KSC2024 한국소프트웨어 종합학술대회

STYLE-VT: Variational Transformer와 대리코드 이론을 적용한 스타일 기반 코드 생성모델

송교원(tim000519@gmail.com) 서동영(lazymoon180@gmail.com) 나준수(najunsoo4439@gmail.com) 양희철(hcyang@cnu.ac.kr)
Division of Computer Convergence, Chungnam National University

1 서론

본 연구의 주 목적은 코드의 생성 과정에 음악의 장르적 특성을 반영하는 것이다. 코드진행은 음악에서 화성의 틀을 구성할 뿐만 아니라, 음악의 장르별 특징 또한 드러낼 수 있는 중요한 요소이다.

연구 내용 및 기존 연구와의 차별점

1. 'Structured', 'Explainable'

Transformer -멜로디와 화음 간의 장거리 의존성을 학습하고, **구조적인** 화음을 생성

Variational Auto Encoder - 장르에 따라 나타나는 특징적인 코드의 분포를 반영.

2. 기존 음악에 대한 편곡

입력 멜로디가 그대로 유지되기 때문에, '편곡' 작업에 활용될 수 있는 방향성을 가짐

3. 작곡가의 의도 반영

멜로디를 유지하며 장르에 어울리는 코드를 생성함으로써 작곡가의 의도를 효과적으로 반영 가능

2 제안 방법

◆ 데이터셋

- MusicXML형식으로 저장되어있는 연구용 공개 데이터셋
- MuseScore에서 직접 악보를 다운받아 MusicXML 형식으로 변환

◆ 음표 및 화음 표현

- 멜로디의 음표는 12개의 음계 클래스와 나머지를 나타내는 1개의 클래스(쉼표)로 구성
- 모든 화음은 Triad코드의 Major, Minor, Diminished와 Seventh코드의 Major, Minor, Dominant로 표현

◆ 배치 처리

- 각 배치의 길이는 8마디로 설정하고, 2마디씩 겹치도록 설정하며, 모든 곡은 4/4박자로 설정한다.

◆ 12키로 변환

- 각 곡을 12키, 즉 모든 조표로 조옮김 하여 사용한다.

Style-VT모델 설계

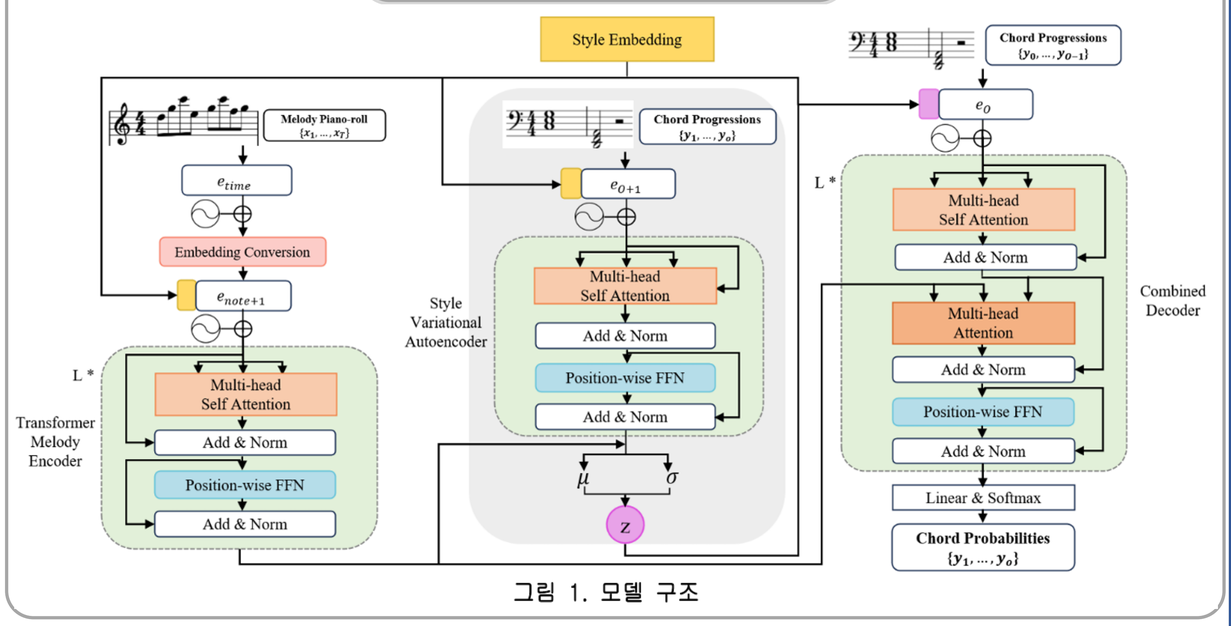


그림 1. 모델 구조

Transformer Melody Encoder

Transformer Melody Encoder(TME)는 입력 x1:T를 받아 멜로디의 각 음표에 대한 context를 캡처한다.

1. 시간 단위 임베딩

입력 x1:T는 임베딩 레이어를 통해 시간 단위 임베딩 벡터 eT로 변환된다.

2. 시간 단위 임베딩을 음표 단위 임베딩으로 변환

시간 단위 임베딩 eT 에 위치 임베딩 wT 를 더한 후, 이 임베딩을 'TimeToNote' 절차를 통해 음표 단위 임베딩 eN으로 변환한다. 그 뒤 장르와 조표를 노트 기반 멜로디 임베딩의 시작에 추가하여 조건부 입력으로 사용한다.

3. 셀프 어텐션 블록

음표 단위 임베딩 en에 위치 임베딩 wN을 더한 후, L개의 멀티 헤드 셀프 어텐션 블록을 통과한다.

Style Variational Encoder

Style Variational Encoder(SVE)는 Transformer 멜로디 인코더의 출력, 코드 입력 y , 그리고 스타일 토큰 c 로부터 잠재 표현 z를 추론한다.

Combined Decoder

오른쪽으로 쉬프트된 목표 화음 시퀀스를 받아 인코더 출력 Enc($x1{:}T$)과의 어텐션을 계산하여 목표 화음을 예측한다.
c와 잠재 변수 z를 추가로 사용하여 조건부 입력으로 활용한다.

1. 임베딩 및 어텐션:

오른쪽으로 쉬프트된 화음 시퀀스 (y0:o-1) 는 임베딩 레이어를 통과한다.

2. 화음 예측

최종 선형 레이어와 소프트맥스 활성화를 통해 화음의 확률을 예측한다.

Loss Function

본 연구에서는 대리 코드 이론 (Chord Substitiution Theory)를 적용하여 모델의 목적 함수를 재정의하였다.
Style-VT 를 학습하기 위한 목적 함수는 다음과 같이 정의된다.

❖ $\mathcal{L}_{style-\mathrm{VT}} = \mathrm{RE} + \lambda_S \mathrm{SE} + \lambda_K \mathrm{KLD}$

여기서 RE 는 모델 출력과 목표 값 간의 차이를 측정하는 재구성 오류(Reconstruction Error)이며, 이는 다음과 같다.

❖ $\mathrm{RE} = \mathbb{E}_q(z|x,y,c)\,[-\log p_\theta(y|x,z,c)]$

SE 는 대리 코드가 재구성 과정에 미치는 영향을 반영한 대체 재구성 오류(Substitution Reconstruction Error)로, 다음과 같다.

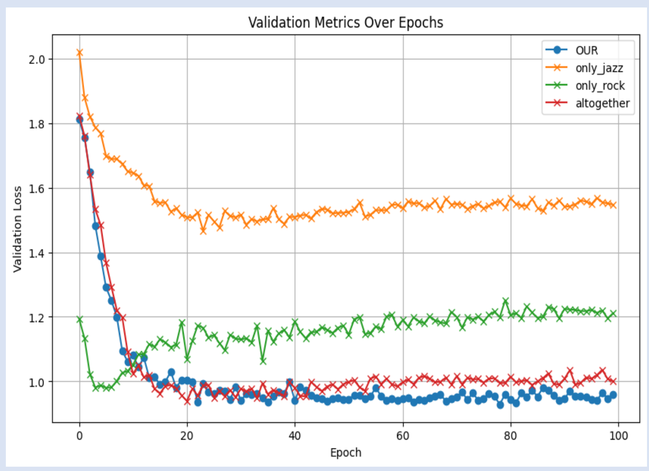
❖ $\mathrm{SE} = -\sum_{i=1}^{N}\sum_{k=1}^{c} v_{yk}\log(p_{ik})$

KLD 는 잠재 변수의 분포를 정규화 하는 쿨백-라이블러 발산으로, 다음과 같이 정의된다.

❖ $\mathrm{KLD} = \mathrm{KL}\left(q_\phi(z|x,y,c)\,||\,p(z)\right)$

3 실험 결과

모델별 Validation Loss 비교



제안 모델인 Chord Generation Variational Transformer (이하 CGVT)와 비교 모델의 테스트 데이터에 대한 epoch별 예측 성능을 그래프로 나타낸 것이다.

비교 대상 모델은 각각 Jazz 데이터셋만 학습한 모델, Rock 데이터셋만 학습한 모델, 장르를 고려하지 않고 모든 데이터를 학습한 모델이다. 위 모델 모두 Transformer 기반의 모델로, 장르 임베딩 값을 이어 붙이지 않았다는 점을 제외하면 기본 구조가 완전하게 동일한 모델이다.

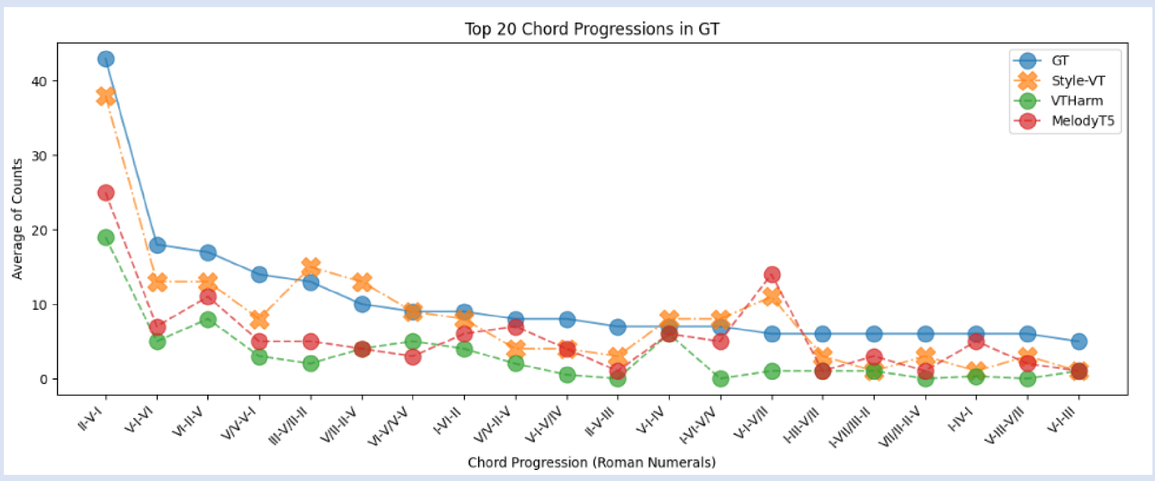
실험 결과 장르 정보를 임베딩하는것이 모델 학습에 영향을 준다는 것을 알 수 있으며, 즉 장르별로 사용되는 코드 빈도의 분포가 다르다는 것을 보인다.

모델별 Metric 비교

Metric / Model	Harmonic Similarity			Diversity and Coherence					
	TPSD↓	DICD↓	LD↓	CHE↑	CC↑	CTR↑	PCS↑	CTD↓	MTD↓
GT	-	-	-	1.629	6.688	1.319	0.385	0.623	1.403
VTHarm [8]	2.987	144.567	0.905	1.9571	8.0039	1.2162	0.3671	0.6862	1.4321
MelodyT5 [21]	2.695	142.047	0.907	1.538	6.238	1.210	**0.441**	0.695	**1.4054**
Style-VT (w/o style embedding, $\lambda_S = 0.1$)	2.799	135.533	0.862	**1.9660**	**8.0643**	1.2301	0.3798	0.7053	1.4297
Style-VT (style embedding, $\lambda_S = 0$)	2.903	134.795	0.867	1.9258	7.693	1.2136	0.3763	**0.6653**	1.4138
Style-VT (style embedding, $\lambda_S = 0.1$)	**2.647**	**131.493**	**0.843**	1.9426	7.8536	**1.2632**	0.3858	0.6942	1.4226

CHE와 CC는 화음의 다양성을 측정한다. CTD는 화음 전환의 일관성을 측정하고, CTR, PCS, MTD는 멜로디와 화음 간의 일관성을 측정한다. 제안 모델 CGVT는 코드의 다양성을 나타내는 지표인 CHE와 CC에 있어서 높은 수준의 결과를 보였다. 이는 이 모델이 높은 화음 다양성을 가지고 있음을 나타낸다. 또한 일관성을 나타내는 지표들의 값 또한 최상, 혹은 두 번째로 높은 결과를 보인다. 일반적으로 화음의 일관성과 다양성 간의 트레이드 오프가 있다고 알려져 있음에도, 본 모델은 높은 수준의 화음 생성 성능을 보인다.

모델별 장르의 코드 진행 비율 비교

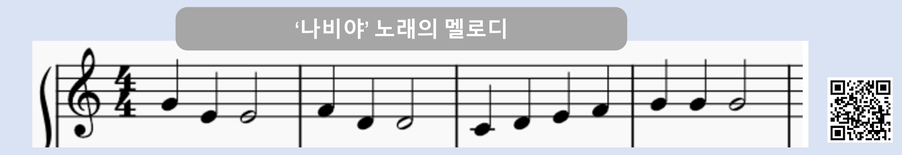


장르별 특성이 반영되었는지 평가하기 위해, 생성된 음악의 코드 진행 빈도를 산출하였다. Ground Truth는 사람이 작곡한 원곡 재즈에서 나타난 코드 진행의 빈도이며, Style-VT는 장르 임베딩을 재즈 심벌로 지정해주고 생성하였을 때 나타난 코드 진행의 빈도이다. VTHarm은 선행연구의 모델이며, 모든 모델에 대해서 같은 테스트 셋이 사용되었으며, 테스트 셋은 학습에 사용되었던 데이터와 중복되지 않았다. 이 지표를 통해 장르 임베딩을 도입한 Style-VT 모델은 해당 장르의 코드 분포를 잘 모방하여 코드를 생성한다는 것을 알 수 있다.

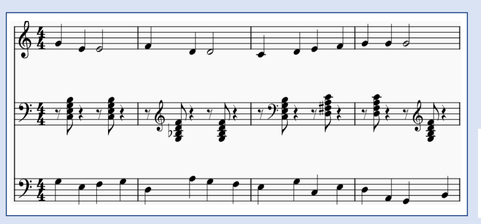
4 모델 음악 생성 결과 / 향후 연구 계획

모델 음악 생성 결과

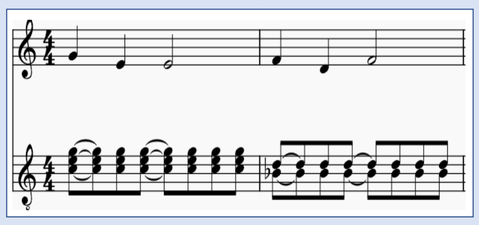
'나비야' 노래의 멜로디



↓ "재즈" 장르로 생성



↓ "락" 장르로 생성



Style-VT는 멜로디와 장르 정보를 입력으로 받아, 멜로디에 어울리는 코드를 입력받은 장르의 코드 분포를 반영하여 생성해준다.
즉, 장르에 어울리는 코드가 생성되는 것이다. 우리는 이에 더 나아가 생성된 코드에 박자를 부여하거나 코드 구성음을 스케일링 하는 등의 기법을 통해 장르가 가지는 특징을 더 뚜렷하게 드러내고자 한다.

위의 최종 산출물 음원은 우리의 연구를 통해 실제로 생성된 악보로, 멜로디에 대한 장르를 재즈로 선택하여 모델에 입력하면 내부 파일이 자동적으로 코드에 박자를 부여하고 구성음을 스케일링하여 위의 그림과 같은 악보가 최종적으로 산출된다.

본 프로젝트의 향후 연구 방향은, 재즈 뿐만이 아니라 락이나 블루스 등 다양한 장르에 대하여 생성된 코드진행에도 장르의 특징을 더 잘 반영할 수 있는 추가적인 방법을 고안하여, 음악의 작곡 혹은 더 나아가 편곡까지 아우를 수 있는 프로그램을 생성하고자 한다.