BERT を用いた音楽構造理解モデルにおけるエンコード方式の提案

荒木 真† 森 康真‡ 田村 慶一‡

†‡広島市立大学大学院情報科学研究科 〒731-3194 広島市安佐南区大塚東三丁目 4 番 1 号

†E-mail: †mh67001@e.hiroshima-cu.ac.jp, ‡{mori, ktamura}@hiroshima-cu.ac.jp

あらまし 音楽構造理解とは MIDI データ等の記号データから音楽構造の関係性を理解することを指し,メロディ分類,ジャンル分類といった分類タスクに応用する研究が挙げられる.これらの音楽タスクには分類のためのデータ付きラベルが必要であり,学習データを確保することが困難であることが課題である.そこで,ラベルを必要としない BERT を用いた大規模な事前学習モデル構築が音楽構造を理解することに貢献し,ファインチューニングを用いて様々なタスクに応じたモデル構築が行う事ができる.本研究では,BERT モデルにおいて,計算コストの観点から入力長を削減するデータ構造と,そのデータ構造に考慮したマスク戦略を用いて音楽構造理解モデルを構築し,音楽分類タスクによるモデルの評価を行い,データ構造による計算コストを考察する.

キーワード 音楽,データ分類,自然言語処理応用,BERT,音楽構造理解

1. はじめに

音楽の構造には音階,和音の構成,コード進行などの音楽理論に基づいたルールを持っていることから機械学習等の手法を用いてその構造を分析することが可能と考えられている[1]. 2000年代頃の先行研究では音楽データとして周波数方式のオーディオ信号を対象とした研究[2,3]が主に行われていた.一方で電子楽器などの演奏データを記号トークンとして扱うMIDI規格のデータを対象とした研究[4,5]が近年になって研究されてきている.MIDIデータは音楽をトークンデータとして扱う事が可能であり，扱うデータ量としてオーディオ信号より小さい点が挙げられる[6]．しかし，MIDI規格のデータを対象とした場合,ラベル付きデータセットが少ない事[7]が課題としてあり，効果的な教師あり学習が困難である．ラベル付きデータの不足に対しての対策として大量のラベル無しデータによる事前学習と少量のラベル付きデータによるファインチューニングの2段階に分ける方法がある[8]．この方法は自然言語処理分野(NLP)において用いられる手法でありTransformer[9]を用いたモデルが分類や生成などのタスクにおいて主流となっている[10]．しかし,NLPで用いられる手法は文章によるテキストデータを対象としたモデルであり，MIDIデータを直接適用する事は困難である．MIDIデータによるトークンデータは音楽の要素である音程やその音符の持続時間と強弱,小節間の位置等の情報を含んでおり，音楽情報を保持する符号方式を考慮する必要がある.また,Transformerを用いたモデルは自己注意機構self-Attentionによって単語間の類似度を入力されるすべての単語ごとに求める手順がある．これにより,入力長が大きくなる事による二次元的計算コストが要求されることから入力長は可能な限り削減する必要がある．本論文では，対象データとしてMIDIデータを扱い,音符データを複数にまとめて表記するMeasure方式の提案とBERTを用いた音楽構造理解モデルを用いた楽曲の分類タスクを行い,従来の符号方式との比較を行い,BERTによる計算コストの関係を考察する.

2. 関連研究

自然言語処理分野でのBERT[11]システムは文章の関係性を学習した事前学習モデルを用いる事でファインチューニングによる調整を行うことで様々なタスクに応用できる事が可能となった[12,13].これは事前学習にはラベル付きデータを必要としない事から,大規模な学習を行う事が可能である,ラベル付きデータが少量であっても学習が可能な事が利点として挙げられる.しかし,音楽におけるBERTシステムを用いた研究は少なく,音楽用の事前学習モデルとしてはピアノトラックのみを対象として約4000曲のMIDIデータから事前学習を行ったMIDIBERT[14]やマルチトラックに対応した100万以上のMIDIデータから事前学習を行ったMusicBERT[15]が先行研究として挙げられる.この二つの研究は事前学習を行い音楽用の事前学習モデルの構築とファインチューニングを用いてメロディ単位の分類と楽曲単位の分類を行うことでモデルの評価を行っている.これらの研究はデータセットやパラメータなどの情報が公開されており,このBERTを用いた音楽構造理解モデルのベンチマークとして捉えることができる.我々の研究では計算資源の関係から,MIDIBERTを主に参考とし,実験設定のパラメータなどもこの文献によるものとする.

3. 手法

本章では,MIDI データをトークンデータに対応させるエンコード方式と BERT による学習手法であるマスク戦略とモデルの構造について述べる.

3.1 エンコード方式

BERT に MIDI データを入力するには数値データに変換するエンコードを行う必要がある.文章におけるエンコードは一連のトークン列として扱えば文章を表現できるが,音楽では同時刻に複数の音符が演奏されることから,時間の情報を考慮する必要がある.従来の方式では MIDI データを 4 つの要素別に扱う方法が提案されており,各要素の説明は以下に示す.

- **Bar**. 小節の区切りを表す要素で,‘Bar new’と‘Bar cont’はそれぞれ小節の始まりと途中であることを示す.
- **Position**. 小節の位置を表す要素で,16分割で‘Position 9/16’は3泊目を示す.
- **Pitch**. 音符の音程を表す要素で,MIDI規格に従い,最も低い音程を‘Pitch 22’,最も高い音程を‘Pitch 107’で表される.
- **Duration**. 音符の持続時間を表す要素で,0~63の範囲で表され,最大‘Duration 32’は全音符を示す.

3.1.1 従来のエンコード法

REMI[16]とCP[17]と呼ばれるエンコード方式があり,従来モデルのMIDIBERTにもこのエンコード方式が用いられている.REMI方式は各要素を文章のように一連のトークン列として表記し,CP方式では次元を増やすことで上記の4要素を持つ音符情報として表記される.また,BERTの入力には固定長のデータにしなければいけない事から,データが固定長になるようにパディング処理を行っている.REMI方式とCP方式の比較についてはCP方式の方が学習速度と精度において上回る結果が示されており,REMI方式に関しては本実験では省略し,CP方式のデータ図を図1に示す.

3.1.2 Measure方式

Transformerを用いたモデルは自己注意機構Attentionによって単語間の類似度を入力されるすべての単語ごとに求める手順がある. この手順はこれにより,入力長が大きくなる事による二次元的計算コストが要求されることから入力長は可能な限り削減する必要がある[16].そこでCP方式と同様に次元を増やす操作を行い(入力長,音符数*4)の次元数で表記するMeasure方式を提案する.これはCP方式が2次元目に1つの音符を表記する方式に対して,Measure方式では複数の音符を表記する方式である事を意味する.本実験ではパディング処理による使用しないデータが極端に増えない様に音符数を4つとし,(入力長,16)の次元数のデータを図2に示す.これは複数の音符を一定の纏まりで表現することで,GPUの並列処理による計算資源を効率化と計算速度の向上を図る.

3.2 マスク戦略

BERTによる事前学習にはMasked Language Model(MLM)[11]と呼ばれる学習手法を用いている.先行研究[15]では,マスク処理を行うトークンをランダムに行うマスク戦略に加え,1音符全てをマスクする‘octuple masking’と小節毎に同要素をマスクする‘bar-level masking’を用いており,各タスクにおいて若干の精度ではあるが,‘bar-level masking’が良いモデルとしている.しかし,若干の精度差による結論になることに加え,CP方式との比較から音符レベルのマスク戦略を採用する.

3.3 モデル構造

本実験で扱うモデルはMIDIBERT[14]を参考とし,111Mのパラメータをもつ事前学習モデルを構築した.各入力トークンは埋め込み層を通して256個の分散表現に変換され,12個のTransformer層による自己注意機構による相対位置エンコーディング[18]を行う.その後にマスクされたトークンを予測するように全結合層により各要素の確率分布になる様に学習を行う.次に図3にMeasure方式による事前学習モデル構造を示す.またファインチューニングにおいては音楽構造の関係を事前学習において得ている事から単純なネットワーク構築にしており,事前学習による確率分布の出力を全結合層とDropout層による伝搬を行っている.先行研究と異なる点は事前学習モデルにおいて埋め込み層に入力される情報がCP方式では4つのトークンであるのに対して,Measure方式では,複数の音符を埋め込む必要がある.そこで,埋め込み層を16個のトークンを入力するために,16個の埋め込み層を用意し,256個の分散表現を16個持つ(16,256)の次元数となる様に構築した.このモデル構造の詳細を図3に示す.その他のモデルの構築は参考文献[14]と同等の設定で行っている.

図1 従来方式CP方式

図2 Measure方式

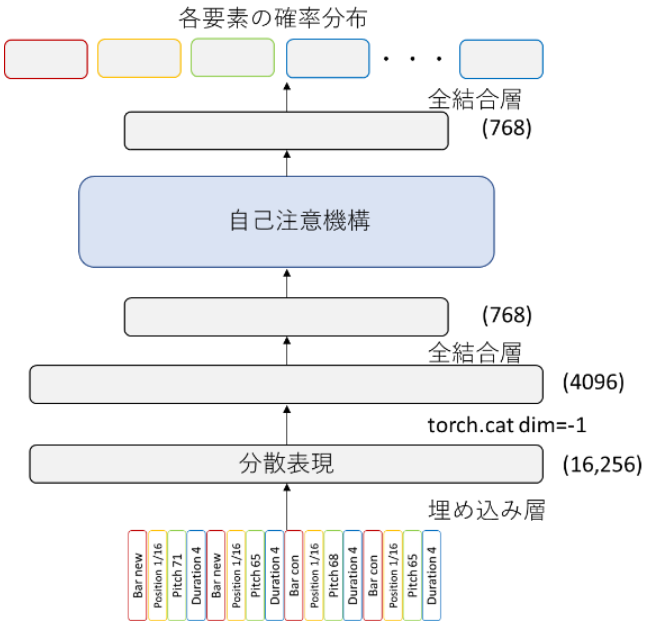

図3 Measure方式による事前学習モデル構造(括弧は次元数を示す)

4.実験詳細

本実験ではタスクとして作曲者分類と感情分類を選択した.これらのタスクについてはMIDIBERTと同様に設定したものであり,詳細は省略する.データセットはMIDIBERTで用いられるPOP909,Pianist8,EMOPIAを対象とした.各データセットの概要を示す.

- **POP909.**
  909曲のポップソングピアノカバーから構成される楽曲の主旋律,伴奏,その他の旋律のトラックに分けられる.
- **Pianist8.**
  作曲者分類に用いられる有名ピアニストのオリジナルピアノ曲で構成される.曲数は411曲で8人の分類となる
- **EMOPIA.**
  楽曲の感情関連タスクのためにYOUTUBEから収集したPOPピアノ曲の新しいデータセットである(日本のアニメ, 韓国と西洋pop,映画サウンドトラック, 個人的な楽曲 387曲).感情ラベルは感情価と覚醒の軸があり4分類となる.

事前学習において使用されるモデルはMeasure方式によるBERTモデルと先行研究MIDIBERTモデル[14]を比較のために用いる.用いるデータセットの85%をトレーニングデータとし,15%をテストデータとする.また,マスク戦略において従来の割合と同様に全体の15%の内から80%をマスクトークンとし,残りの10%をランダムに置き換える処理,最後の10%には処理をしない操作を行う.学習で使用するGPUにはNVIDIA TITAN RTXを用い,機械学習ライブラリはPyTorchである.最適化手法はAdamWOptimizerを用い,学習率は0.01とする.扱うデータセットは上記の全データセットによる2376曲で最大500エポック数として学習を行い,ファインチューニングには作曲者分類にPOP909,感情分類にEMOPIAによるラベルにおいて教師あり学習を行う.学習において事前学習では30回,ファインチューニングでは5回の交差エントロピー更新が無い場合は学習を停止する.学習に用いる.

5.BERTによる事前学習とファインチューニング

Measure方式とCP方式による学習速度と使用メモリについての結果を示す.また,事前学習のマスク戦略による精度とファインチューニングにおける精度の結果も示す.

5.1 事前学習

事前学習に各エンコード方式おいての精度グラフを図4,図5に示す.CP方式では432エポック数でテストデータに対しての精度が0.7314,Measure方式では407エポック数でテストデータに対しての精度が0.8383で学習を終了した.事前学習時においてはCP方式よりもMeasure方式の精度が高く,音楽構造における音符同士の関係値を学習できている.次に,各分類タスクにおけるファインチューニングの精度と計算コストの指標として学習が終了するまでの時間とNVIDIA System Management Interfaceによる Memory-Usage と GPU-utilizationの値を表1に示す.なお,計算コストの指標は事前学習時の値とする. 次に,分類においては作曲者

分類においてはMeasure方式の方が精度が高く,逆に感情分類においてはCP方式の方が精度が高い結果となっている.精度低下の原因として感情分類に使用したデータセットEMOPIAが他のデータセットに比べ,1曲30秒程の長さで構成されている事が原因として考えられる.EMOPIAのデータセットではMeasure方式にエンコードする際にデータが小さい事でパディング処理する部分が多い事がわかっている.次に計算コストにおいては学習時間において約58%の時間で学習を行えており,この点においてMeasure方式は有効であるといえる.また,Memory-Usageの値はほぼ同等であり,GPU-utilizationの値が22%削減できていることからMeasure方式における計算コストの削減ができている事がわかる.

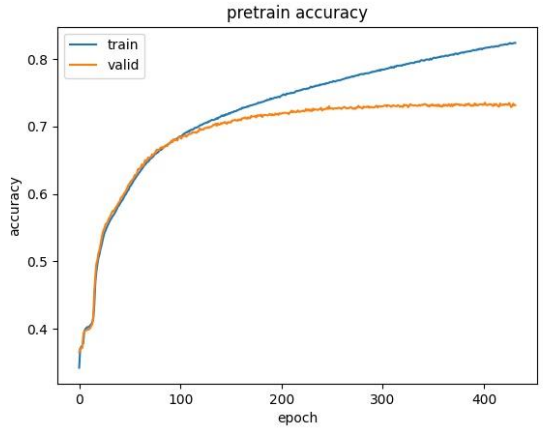

図4 CP方式における事前学習時の精度

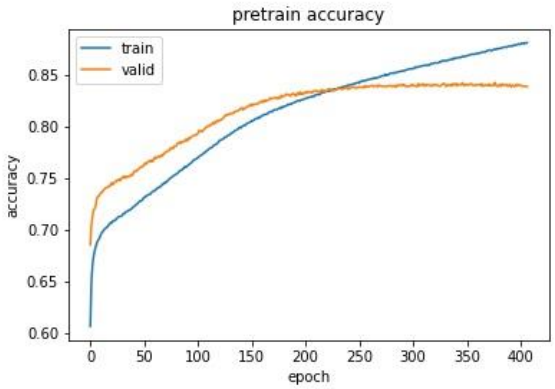

図5 Measure方式における事前学習時の精度

表1 各タスクのファインチューニング精度と計算コスト指標

エンコード方式	作曲者分類	感情分類	学習時間	Memory-Usage	Gpu-Util
CP	0.6984	0.6789	142815s	19623MiB	95%
Measure	0.7551	0.6075	84169s	19857MiB	73%

6.まとめ

本論文では,BERTの2次的計算コストの削減を目的とし,提案手法として音符データを複数にまとめて表記するMeasure方式による音楽構造の関係をもつ音楽構造理解モデルの構築を行った.計算コストの面で貢献があったが,精度においてタスク毎に安定しない事が分かった.また,タスクにおいて特定の分類タスクのみで実験を行ったため,音楽構造理解モデルとして汎用性がある事が議論できなかった.よって今後の課題として,楽曲の推薦やコード認識のような別タスクの追加実験の必要性が挙げられる.

参 考 文 献

[1] J.Briot,G.Hadjeres,and F.Pachet“*Deep Learning Techniques for Music Generation-A Survey*” arXiv preprint arXiv:1709.01620,2017.

[2] T.Li and M.Ogihara,*“Toward intelligent music information retrieval”IEEE Transactions on Multimedi*a,vol.8,no.3,pp.564–574,2006.

[3] M.Levy and M.Sandler,*“Music information retrieval using social tags and audio,”IEEE Transactions on Multimedia*,vol.11,no.3,pp.383–395,2009.

[4] J-P.Briot,G.Hadjeres,and F.Pachet,*“Deep learning techniques for music generation-a survey,”*arXiv preprint arXiv:1709.01620,2017.

[5] E.Waite *et al.*,*“Project Magenta:Generating long-term structure in songs and stories,”*Google Brain Blog,2016.

[6] J.Briot,G.Hadjeres,and F.Pachet,*“Deep learning techniques for music generation-a survey,”*arXiv preprint arXiv:1709.01620,2017.

[7] M.Hamanaka,K.Hirata,and S.Tojo,*“Musical structural analysis database based on GTTM,”* inProc. Int.Soc.Music Information Retrieval Conf.,2014.

[8] X.Han *et al.*,*“Pre-trained models:Past,presentand future,”*arXiv preprint arXiv:2106.07139,2021.

[9] A.Vaswani *et al.*,“*Attention is all you need,”inProc. Advances in Neural Information ProcessingSystems*,2017,pp.5998–6008.

[10] X.Han *et al.*,*“Pre-trained models:Past,presentand future,”*arXiv preprint arXiv:2106.07139,2021.

[11] J.Devlin,M.Chang,K.Lee,and K.Toutanova,*“BERT:Pre-training of deep bidirectional transformers for language understanding,”*in Proc. Conf. North American Chapter of the Association for Computational.

[12] M.Joshi *et al.*,*“SpanBERT:Improving pre-training by representing and predicting spans,”*arXiv preprint arXiv:1907.10529,2019.

[13] Z.Yang *et al.*,"*XLNet:Generalized autoregressive pretraining for language understanding,*"arXiv preprint arXiv:1906.08237,2019.

[14] Yi-Hui Chou,I-Chun Chen,Chin-Jui Chang,Joann Ching,and Yi-Hsuan Yang.*"Midibert-piano:Large scale pre-training for symbolic music understanding"*CoRR,abs/2107.05223,2021.

[15] M.Zeng *et al.*,*"MusicBERT:Symbolic music understanding with large-scale pre-training,"*in Proc.Annual Meeting of the Association for Computational Linguistics,Findings paper,2021.

[16] Y.Huang and Y.Yang,*"Pop Music Transformer:Beat-based modeling and generation of expressive Pop piano compositions,"*in Proc.ACM Multimedia,2020,pp.1180–1188.

[17] W.Hsiao,J.Liu,Y.Yeh,and Y.Yang,*"Compound Word Transformer:Learning to compose full-song music over dynamic directed hypergraphs,"*in Proc.AAAI,2021.

[18] Z.Huang,D.Liang,P.Xu,and B.Xiang,*"Improve transformer models with better relative position embeddings,"*arXiv preprint arXiv:2009.13658,2020.