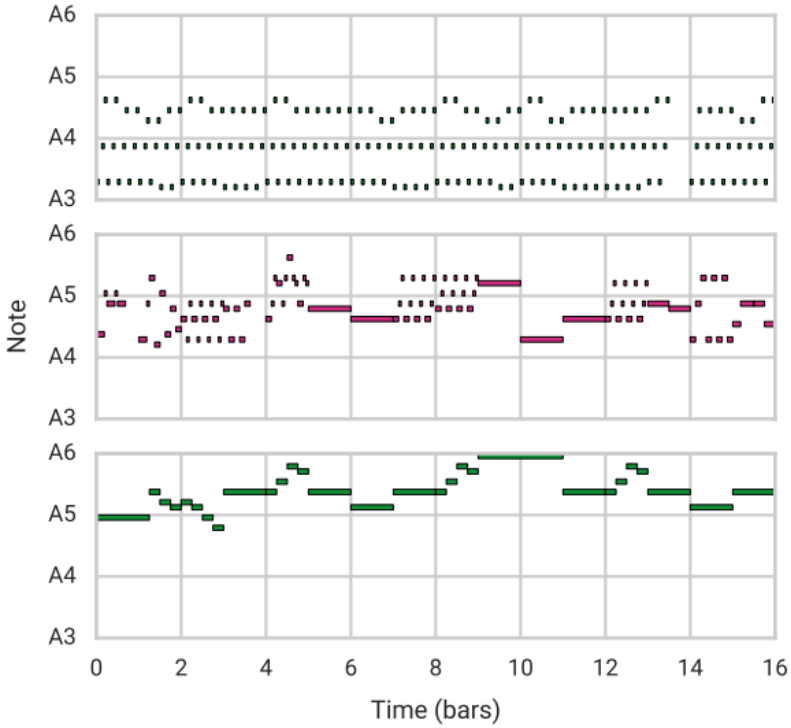
2. MusicVAE 논문 리뷰

목표: MusicVAE 논문을 읽고 모델의 동작 원리와 핵심 아이디어를 이해

A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music

1. Introduction



- **Generative modeling**:
  - data x를 생성하기 위해 사용되는 잠재적인 확률 분포를 추정하는 프레임워크
  - GANs(Generative Adversarial Networks), PixelCNN 및 WaveNet과 같이 deep generative modeling에 매우 다양한 방법이 사용되었다.
  - VAE 모델의 장점:
    - p(z|x)"와 "p(z)"를 명시적으로 모델링한다는 것
    - 여기서 z는 기존 데이터로부터 추론되거나 잠재 공간 상의 분포에서 샘플링할 수 있는 잠재 벡터
      - p(z|x): 주어진 데이터인 x를 기반으로 잠재 변수 z를 추론할 수 있는 확률 분포
      - p(z): 잠재 변수 공간의 확률 분포
    - Latent vector의 역할과 AutoEncoder 모델의 기능에 대해 설명
      - Latent Vector: 주어진 데이터 포인트의 중요한 특성을 포착하고 데이터셋의 다양한 변동 요소를 분리 가능
      - AutoEncoder: Latent vector z를 데이터 공간으로 효율적으로 매핑하는 방법을 제공하는 likelihood p(x|z) 함수를 모델링
  - 대부분의 연구는 이미지와 같이 고정 차원을 가진 연속값 데이터에 집중되어 왔다.

- 순차적인 데이터를 모델링하는 것은 상대적으로 덜 흔함. 특히 음악 악보와 같은 이산 토큰의 순서열을 다루는 경우 AutoRegressive 의 사용이 필요.

  > ☝️ **Why?**
  >
  > Autoregression 이 더 효과적인 이유는 autoencoder가 latent code를 무시하기 때문이다.

- deep latent variable models 은 짧은 sequence에서는 일부 성공을 보여주었지만 아직까지 매우 긴 sequence에 대해서는 성공적으로 적용되지 못했다.
- **hierarchical recurrent decoder**
  - **긴 sequence에 대한 modeling 문제 극복을 위해 도입**
- 이 논문에서는 음악 노트의 시퀀스를 모델링하는 응용에 초점이 맞추고 있다.
  - 서양의 대중 음악은 악곡의 박자와 구절 간의 반복과 변주와 같은 강한 장기 구조를 보여줌
  - 이러한 구조는 또한 계층적
    - ex) 곡 ⇒ 섹션 ⇒ 마디 ⇒ 박자
  - 음악은 기본적으로 다중 스트림 신호
    - 강한 상호 의존성을 가진 여러 명의 연주자가 참여하는 경우가 많다.

2. Background

기본적으로 본 모델은 AutoEncoder 이다.

즉 입력을 정확하게 재구성하는 것이 목표이다. 그러나 새로운 샘플을 생성하고 Latent space에서 보간 및 속성 벡터 연산을 수행할 수 있는 능력도 원한다.

⇒ 이를 위해 VAE 프레임워크를 사용하였다.

- **Variational Autoencoders**
  - AutoEncoder에 적용되는 일반적인 제약 조건은 input에 대한 관련 정보를 낮은 차원의 잠재 코드로 압축한다는 것.
    - VAE는 Latent code z 가 사전 확률 p(z)에 따라 분포된 무작위 변수임을 제약 조건으로 도입

> ☝ **데이터 생성 모델: z ~ p(z), x ~ p(x|z)**

- posterior p(z|x) 을 근사화하는 인코더 qλ(z|x)와 가능도 p(x|z)를 매개변수화하는 디코더 pθ(x|z)로 구성

  ⇒ VAE = Encoder + Decoder

- **VAE 학습 과정은 인코더와 디코더의 매개변수를 최적화하면서 ELBO를 최대화 함 (KL Divergence를 최소화)**

  ⇒ VAE는 데이터의 분포를 모델링하고, 잠재 공간에서의 유의미한 구조와 속성을 학습할 수 있다.

- **β-VAE AND FREE BITS**
  - VAE의 핵심 개념은 ELBO

    > ☝ **ELBO = E[log pθ(x|z)] + KL(qλ(z|x)||p(z))**

  - E[log pθ(x|z)] 는 qλ(z|x)로부터 추출된 z 샘플에 대해 p(x|z)가 높아야 함을 요구 ⇒ 정확한 재구성을 보장
  - KL(qλ(z|x)||p(z)) 는 qλ(z|x)가 사전 분포와 가까워져야 함을 장려 ⇒ p(z)로부터 잠재 벡터를 샘플링하여 실제적인 데이터를 생성할 수 있게 합니다
  - ELBO를 조절하는 방법

    1) KL 가중치 하이퍼파라미터 β를 사용하는 것

    2) KL 정규화 항목을 일정 임계값 이상일 때에만 적용하는 것

- **LATENT SPACE MANIPULATION**
  - AutoEncoder의 광범위한 목표는 데이터의 압축된 표현을 학습하는 것
  - Latent space는 창의적인 응용을 위해 활용될 수 있다.
  - Latent space에서 매핑된 포인트들은 의미적으로 유사한 데이터 포인트들에 매핑되어야 한다.
  - Latent space는 매끄럽고 빈틈 없는 특성을 가져야 하며 의미 있는 의미적 그룹을 구분해야 한다.
  - 앞에서 말한 조건들은 홀드아웃된 테스트 데이터에서 likelihood와 KL divergence 항이 작으면 만족하게 된다.

        - Latent space의 점들 사이에서 보간을 수행하고, 데이터 공간에서 해당 점들이 의미론적으로 의미 있는 방식으로 보간되는지를 테스트할 수 있다.
        - 속성 벡터는 특정 속성을 가진 데이터 집합에 대한 평균 Latent vector로 계산된다.
        - 모델의 성능은 속성을 잘 발견하고, 속성 벡터 조작 결과가 의미 있는지를 테스트함으로써 확인된다.

- **Recurrent VAEs**
    - 인코더는 입력시퀀스를 처리하고 hidden state의 시퀀스를 출력
    - Latent vector z에 대한 분포의 매개변수는 hT의 함수로 설정
    - 디코더는 sampling된 잠재 벡터 z를 사용,디코더의 RNN의 초기 상태를 설정하고, 자기회귀적으로 출력 시퀀스를 생성
    - 표준 VAE와 마찬가지로 사후 분포 qλ(z|x)를 사전 분포 p(z)에 근사하도록 학습
    - **Recurrent Models 의 단점**:
        1. 자체적으로 시퀀스의 강력한 자기회귀 모델로 사용되는 경우가 많다.
        2. 전체 시퀀스를 단일 잠재 벡터로 압축해야 하는데, 이는 **시퀀스 길이가 증가할수록 문제가 발생**할 수 있다.

3. Model

- 모델은 sequential data에 대한 VAE들에서 사용된 기본 구조를 따르나 새로운 **계층적 디코더를 도입**하여 **긴 시퀀스에 대해 훨씬 우수한 성능**을 나타낸다.

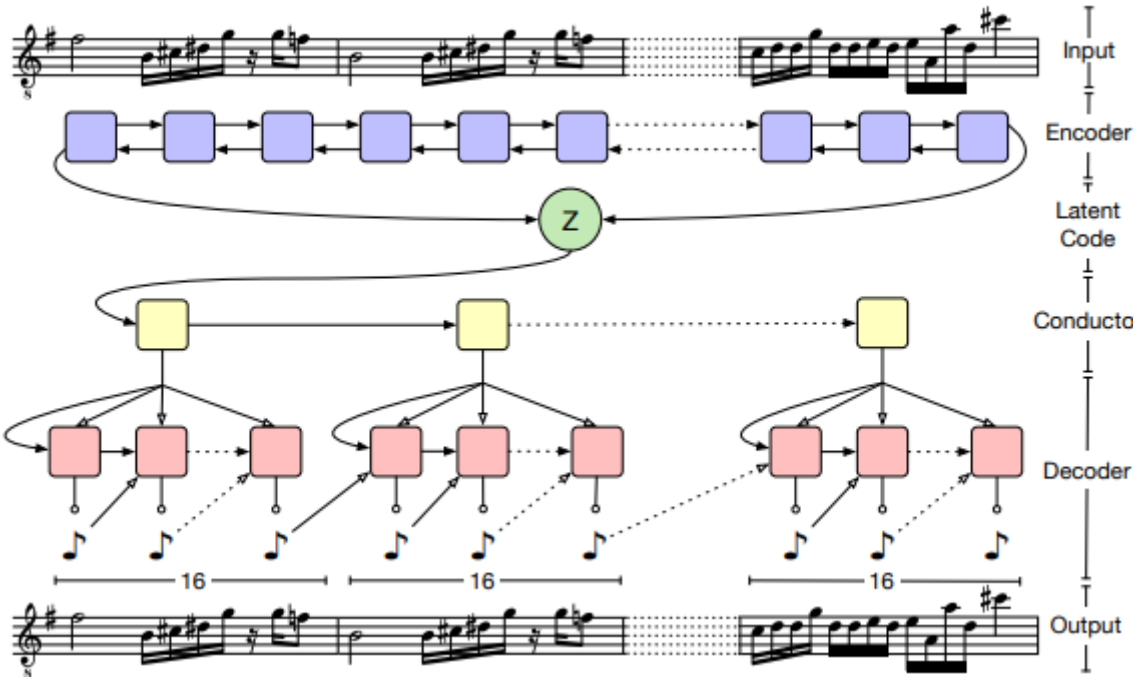


Figure 2. Schematic of our hierarchical recurrent Variational Autoencoder model, MusicVAE.

- **Bidirectional Encoder**
  - Encoder를 위해 우리는 **두개의 양방향 LSTM network를 사용**했다.
  - input seq X={x1,x2, ... , xt} 를 처리하여 두번째 양방향 LSTM 계층에서 최종 상태 벡터 ht -> , <- ht 를 얻게 됨
  - 두 개의 LSTM은 **concatenate** 되어 hT를 생성 후, latent distribution parameter μ 과 σ를 생성하기 위해 **fully connected layer로 공급됨**
  - Whμ, Whσ = 가중치, bμ, b = bias
  - 모든 층에 대해 2048개의 state size, 512 latent dimensions 설정
  - **양방향 순환 인코더의 사용은 장기적인 문맥 정보를 제공**
- **Hierarchical Decoder**
  - decoder RNN은 초기 상태로 Latent vector z가 설정되고, autoregressive하게 출력 시퀀스를 생성
    - ⇒ 긴 시퀀스에 대해서는 간단한 RNN을 디코더로 사용하는 것은 샘플링과 재구성 측면에서 성능이 좋지 않다는 것을 발견
    - Output sequence 가 생성될 때 **latent state의 영향이 사라짐으로써 야기된 문제**라고 생각
  - 이런 문제를 해결하기 위해 계층적 RNN을 사용

- 입력 시퀀스 x를 U개의 하위시퀀스로 나누게 되면, tanh 활성화 함수가 적용된 Fully connected Layer를 통과시킨 z는 각각의 하위 시퀀스와 대응되며 초기 상태의 a “conductor” RNN을 얻는다.
- 그 conductor RNN은 U embedding vecor C를 각 하위 시퀀스마다 한개씩 만든다.
- 현재 모델에서는 2 layer LSTM 을 사용하며 decoder RNN 레이어 하나당 1024 유닛 생성
- Autoregressive RNN decoder는 여전히 “posterior collapse” 문제가 존재

  ⇒ **Latent code를 사용할 수 있도록 decoder의 범위를 제한하는 것이 긴 구조를 모델링 할 때 중요하다는 것을 발견**
- 출력 시퀀스 안에서만 state 를 전달할 수 있도록 하여 decoder 에서 하위 레벨 RNN 의 유효 범위를 줄임
- 각 디코더가 long term context 을 얻는 유일한 방법은 conductor 에 의해 생성된 embedding 을 사용하는 것이며 이는 결국 latent code 에만 의존한다는 것을 의미

- **Multi-Stream Modeling**
  - 음악은 기본적으로 다중 Stream 신호로 구성됨
  - Output tokens 와 관련하여 3개의 각기 다른 distributions을 만든다는 것을 제외하고는 basic MusicVAE 와 동일한 “trio”모델을 소개함으로써 가능성을 탐구
  - 계층적 디코더 모델에서 이러한 별개의 stream 을 계층의 직교 차원으로 간주하고 각 계층에 대해 별도의 decoder RNN 을 사용
  - Conductor RNN 의 embedding 은 별도의 fully connected layer 를 통해 각 계층의 RNN 상태를 초기화 한 다음 tanh activation 을 수행
  - baseline인 flat 디코더에서는 단일 RNN을 사용하고, 악기별 softmax를 도출하기 위해 출력 결과를 분리

4. Experiments

- 음악 데이터에 대해 일련의 양적 및 질적 연구 수행
- 먼저 간단한 순환 VAE가 음악 음표의 짧은 시퀀스를 효과적으로 생성하고 보간할 수 있는 지를 보여주고 그런 다음 효과적으로 모델링하기 위해 신규 계층적 디코더가 필요한 상당히 긴 음표 시퀀스로 이동.
- 이 주장을 검증하기 위해, 기준과 비교하여 데이터의 재구성, 보간, 속성 모델링 능력이 현저하게 향상되었음을 양적으로 입증

- 제안된 모델이 샘플의 지각된 품질을 현저하게 향상시킨다는 것을 보여주는 청취 연구를 수행
- **Data and Training**
  - 각 악기에서 연주할 음표에 대한 지시사항과 박자 정보를 포함하고 있는 150만개의 MIDI 파일을 사용
  - 2-bar, 16-bar 멜로디, 2-bar와 16-bar 드럼 패턴, 그리고 멜로디 라인, 베이스 라인, 드럼 패턴으로 구성된 16-bar의 “Trio” 시퀀스의 훈련 데이터를 MIDI 파일로부터 수집
  - 단음표 멜로디와 베이스라인을 16분음표 이벤트의 시퀀스로 모델링 ⇒ 130 차원의 출력 공간이 생성
    - 멜로디: 128개의 MIDI 피치에 대한 **“note-on” 토큰** 128개와 **note-off**, **rest 토큰**으로 구성됨
    - 드럼 패턴: 61개의 드럼 클래스를 9개의 대표 클래스로 매핑하고, 512개의 범주형 토큰으로 가능한 모든 연주 조합을 나타냄
  - 타이밍에 대해서는 모든 경우에 16분음표 간격으로 음표를 양자화하여 각 마디가 16개의 이벤트로 구성되도록 했다.
  - 모든 모델은 **Adam을 사용하여 학습**되었으며, 학습률은 지수적 감소율 0.9999로 10^(-3)에서 10^(-5)로 줄여가며 배치 크기는 512로 설정
- **Short Sequences**
  - 음악 시퀀스를 순환성 VAE로 모델링 하는 것이 가능함을 입증하기 위해, 먼저 2-bar(T=32)의 단음표 음악 시퀀스를 flat decoder로 모델링

    ⇒ 입력을 매우 정확하게 재구성 하는 것을 발견
    - **posterior collapse** 나 **exposure bias** 문제 없이 Latent code를 효과적으로 사용하는 방법을 학습
  - 그러나 16-bar 시퀀스를 재구성하기 어려웠다.(긴 시퀀스)
    - **Teacher-forced 와 샘플링 재구성 정확도 사이의 불일치가 27 % 이상 증가**

> ☝️ 계층적 티코더를 설계한 동기!!

- **Reconstruction Quality**

- Hierarchical decoder 가 16-bar 멜로디와 드럼 패턴에 대해 더 나은 재구성 정확도를 제공하는지 여부 평가
- flat decoder를 사용했을 때는 posterior collapse 징후가 보이며, teacher-forcing이 추론에서 제거될 때 **정확도가 약 27-32% 감소**
- 반면, Hierarchical decoder는 샘플링의 정확도와 teacher-forcing의 차이는 약 **5-11%로 유지**
- Hierarchical 모델은 flat model보다 **더 높은 정확도**를 달성하면서 teacher-forcing과 샘플링 **성능 간의 차이가 훨씬 작은 것을 확인**시켜줬다.

- **Interpolations**
  - 보간 실험에서 계층적 디코더는 부드럽게 변하는 의미 있는 보간을 제공
  - 데이터 공간에서의 보간은 원래 멜로디보다 낮은 확률을 가지며, flat model은 좋은 결과를 보여주지만 hierarchical model 보다 일관성이 떨어진다.
  - hierarchical decoder를 사용한 보간은 의미 있는 음악적 특징을 합성하는 반면, 데이터 보간은 조화 및 리듬적인 불일치를 초래한다.

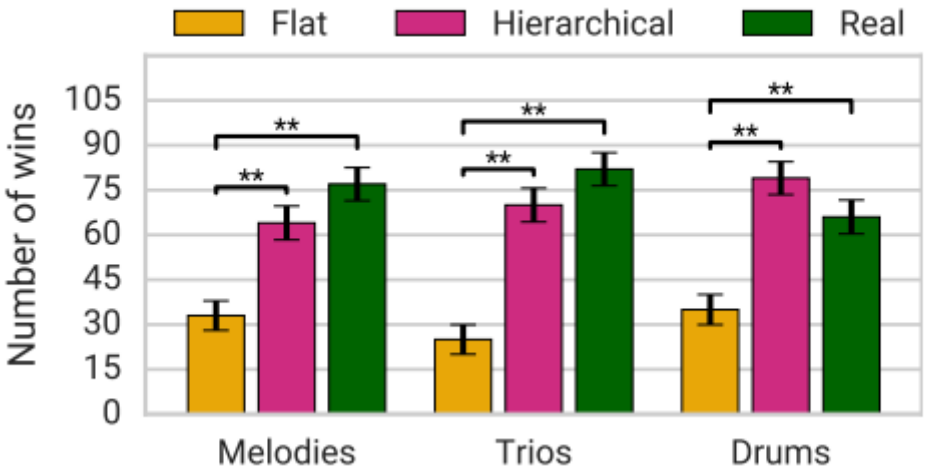
> ### Q. 보간의 개념은?
>
> - 두 개 이상의 점, 값 또는 시퀀스 사이에서 중간 값을 추정하는 과정
> - 음악 분야에서는 보간이 음악적인 속성을 유지하면서 두 개 이상의 음악 시퀀스 사이에서 부드럽게 연결되는 중간 시퀀스를 생성하는 기술

- **Attribute Vector Arithmetic**
  - 잠재 공간의 구조를 활용하여 “속성 벡터”를 사용하여 주어진 시퀀스의 속성을 변경 가능
    - 다섯 가지의 속성 정의: diatonic membership, note density, average interval, 16 번째 8번째 note syncopation 을 정의
  - 이를 테스트하기 위해 먼저 37만 개의 무작위 학습 예제에서 이러한 속성을 측정
  - 각 속성에 대해 해당 속성을 나타내는 양에 따라 집합을 정렬하고 사분위수로 분할, 상위 사분위수의 평균 잠재 벡터에서 하위 사분위수의 평균 잠재 벡터를 빼서 속성 벡터를 계산

- 특정한 속성 벡터를 적용하는 것이 목표 속성에 의도된 변화를 일관되게 생성한다는 것을 발견
- 한 속성을 증가시키면 다른 속성이 감소하는 경우도 발견

  ⇒ 이는 휴리스틱이 중복된 특성을 포착하기 때문이라고 생각한다.

- **Listening Tests**



- 청취 테스트 결과, 계층적 디코더를 사용하면 샘플 품질이 크게 향상되었음을 명확하게 보여줌. 모든 경우에 **계층적 모델**은 평면 모델보다 **훨씬 더 많은 비율로 선호**되었으며, 평가 데이터와 동일한 비율로 선호

5. Conclusion

- MusicVAE 모델을 제안하며 flat baseline 보다 **현저히 우수한 성능을 달성한다는 것을 정량적 및 정성적 실험을 통해 철저히 입증**했다.