. Isolating Sources of Disentanglement in VAEs, 2018.

[6] Matthew D. Hoffman and Matthew J. Johnson. ELBO surgery: yet another way to carve up the variational evidence lower bound, 2016.

[7] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes, 2013.

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever and Pieter Abbeel, InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, 2016.

[9] Adji B. Dieng, Yoon Kim, Alexander M. Rush and David M. Blei. Avoiding Latent Variable Collapse With Generative Skip Models, 2019.

[10] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz and Lawrence Carin. Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing, 2019.

**부록**

**부록 A**

**A-1:**

(A-1-1)

(A-1-2)

(A-1-3)

(A-1-4)

(A-1-2)의 항은 데이터 분포 고유의 엔트로피입니다. 따라서 상수 취급하여 무시됩니다. (A-1-3)의 부등식은 쿨백-라이블러 발산함수는 언제나 0 이상이라는 성질(non-negativity)에 근거하여 식 로부터 유도됩니다.

**A-2:**

N은 데이터 집합의 크기입니다. K는 데이터 집합의 인덱스 변수입니다. 를 만족하는 이산 균일 분포를 따르며 각각의 확률 변수들과 의 관계를 갖습니다.

(A-2-1)

(A-2-2)

(A-2-3)

(A-2-4)

(A-2-5)

= (A-2-6)

여기서 k에 의해 x가 특정된다는 점에서 =라는 점이 활용됩니다. (A-2-3)의 부등식은 (A-1-3)의 부등식과 같은 방식으로 유도됩니다.

부록 A의 증명은 논문 **[1]**을 참고하였습니다.

**부록 B**

부록 B에서는 **[그림 1]**의 관계에 의해 이 만족한다는 점을 이용합니다.

**B-1:**

(B-1-1)

(B-1-2)

(B-1-3)

(B-1-4)

(B-1-5)

(B-1-6)

여기서 (B-1-5)의 부등식은 (A-1-3)의 부등식과 같은 방식으로 유도됩니다. (B-1-6)의 부등식은 부록 (A-1)에서 증명된 방식을 에 적용하여 유도됩니다.

**부록 C**

부록 C에서는 식 **(9)**, **(10)**을 다룹니다.

**C-1:**

(C-1-1)

(C-1-2)

(C-1-3)

(C-1-4)

(C-1-5)

(C-1-6)

(C-1-7)

(C-1-8)

여기서 (C-1-5)의 부등식은 젠센(Jensen) 부등식으로부터 유도됩니다.

**C-2:**

(C-2-1)

(C-2-2)

(C-2-2)

(C-2-3)

(C-2-4)

(C-2-5)

(C-2-6)

(C-2-5)의 부등식은 항이 0보다 크거나 같다는 점에서 유도됩니다.