Vol.2020-MUS-129 No.3
2020/11/2

大局的構造に基づく正則化を用いた自己注意機構付き深層ドラム採譜

石塚 崚斗[1,a)]　錦見 亮[1,b)]　中村 栄太[1,c)]　吉井 和佳[1,d)]

概要：本稿では，音響音楽信号からドラムのオンセット時刻をテイタム単位で推定する手法を述べる．自動ドラム採譜では，フレーム単位で設計された深層ニューラルネットワーク (deep neural network; DNN) により，スペクトログラムを入力としてドラムのオンセット時刻を出力する手法が盛んに研究されてきたが，記号単位でドラム譜の推定を行う研究はまだ少ない．フレーム単位の入力からテイタム単位の出力を行う機構 (採譜モデル) として，エンコーダ・デコーダモデルがある．しかし，DNN の学習に用いるペアデータの量が限られているために，採譜モデルが音楽的に不自然なドラムパターンを生成してしまうという問題点が残されている．このような背景から，我々はフレーム単位の入力特徴量からテイタム単位のドラム譜を推定する採譜モデルを設計する．さらに，採譜モデルの推定結果を音楽的に妥当なパターンに誘導するため，大規模ドラム譜データを用いて学習した言語モデルによる評価値を，採譜モデルの学習時に正則化項として組み込む手法を提案する．このとき，採譜モデルに自己注意機構を導入し，言語モデルに Masked language model (MLM) を利用することで，双方向の文脈から楽曲の長期的な構造を学習することができる．標準データセットを用いた実験により，提案法の効果を示す．

1. はじめに

自動音楽採譜とは音楽音響信号から楽譜を推定するタスクであり，作曲・編曲などの創作活動の補助に応用することができる．ドラムがポピュラー音楽の音楽的な構造を支える重要な楽器であることから，自動ドラム採譜は自動音楽採譜の中でも特に重要な役割を果たしている．本稿では，ドラムの中でも主要な bass drum (BD)，snare drum (SD)，hi-hats (HH) の 3 楽器を扱う．一般的には，自動ドラム採譜ではフレーム単位の入力特徴量からドラムが演奏されている秒数 (オンセット時刻) の推定を行うが，記号単位の推定を行う研究はまだ少ない．

自動ドラム採譜では，深層ニューラルネットワーク (deep neural network; DNN) と非負値行列因子分解 (non-negative matrix factrization; NMF) に基づく手法が代表的であり，入力特徴量であるスペクトログラムからフレーム単位でオンセット確率値を出力し，ピークピッキングを用いてオンセット時刻を検出する [1]．畳み込みニューラルネットワーク (convolutional neural networks; CNN) は時間-周波数領域に着目して特徴量を抽出することにより，高い精度を達成している [2–4]．再帰型ニューラルネットワーク (recurrent neural networks; RNN) を用いることで，時系列を考慮してドラム譜を推定できる [5–7]．

記号単位のドラム譜を推定するためには，フレーム単位の入力特徴量からテイタム単位のオンセット確率値を推定する機構が必要になる．ある入力系列から異なる長さの系列を出力する機構に，エンコーダ・デコーダモデルがある．このモデルは，エンコーダが音響的な特徴量を抽出し，デコーダが言語的に妥当な推論を行う機構とみなすことができる [8] だけでなく，実装が容易で推論時間も短いという利点がある．そこで，本提案手法では，フレーム単位の入力特徴量からテイタム単位のドラム譜を推定する機構 (採譜モデル) として，エンコーダデコーダモデルを採用する．エンコーダが局所的な特徴量に着目して潜在表現ベクトルを抽出し，デコーダがテイタム単位で楽曲の構造を捉える機構として働くため，採譜モデルはエンコーダ・デコーダモデルとして音響的な側面と音楽的な側面の両者を同時に学習することが期待できる．なお，ビート時刻は予め推定しておき，max-pooling を用いてフレーム・テイタム間の対応づけを行うものとする．

本提案法におけるエンコーダには，特徴量抽出器として CNN を用いる (図 1「採譜モデル」上部)．デコーダには，時系列を考慮してドラム譜を推定するため，RNN と自己

[1] 京都大学大学院情報学研究科
[a)] ishizuka@sap.ist.i.kyoto-u.ac.jp
[b)] nishikimi@sap.ist.i.kyoto-u.ac.jp
[c)] enakamura@sap.ist.i.kyoto-u.ac.jp
[d)] yoshii@kuis.kyoto-u.ac.jp

注意機構を用いることが考えられる．機械翻訳や自動音声認識の分野でよく利用される自己注意機構は，RNN よりも広範な構造を学習することができるうえに，並列計算による高速化を可能にするため，本提案法ではデコーダとして自己注意機構に基づく Transformer エンコーダを採用する (図 1「採譜モデル」下部)．具体的には，学習データとして 16 小節の系列を扱い，注意重みを計算することで長期的な依存関係を明示的に学習することが可能になる．

DNN ベースの自動ドラム採譜システムでは，準備することができるペアデータの量が限られているために，繰り返し構造に代表されるような楽曲の長期的な特徴を学習することが難しく，音楽的に不自然なパターンをしばしば生成してしまうという問題点が指摘されている．この問題を解決する手法の一つに，転移学習 [9,10] の利用が挙げられる．大規模ドラム譜データからテイタム単位でオンセットの確率分布を学習した言語モデルの知識を，DNN ベースの採譜モデルに転移することで，音楽的な妥当性を担保しながらドラム譜を推定することができる．従来は単方向の言語モデルが設計されてきたが，Masked language model (MLM) を用いることで双方向の文脈を利用してドラム譜の音楽的な妥当性を評価することができる．

このような背景から，本提案法において採譜モデルの学習過程で MLM による正則化を行う (図 1 赤線「正則化」部)．採譜モデルの正則化付き学習では，モデルが出力するテイタム単位オンセット確率値と正解ラベル間のクロスエントロピー $\mathcal{L}_{\text{tran}}$ と，MLM により計算されたオンセット確率値の損失 $\mathcal{L}_{\text{lang}}$ との重み付け和 $\mathcal{L}_{\text{total}}$ を目的関数として最小化する．このとき，オンセット確率値は gumbel-sigmoid trick [11] により微分可能な形で二値化されるため，誤差逆伝播に基づくネットワークの最適化が可能になる．なお，これらのフレームワークは，我々が最近提案した手法 [12] の拡張になっている．

2. 関連研究

本章では，自動ドラム採譜 (2.1 節) の関連研究を述べた後に，エンコーダ・デコーダモデル (2.2 節)， 言語モデル (2.4 節)，転移学習 (2.3 節)，自己注意機構 (2.5 節) に関する関連研究を述べる．

2.1 自動ドラム採譜

NMF は，振幅スペクトログラムを基底行列とアクティベーション行列に分解する手法として自動ドラム採譜に応用されてきた [13–17]．しかし，基底行列の表現力の乏しさから，最近では局所的な特徴量を自動で抽出する CNN が利用されるようになり，自動ドラム採譜に限らず自動音楽採譜で広く活用されている [2–4, 18, 19]．フレーム単位の時系列依存性を学習するために，RNN もよく利用されている [5,7]．このように，自動ドラム採譜ではフレーム単位で設計された DNN ベースの手法がよく用いられているが，記号単位で推定を行う手法はまだ少ない．

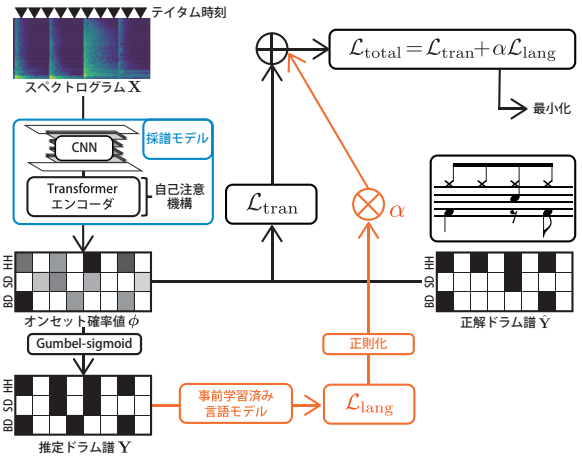

図 1　本提案法の概略図．

2.2 エンコーダ・デコーダモデル

機械翻訳や文章要約，画像認識などの分野において，ある系列を異なる系列に変換する手法が盛んに提案されている．エンコーダ・デコーダモデルは，長さの異なる系列を扱う最も基本的なモデルである．このモデルは，エンコーダが音響的な特徴量を抽出し，デコーダが言語的に妥当な推論を行う機構とみなすことができる [8]．エンコーダ・デコーダモデルは，実装が容易で推論時間も短い一方で，学習に十分なペアデータが必要となる．そこで，大量のテキストデータによって学習されたモデルの知識を転移学習に基づいて活用しようとする手法が提案されてきた [9, 10]．

2.3 転移学習

転移学習は関連するドメインから効率的に知識を利用することを目的とし，多くの分野で幅広く応用されている [20]．転移学習の扱う対象はペアデータだけでなく，ラベルなしデータに対しても適用することができ，対象とする問題に応じて様々なデータセットの組み合わせが利用される [21–23]．いくつかの研究では，teacher-student の枠組みで student モデルが teacher モデルと同等もしくは上回る性能を発揮することが報告されている [24, 25]．転移学習が比較的軽量な student モデルの学習に利用される場合は，特に知識蒸留と呼ばれる [21]．

最近，大規模ラベルなしデータから学習された他ドメインの知識を利用する研究が盛んに行われている．ASR では，言語として自然な単語列を推定するために，デコード時に言語モデルが利用されている．しかし，この手法では推論に長い時間を要する．このような背景から，ASR では知識蒸留に基づく手法が提案されている．Cold fusion [26] と呼ばれる手法では，End-to-end 型モデルの学習時に事前学習済み言語モデルの隠れ状態を組み込むことで，転移学習を行なっている．知識蒸留に基づく手法は推論時に言語モデルを必要としないため，高速に動作する利点が

ある [9]．知識蒸留は自動ドラム採譜でも利用されており，Wu ら [27] は DNN ベースの student モデルに対して NMF ベースの teacher モデルの知識を埋め込むことで，ラベルなしデータの活用方法を示した．石塚ら [12] は，CRNN ベースの採譜モデルに対して GRU ベースの事前学習済み言語モデルによる正則化を施すことで，音楽的な妥当性を担保しながらドラム採譜を行う方法を提案した．

2.4 言語モデル

DNN ベースの自動ドラム採譜システムでは，準備することができるペアデータの量が限られているために，音楽的に不自然な出力をしばしば推定してしまうという問題が残されている．この問題を解決する手法の一つとして，音楽的な妥当性を評価する言語モデルの利用が挙げられる．従来，言語モデルはフレーム単位で設計されてきた．例えば，Raczyński ら [28] は，deep belief network で設計された言語モデルを用いてコードの遷移を確率的にモデル化することで，NMF で設計された採譜モデルの精度を向上させた．Sigtia ら [29] は，DNN によって推定された事後確率から音楽的に妥当なコード系列を推定するために，RNN で設計された言語モデルを用いた．しかし， [30,31] で指摘されているように，音楽的な構造を学習するためには，言語モデルをテイタム単位で設計する必要がある．

最近になって，自動音楽採譜の分野ではテイタム単位で設計された言語モデルの利用が検討されている．Korzeniowski [32] らは，N-gram を用いてシンボル単位の言語モデルを設計し，DNN ベースのコード認識モデルの精度を向上させた．Korzeniowski [33] らは, フレーム単位で設計された言語モデルがデコード時にスムージングの効果しかもたらしていないという問題意識の下，コードの継続時間をモデル化することで RNN で設計されたシンボル単位の言語モデルをコード認識に活用した．Ycart ら [34] は，LSTM (long short-term memory) の言語モデルとしての能力を幅広い観点から調査し，16 分音符単位で設計された LSTM 言語モデルが音符の遷移などを表現することができることを示した．自動ドラム採譜では，Thompson ら [35] がサポートベクターマシンを用いてドラムパターンをいくつかに分類した辞書を作り，そのテンプレートとのマッチングに基づく言語モデルを提案した．上田ら [15] は，DNN ベースの言語モデルをドラム譜の事前分布として利用することで，ベイズ推論を行う手法を提案した．しかし，いずれの手法も採譜モデルが楽曲の長期的な構造を学習しにくいという問題点が残されている．

2.5 自己注意機構

注意機構に基づくエンコーダ・デコーダモデルは，デコード時にエンコーダの最終出力である固定長の埋め込みを用いる代わりに全ての埋め込みの重み付け和を用いることで，より長期的な系列を扱うことを可能にした [36,37]．注意機構はクエリ，キー，バリューという 3 つのベクトルを用いて記述することができる．クエリは入力系列，キーとバリューは出力系列に対応するベクトルである．注意機構に基づくエンコーダ・デコーダモデルでは，クエリベクトルとキーベクトルを用いてスコアが計算され，スコアに基づいて算出された注意重みに従ってバリューベクトルから各要素が抽出される．計算の容易さから，スコアの計算には内積が用いられることが多い．入力系列と出力系列が同じ系列を指す場合，注意機構は特に自己注意機構と呼ばれる．Transformer は (自己) 注意機構のみで構成されたエンコーダ・デコーダモデルであり，スコアは内積によって計算される [38]．従来の RNN を用いたエンコーダ・デコーダモデルとは異なり，学習時には非再帰的に並列計算を行う．Transformer を用いることで，長期的な時系列依存性を学習するために深いネットワークを構築する必要がなくなり，並列計算による高速化が可能になった．

最近，言語モデルの設計として，事前学習を用いる手法が盛んに提案されている．ELMO (embeddings from language models) [39] は特徴量に基づく事前学習モデルの一つであり，順方向 RNN と逆方向 RNN を最終層で結合することで双方向の推論を可能にしているが，順方向の推論と逆方向の推論が分離されている．さらに，学習が再帰的に行われるため，計算に時間を要する．GPT (generative pretrained Transformer) [40] と BERT (bidirectional encoder representations from Transformers) [41] はファインチューニングに基づく事前学習モデルの一つである．GPT は Transformer の注意機構にマスクをかけて未来の情報に注意がかからないようにすることで自己注意機構を用いた言語モデルを実現したが，その推論は単方向に限られている．BERT の事前学習は MLM と Next sentence prediction のマルチタスク学習によって構成されており，双方向の文脈を同時に学習して推定単語列のスコア (擬似パープレキシティ) を計算することができる [42]．

自己注意機構は，音楽生成の分野でもよく利用されている．Music Transformer [43] は，楽曲の長期的な構造を学習するために相対位置に基づく自己注意機構を導入し，計算量を削減する新しいアルゴリズムを提案した．Pop Music Transformer [44] は Transformer-XL を利用し，音楽のリズムや調和を表現する新しいデータ構造を提案した．Transformer Variational AutoEncoder [45] は，階層的な変分オートエンコーダに自己注意機構を導入することで，局所的・長期的な楽曲構造を同時に学習することを可能にした．Harmony Transformer [46] は，コードの認識と長期的な遷移を自己注意機構を用いて同時に学習することで，コード認識の性能を向上させた．しかし，自動ドラム採譜の文脈では，ドラムが繰り返し構造を表しやすい楽器でありながらも，自己注意機構を用いる研究はまだ少ない．

3. 提案手法

本章では，音楽音響信号から得られたメル周波数スペクトログラムを入力として，テイタム単位のドラム譜を推定する提案法について，問題設定を述べる (3.1 節)．図 1 に示す通り，我々の提案法はテイタム単位のオンセット確率値を推定する DNN ベースの採譜モデル (3.2 節) を利用し，ドラム譜の音楽的な妥当性を評価する言語モデル (3.3 節) を大規模ドラム譜により事前に学習しておくことで，採譜モデルの学習時に正則化を施す (3.4 節)．

3.1 問題設定

我々は，メル周波数スペクトログラム $\mathbf{X} \in \mathbb{R}_{+}^{F \times T}$ からドラム譜 $\mathbf{Y} \in \{0,1\}^{M \times N}$ を推定する問題に取り組む．ここで，M は扱うドラムの楽器数 (本稿では BD，SD，HH であるため $M = 3$)，N はテイタム数（本稿では 16 小節の系列を扱うため $N = 256$)，F は周波数ビン数，T は時間フレーム数を表す．本稿では，全てのオンセット時刻がビート時刻を四等分した 16 分音符単位のテイタム時刻 $\mathbf{B} = \{b_n\}_{n=1}^{N}$ に沿っているものと仮定し，ビート時刻は予め推定された結果を利用する．ただし，$1 \leq b_n \leq T$ かつ $b_n < b_{n+1}$ とする．

3.2 採譜モデル

採譜モデルは，テイタム単位のオンセット確率値 $\boldsymbol{\phi} \in [0,1]^{M \times N}$ を推定する．ここで，$\boldsymbol{\phi} \in [0,1]^{M \times N}$ はドラム m のテイタム n におけるオンセット確率値を表す．$\boldsymbol{\phi} \in [0,1]^{M \times N}$ を閾値 $\delta \in [0,1]$ によって二値化することで，ドラム譜 $\mathbf{Y} \in \{0,1\}^{M \times N}$ を推定する．採譜モデルには，フレーム単位の CNN，フレーム・テイタム間のアライメントを行う max-pooling，自己注意機構に基づくテイタム単位の Transformer エンコーダを用いる．最初に，CNN はメル周波数スペクトログラム $\mathbf{X}$ を入力として受け取り，フレーム単位の潜在ベクトル表現系列 $\mathbf{F} \in \mathbb{R}^{D_F \times T}$ に変換し，テイタム時刻 $\mathbf{B}$ に基づく max-pooling を利用してテイタム単位の潜在ベクトル表現系列 $\mathbf{G} \in \mathbb{R}^{D_F \times N}$ に変換する (図 2)．ここで，D_F は潜在ベクトル表現の次元数とする．具体的には，以下の操作を行う．

$$G_{d,n} = \max_{\frac{b_{n-1}+b_n}{2} \leq t < \frac{b_n+b_{n+1}}{2}} F_{d,t} \tag{1}$$

ここで，$b_0 = b_1$ かつ $b_{N+1} = b_N$ である．Transformer エンコーダは，$\mathbf{G}$ に Positional encoding を加えた $\mathbf{G}'$ を入力として受け取り，オンセット確率値 $\boldsymbol{\phi}$ を推定する (図 3)．ここでは， [38] で述べられている Positional encoding を採用し，Multi-head 注意機構 (図 3 青枠部) の head の数は $h \in \mathbb{N}$ とする．Transformer エンコーダにおいて，クエリベクトル $\mathbf{Q}_i \in \mathbb{R}^{D_K \times N}$，キーベクトル $\mathbf{K}_i \in \mathbb{R}^{D_K \times N}$，バリューベクトル $\mathbf{V}_i \in \mathbb{R}^{D_K \times N}$ は以下のように計算される．

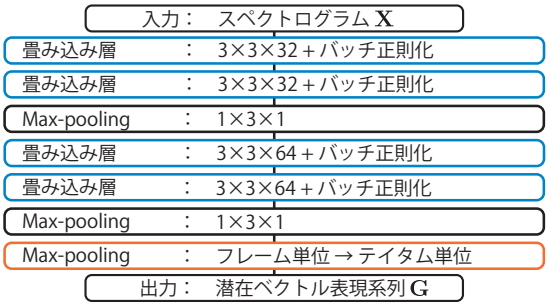

図 2 CNN の概略図.

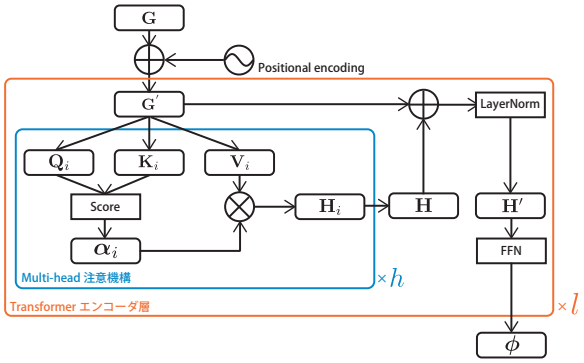

図 3 自己注意機構の概略図.

$$\mathbf{Q}_i = [\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,N}] = \mathbf{W}_i^{(Q)}\mathbf{G}' + \mathbf{b}_i^{(Q)} \tag{2}$$

$$\mathbf{K}_i = [\mathbf{k}_{i,1}, \ldots, \mathbf{k}_{i,N}] = \mathbf{W}_i^{(K)}\mathbf{G}' + \mathbf{b}_i^{(K)} \tag{3}$$

$$\mathbf{V}_i = [\mathbf{v}_{i,1}, \ldots, \mathbf{v}_{i,N}] = \mathbf{W}_i^{(V)}\mathbf{G}' + \mathbf{b}_i^{(V)} \tag{4}$$

ただし，$1 \leq i \leq h$ は Multi-head 注意機構における head のインデックス，D_k は潜在ベクトル表現の次元数，$\mathbf{q}_i \in \mathbb{R}^{D_K}$，$\mathbf{k}_i \in \mathbb{R}^{D_K}$，$\mathbf{v}_i \in \mathbb{R}^{D_K}$ はそれぞれクエリ，キー，バリューベクトル，$\mathbf{W}_i^{(Q)} \in \mathbb{R}^{D_K \times D_F}$，$\mathbf{W}_i^{(K)} \in \mathbb{R}^{D_K \times D_F}$，$\mathbf{W}_i^{(V)} \in \mathbb{R}^{D_K \times D_F}$ はそれぞれ重み行列，$\mathbf{b}_i^{(Q)} \in \mathbb{R}^{D_K \times N}$，$\mathbf{b}_i^{(K)} \in \mathbb{R}^{D_K \times N}$，$\mathbf{b}_i^{(V)} \in \mathbb{R}^{D_K \times N}$ はそれぞれバイアスベクトルである．なお，本稿では [38] に従って $D_k = \frac{D_F}{h}$ に設定した． 注意重み $\boldsymbol{\alpha}_i \in \mathbb{R}^{N \times N}$ は，$\mathbf{G}'$ 自身の関連度を表す正規化された行列であり，以下のように計算される．

$$e_{i,n,n'} = \mathrm{Score}\left(\mathbf{q}_{i,n}, \mathbf{k}_{i,n'}\right) \tag{5}$$

$$\alpha_{i,n,n'} = \frac{\exp\left(e_{i,n,n'}\right)}{\sum_{n'=1}^{N} \exp\left(e_{i,n,n'}\right)} \tag{6}$$

ここで，n と n' はそれぞれ $\mathbf{Q}_i$ と $\mathbf{K}_i$ のテイタム方向に関するインデックスを示す．本稿では，Score の計算に内積を用いる．

$$\mathrm{Score}\left(\mathbf{q}_{i,n}, \mathbf{k}_{i,n'}\right) = \frac{\mathbf{q}_{i,n}^T \mathbf{k}_{i,n'}}{\sqrt{D_K}} \tag{7}$$

ただし，T はベクトルの転置を表す．バリューベクトルと注意重みの内積を潜在表現ベクトルの次元方向に結合することで，潜在ベクトル表現系列 $\mathbf{H} \in \mathbb{R}^{D_F \times N}$ を得る．

$$\mathbf{H}_i = \mathbf{V}_i \boldsymbol{\alpha}_i^T \tag{8}$$

$$\mathbf{H} = [\mathbf{H}_1, \ldots, \mathbf{H}_h] \tag{9}$$

$\mathbf{H}$ に残差接続として入力 $\mathbf{G}'$ を加えた後に，Layer normarization [47] を施して $\mathbf{H}' \in \mathbb{R}^{D_F \times N}$ を得る．

$$\mathbf{H}' = \mathrm{LayerNorm}(\mathbf{H} + \mathbf{G}') \tag{10}$$

$\mathbf{H}'$ は，ReLU (rectified linear unit) を通して 2 層のフィードフォワードネットワークに通される．

$$\mathrm{FFN}_j(\mathbf{H}') = \mathbf{W}_2^{(H')} \max\left(0, \mathbf{W}_1^{(H')}\mathbf{H}' + \mathbf{b}_1^{(H')}\right) + \mathbf{b}_2^{(H')} \tag{11}$$

ただし，j は Transformer エンコーダの層に関するインデックス，$\mathbf{W}_1^{(H')} \in \mathbb{R}^{D_{\mathrm{FFN}} \times D_F}$，$\mathbf{W}_2^{(H')} \in \mathbb{R}^{D_F \times N_{\mathrm{FFN}}}$ は重み行列，$\mathbf{b}_1^{(H')} \in \mathbb{R}^{D_{\mathrm{FFN}} \times D_F}$，$\mathbf{b}_2^{(H')} \in \mathbb{R}^{D_F \times N}$ はバイアスベクトル，D_{FFN} は線形変換後の潜在表現ベクトルの次元数である．式 (2)〜式 (11) で表される処理を一つの層として設定し，残差接続として初期の入力 $\mathbf{G}'$ の代わりに前の層の出力 $\mathrm{FFN}_j(\mathbf{H}')$ を加えて l 回繰り返す (図 3 赤線部)．最終的に，オンセット確率値 $\boldsymbol{\phi}$ は以下のように計算される．

$$\boldsymbol{\phi} = \sigma\left(\mathbf{W}_3^{(H')}\mathrm{FFN}_l(\mathbf{H}') + \mathbf{b}_3^{(H')}\right) \tag{12}$$

ただし，$\sigma(\cdot)$ はシグモイド関数，$\mathbf{W}_3^{(H')} \in \mathbb{R}^{M \times D_F}$ は重み行列，$\mathbf{b}_3^{(H')} \in \mathbb{R}^{M \times N}$ はバイアスベクトルである．

3.3 言語モデル

言語モデルは，ドラム譜 $\tilde{\mathbf{Y}} \in \mathbb{R}^{M \times N}$ の生成確率を評価する．本稿では，双方向言語モデルを提案する．

3.3.1 双方向言語モデル

楽曲の長期的な構造を学習する双方向言語モデルとして，採譜モデル同様 Transformer エンコーダを用いる． 事前学習タスクとして，BERT で利用されている MLM を採用する (図 4 上部)．MLM では，学習時に全テイタム数の 15%をランダムに選び，[Mask] トークンを挿入する．学習時は，以下の $\mathcal{L}_{\mathrm{lang}}^{\mathrm{bi}}(\tilde{\mathbf{Y}})$ を最小化する．

$$\hat{p}(\tilde{\mathbf{Y}}_n) = p(\tilde{Y}_{:,n} | \tilde{Y}_{:,1:n-1}, \tilde{Y}_{:,n+1:N}) \tag{13}$$

$$\mathcal{L}_{\mathrm{lang}}^{\mathrm{bi}}(\tilde{\mathbf{Y}}) = -\sum_{n=1}^{N} \log \hat{p}(\tilde{\mathbf{Y}}_n) \tag{14}$$

3.3.2 単方向言語モデル

単方向言語モデルの学習時には，負の対数尤度 $\mathcal{L}_{\mathrm{lang}}^{\mathrm{uni}}(\tilde{\mathbf{Y}}) = -\log p(\tilde{\mathbf{Y}})$ を最小化する．

$$\mathcal{L}_{\mathrm{lang}}^{\mathrm{uni}}(\tilde{\mathbf{Y}}) = -\sum_{n=1}^{N} \log p(\tilde{Y}_{:,n} | \tilde{Y}_{:,1:n-1}) \tag{15}$$

ここで，":" は全ての要素を表す．単方向言語モデルには，[12] で提案されているスキップ型 Bi-gram (図 4 中部) と GRU を用いる (図 4 下部)．単方向言語モデルの損失 $\mathcal{L}_{\mathrm{lang}}^{\mathrm{uni}}(\mathbf{Y})$ は負の対数尤度であるのに対し，双方向言語モデルの損失 $\mathcal{L}_{\mathrm{lang}}^{\mathrm{bi}}(\mathbf{Y})$ は負の対数尤度と厳密には異なることに注意されたい．

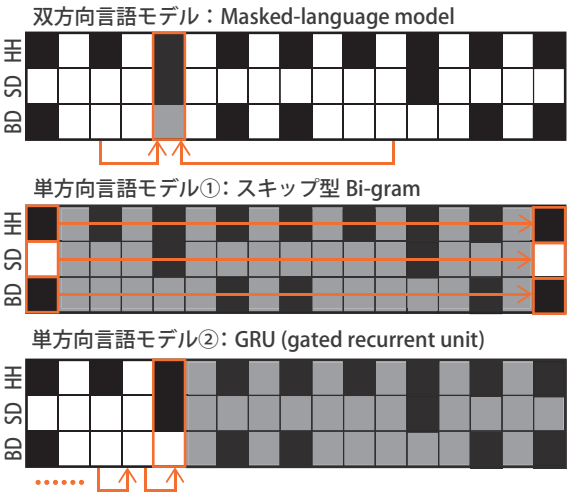

図 4　言語モデルの概略図．

3.4 正則化

採譜モデルの学習時は，正解ラベル $\hat{\mathbf{Y}}$ が与えられた下で，$\hat{\mathbf{Y}}$ の負の対数尤度を最小化する．

$$\begin{aligned}&\mathcal{L}_{\mathrm{tran}}(\boldsymbol{\phi}|\hat{\mathbf{Y}}) \\ &= -\sum_{m=1}^{M}\sum_{n=1}^{N}\left(\gamma \hat{Y}_{m,n} \log \phi_{m,n} + (1-\hat{Y}_{m,n}) \log(1-\phi_{m,n})\right)\end{aligned} \tag{16}$$

ただし，$\gamma > 0$ はオンセットとオフセットの不均衡を調節するための重みである．$\mathcal{L}_{\mathrm{tran}}$ は $\boldsymbol{\phi}$ と $\hat{\mathbf{Y}}$ のクロスエントロピー，つまり音響的な損失しか考慮していないために，$\boldsymbol{\phi}$ を二値化して得られる $\mathbf{Y}$ の音楽的な妥当性は担保されていない．そこで，本提案法では以下で表される損失 $\mathcal{L}_{\mathrm{total}}$ を最小化することで誤差逆伝播に基づいて採譜モデルの最適化を行う．

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{tran}}(\boldsymbol{\phi}|\hat{\mathbf{Y}}) + \alpha \mathcal{L}_{\mathrm{lang}}^{*}(\mathbf{Y}) \tag{17}$$

ここで，$\alpha > 0$ は事前学習済み言語モデルによる正則化項の重み，*は「uni」または「bi」を表す．学習時には，$\boldsymbol{\phi}$ を閾値によって二値化する代わりに，誤差逆伝播を行うために微分可能な形で二値化する．そこで， [12] と同様に gumbel-sigmoid trick [11] と呼ばれる手法を用いる．事前学習済み言語モデルのパラメータは，採譜モデルの学習時は固定しておく．

4. 評価実験

本章では，実験設定 (4.1 節) を述べた後に，言語モデルの評価 (4.2 節) と提案法の有効性 (4.3 節) を報告する．

4.1 実験設定

言語モデルの学習には，J ポップとビートルズ計 510 曲のドラム譜を用いた．ハイパーパラメータは Optuna [48] を用いて 3 分割交差検証に基づくベイズ最適化を行なった．その結果，GRU の層数は 3，潜在ベクトル表現の次元

数は 98，Transformer エンコーダのパラメータは $h = 4$，$D_F = 112$，$D_{\mathrm{FFN}} = 448$，$l = 8$ となった．採譜モデルの評価には RWC ポピュラー音楽データベース [49] (RWC) でドラムが含まれる 89 曲のうち，アノテーションが正確な計 69 曲を用い，10 分割交差検証を行なった．いずれの交差分割検証においても，学習データのうちランダムに選んだ 15% を検証データとして利用した．いずれのデータセットにおいても，学習データには正解ビートを用いて 16 小節ごとに分割したデータを利用し，バッチサイズは 10 に設定した．RWC の楽曲に対して，シフト幅 441 (10 ms)，窓幅 2048 (約 46 ms) の短時間フーリエ変換を用いて振幅スペクトログラムを作成した．さらに，バンド数 80，最低周波数 20Hz，最高周波数 20000Hz のメルフィルタバンクを利用して，入力特徴量のメル周波数スペクトログラムを作成した．学習データには Spleeter [50] による分離音源も用い，混合音から得られたメル周波数スペクトログラムと合わせて入力特徴量を 2 チャネルに設定した．ビート推定には Madmom [51] を利用し，評価尺度には，以下で定義される F 値を利用した．

$$\mathcal{P} = \frac{N_c}{N_e} \quad \mathcal{R} = \frac{N_c}{N_g} \quad \mathcal{F} = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P} + \mathcal{R}} \tag{18}$$

ここで，N_e は推定されたビート数，N_g は正解のビート数，N_c は正解したビート数を表している．推定されたビート時刻には 50ms の許容誤差を設定し，mir_eval [52] を利用して $\mathcal{P}, \mathcal{R}, \mathcal{F}$ を計算した．Madmom によるビート推定の精度は 96.4%であり，RWC に対しては高い精度でビートを推定できていることが分かった．

4.2 言語モデルの評価

言語モデルの予測性能を評価するため，RWC に対してパープレキシティを計算した．事前学習済み双方向言語モデルの尤度は以下のように定義される [42].

$$p(\tilde{\mathbf{Y}}) = \frac{1}{Z} \prod_{n=1}^{N} \hat{p}(\tilde{Y}_n) \tag{19}$$

$$Z = \sum_{\tilde{\mathbf{Y}} \in \Theta} \hat{p}(\tilde{Y}_n) \tag{20}$$

ただし，Θ はドラムパターンの全ての組み合わせの総数を表す．しかし，計算量の問題から Z を算出することが難しい．そこで，単方向言語モデル同様に，双方向言語モデルでも以下のように定義されるパープレキシティを利用する．

$$\mathrm{PPL}^{*}(\tilde{\mathbf{Y}}) = 2^{\frac{1}{N}\mathcal{L}^{*}_{\mathrm{lang}}(\tilde{\mathbf{Y}})} \tag{21}$$

ただし，「*」は uni もしくは bi を表す．$\mathrm{PPL}^{\mathrm{bi}}$ はパープレキシティの定義と厳密には異なるため，$\mathrm{PPL}^{\mathrm{uni}}$ と単純な比較ができないことに注意されたい．スキップ型 Bi-gram と GRU を用いて $\mathrm{PPL}^{\mathrm{uni}}$ を計算したところ，スキップ型 Bi-gram は 1.326，GRU は 1.266 となった．この結果から，スキップ型 Bi-gram に比べて GRU の方が，評価データに対する言語モデルとして適切に機能していることが分かる．また，MLM を用いて $\mathrm{PPL}^{\mathrm{bi}}$ を計算したところ，1.084 となった．この結果から，正則化項として言語モデルによる評価値を利用する際には，同程度のオーダの重みパラメータが利用されることが予測される．

4.3 実験結果

CNN の構造は [53] に基づいた．畳み込み層のカーネルサイズは 3×11 であり，時間方向の次元数が一定になるようにゼロパディングを施した．ネットワークの最適化には AdamW [54] (lr=10^{-3}, $\beta_1 = 0.9$，$\beta_2 = 0.98$，$\epsilon = 10^{-9}$) を利用し，検証データの損失 $\mathcal{L}_{\mathrm{total}}$ が 30 回連続で下がらなかった場合に学習を停止した．Transformer の学習率は，[38] に従って変化させた．なお，*warmup_steps* は 4000 に設定した．全ての学習段階で過学習を防ぐために重み正則化 (λ=10^{-4}) を施し，ドロップアウトを CRNN の全結合層の直前 ($p = 0.2$) と Transformer の各レイヤー ($p = 0.1$) に適用した．CNN と GRU の重みは [55]，全結合層の重みは Uniform$(0, 1)$, バイアスは 0 にそれぞれ初期化した．比較のため，フレーム単位で設計する CRNN として最高精度を達成したとされている従来法 [7] の精度も評価した．このモデルの学習方法は，[12] に従った．ハイパーパラメータは，言語モデル同様検証データに対して Optuna [48] を用いてベイズ最適化し，$h = 8$，$D_K = 120$，$D_{\mathrm{FFN}} = 480$，$l = 7$ に設定した．α，β，γ の最適化の結果は表 1 に示す．

提案法の評価は，ビート推定と同様に式 (18) に従い，許容誤差も 50 ms に設定した．提案法の採譜精度を表 1 に示す．採譜モデルに CRNN，Transformer エンコーダを使った場合のいずれも，言語モデルによる正則化を施すことで採譜精度が向上し，MLM の正則化に基づく採譜モデルが最も全体の精度向上幅が大きかった．BD と SD の精度は採譜モデルに Transformer エンコーダ，言語モデルに MLM を用いた場合が最も高く，HH の精度は採譜モデルに CRNN，言語モデルに GRU を用いた場合が最も高く，全体の精度は採譜モデルに CRNN，言語モデルに MLM を用いた場合が最も高かった．この結果は，繰り返し構造を学習するために，BD や SD が長期的な履歴を必要とするのに対し，HH が短期的な履歴で十分であることを示唆している．最後に，正則化によって結果が改善した RWC の楽曲 (RWC-MDB-P-2001 No.40) の一部を図 5 に示す．採譜モデルには Transformer エンコーダ，言語モデルには MLM ($\alpha = 1.312$) を利用し，図中の F 値は楽曲全体の採譜精度を表す. 左図を見ると，MLM による正則化に基づいて，各小節の 4 拍目に SD が演奏されやすいという規則性を採譜モデルが学習していることが確認できた．一方で，右図のように双方向の長期的な履歴に引きずられて HH を単純化しすぎてしまう例も見られた．

表 1　従来法と提案法のドラム採譜精度 (F 値) .

手法	言語モデル	採譜モデル		Madmom BD	Madmom SD	Madmom HH	Madmom Total	Ground-truth BD	Ground-truth SD	Ground-truth HH	Ground-truth Total
従来法	-	CRNN [7]	($\beta = 8.270$)	83.8%	72.1%	70.8%	75.5%	83.7%	72.2%	70.8%	75.6%
提案法	-	CRNN	($\gamma = 0.610$)	87.5%	74.3%	79.8%	80.5%	87.9%	74.0%	79.0%	80.3%
	単方向	+ Bi-gram	($\alpha = 1.370$)	87.5%	76.8%	81.2%	81.9%	87.6%	76.4%	80.6%	81.6%
		+ GRU	($\alpha = 0.036$)	87.4%	77.6%	**81.4%**	82.1%	87.3%	76.8%	**80.7%**	81.6%
	双方向	+ MLM	($\alpha = 0.264$)	88.2%	78.2%	80.2%	**82.2%**	88.5%	78.4%	79.7%	**82.2%**
	-	Transformer	($\gamma = 0.620$)	87.9%	77.1%	71.9%	78.9%	88.2%	76.7%	72.6%	79.2%
	単方向	+ Bi-gram	($\alpha = 1.103$)	87.8%	76.3%	73.5%	79.2%	88.1%	76.8%	74.2%	79.7%
		+ GRU	($\alpha = 0.045$)	88.1%	77.4%	73.0%	79.5%	88.3%	78.2%	74.3%	80.3%
	双方向	+ MLM	($\alpha = 1.312$)	**88.7%**	**79.2%**	73.4%	80.4%	**89.0%**	**79.6%**	74.7%	81.1%

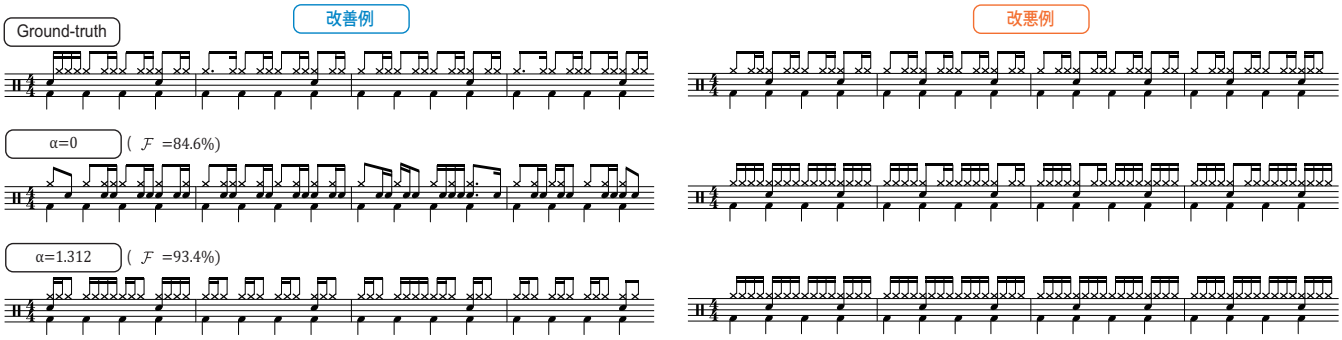

図 5　採譜モデルに正則化を施すことで推定結果が改善 (左図)，改悪 (右図) した具体例.

5. おわりに

本稿では，テイタム単位のドラム採譜において自己注意機構を導入すると同時に，双方向言語モデルにより長期的なドラム譜を学習する手法を提案した．採譜モデルは，フレーム単位の特徴量抽出器である CNN に基づくエンコーダとテイタム単位で音符の時系列依存性を表現する Transformer エンコーダに基づくデコーダで構成され，音響・音楽的な側面を同時に学習する．評価実験により，MLM による正則化は，採譜モデルの精度と推定結果の音楽的な自然さを向上させることが明らかになった．

今後は，自動ドラム採譜の目的が記号単位の楽譜を出力することから，従来のようなフレーム単位の評価ではなく記号単位の評価尺度を導入する予定である．他にも，ドラムのオンセット時刻とビート時刻に相互依存性があるとすれば，マルチタスク学習の枠組みでビート・ダウンビート推定を提案法に組み入れることが考えられる．

謝辞　本研究の一部は，JST ACCEL No.JPMJAC1602，JSPS 科研費 No.16H01744，19K20340 および 19H04137 の支援を受けた．

参考文献

[1] Wu, C.-W., Dittmar, C., Southall, C., Vogl, R., Widmer, G., Hockman, J., Müller, M. and Lerch, A.: A Review of Automatic Drum Transcription, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
[2] Jacques, C. and Roebel, A.: Automatic Drum Transcription with Convolutional Neural Networks, *DAFx*.
[3] Gajhede, N., Beck, O. and Purwins, H.: Convolutional Neural Networks with Batch Normalization for Classifying Hi-hat, Snare, and Bass Percussion Sound Samples, *Proceedings of the Audio Mostly 2016*.
[4] Southall, C., Stables, R. and Hockman, J.: Automatic Drum Transcription for Polyphonic Recordings Using Soft Attention Mechanisms and Convolutional Neural Networks, *ISMIR*.
[5] Vogl, R., Dorfer, M. and Knees, P.: Recurrent Neural Networks for Drum Transcription, *ISMIR*.
[6] Stables, R., Hockman, J. and Southall, C.: Automatic Drum Transcription using Bi-directional Recurrent Neural Networks, *ISMIR*.
[7] Vogl, R., Dorfer, M., Widmer, G. and Knees, P.: Drum Transcription via Joint Beat and Drum Modeling Using Convolutional Recurrent Neural Networks, *ISMIR*.
[8] Chorowski, J., Bahdanau, D., Cho, K. and Bengio, Y.: End-To-End Continuous Speech Recognition Using Attention-Based Recurrent NN: First Results, *NIPS Workshop on Deep Learning* (2014).
[9] Kubin, G. and Kacic, Z.: Learn Spelling from Teachers: Transferring Knowledge from Language Models to Sequence-to-Sequence Speech Recognition, *Interspeech*.
[10] Chen, Y.-C., Gan, Z., Cheng, Y., Liu, J. and Liu, J.: Distilling the Knowledge of BERT for Text Generation, *arXiv* (2019).
[11] Tsai, Y.-H., Liu, M.-Y., Sun, D., Yang, M.-H. and Kautz, J.: Learning Binary Residual Representations for Domain-Specific Video Streaming, *AAAI*.
[12] Ishizuka, R., Nishikimi, R., Nakamura, E. and Yoshii, K.: Tatum-Level Drum Transcription Based on a Convolutional Recurrent Neural Network with Language Model-Based Regularized Training, *arXiv* (2020).
[13] Wu, C.-W. and Lerch, A.: Drum Transcription Using Partially Fixed Non-Negative Matrix Factorization with

Template Adaptation, *ISMIR*.
[14] Roebel, A., Pons, J., Liuni, M. and Lagrangey, M.: On Automatic Drum Transcription Using Non-Negative Matrix Deconvolution and Itakura Saito Divergence, *ICASSP*.
[15] Ueda, S., Shibata, K., Wada, Y., Nishikimi, R., Nakamura, E. and Yoshii, K.: Bayesian Drum Transcription Based on Nonnegative Matrix Factor Decomposition with a Deep Score Prior, *ICASSP*.
[16] Dittmar, C. and Gärtner, D.: Real-Time Transcription and Separation of Drum Recordings Based on NMF Decomposition, *DAFx*.
[17] Paulus, J. and Klapuri, A.: Drum Sound Detection in Polyphonic Music with Hidden Markov Models, *EURASIP Journal on Audio, Speech, and Music Processing*.
[18] Schlüter, J. and Böck, S.: Improved Musical Onset Detection with Convolutional Neural Networks, *ICASSP*.
[19] Wang, Q., Zhou, R. and Yan, Y.: A Two-Stage Approach to Note-Level Transcription of a Specific Piano, *Applied Sciences*.
[20] Weiss, K., Khoshgoftaar, T. M. and Wang, D.: A Survey Of Transfer Learning, *Journal of Big data*.
[21] Hinton, G., Vinyals, O. and Dean, J.: Distilling the Knowledge in a Neural Network, *arXiv* (2015).
[22] Zagoruyko, S. and Komodakis, N.: Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks Via Attention Transfer, *ICLR* (2017).
[23] Yim, J., Joo, D., Bae, J. and Kim, J.: A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning, *CVPR*.
[24] Mobahi, H., Farajtabar, M. and Bartlett, P. L.: Self-Distillation Amplifies Regularization in Hilbert Space, *arXiv* (2020).
[25] Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L. and Anandkumar, A.: Born Again Neural Networks, *ICML*.
[26] Sriram, A., Jun, H., Satheesh, S. and Coates, A.: Cold Fusion: Training Seq2seq Models Together with Language Models, *Interspeech*.
[27] Wu, C.-W. and Lerch, A.: Automatic Drum Transcription Using the Student-Teacher Learning Paradigm with Unlabeled Music Data, *ISMIR*.
[28] Raczyński, S. A., Vincent, E. and Sagayama, S.: Dynamic Bayesian Networks for Symbolic Polyphonic Pitch Modeling, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21, No. 9.
[29] Sigtia, S., Boulanger-Lewandowski, N. and Dixon, S.: Audio Chord Recognition with a Hybrid Recurrent Neural Network, *ISMIR*.
[30] Korzeniowski, F. and Widmer, G.: On the Futility of Learning Complex Frame-Level Language Models for Chord Recognition, *ISMIR*.
[31] Ycart, A., McLeod, A., Benetos, E., Yoshii, K. et al.: Blending Acoustic and Language Model Predictions for Automatic Music Transcription, *ISMIR*.
[32] Korzeniowski, F. and Widnaer, G.: Automatic Chord Recognition with Higher-Order Harmonic Language Modelling, *EUSIPCO*.
[33] Korzeniowski, F. and Widmer, G.: Improved Chord Recognition by Combining Duration and Harmonic Language Models, *ISMIR*.
[34] Ycart, A., Benetos, E. et al.: A Study on LSTM Networks for Polyphonic Music Sequence Modelling, *ISMIR*.
[35] Thompson, L., Mauch, M., Dixon, S. et al.: Drum Transcription Via Classification of Bar-Level Rhythmic Patterns, *ISMIR*.
[36] Luong, T., Pham, H. and Manning, C. D.: Effective Approaches to Attention-based Neural Machine Translation, *EMNLP*.
[37] Bahdanau, D., Cho, K. and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, *ICLR* (2015).
[38] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention Is All You Need, *NIPS*.
[39] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L.: Deep Contextualized Word Representations., *NAACL-HLT*.
[40] Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I.: Improving Language Understanding by Generative Pre-Training (2018).
[41] Devlin, J., Chang, M., Lee, K. and Toutanova, K.: BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, *NAACL-HLT*.
[42] Chen, X., Liu, X., Wang, Y., Ragni, A., Wong, J. H. and Gales, M. J.: Exploiting Future Word Contexts in Neural Network Language Models for Speech Recognition, *TASLP*.
[43] Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Hawthorne, C., Dai, A. M., Hoffman, M. D. and Eck, D.: Music Transformer: Generating Music with Long-Term Structure, *arXiv* (2018).
[44] Huang, Y.-S. and Yang, Y.-H.: Pop Music Transformer: Generating Music with Rhythm and Harmony, *arXiv* (2020).
[45] Jiang, J., Xia, G. G., Carlton, D. B., Anderson, C. N. and Miyakawa, R. H.: Transformer VAE: A Hierarchical Model for Structure-Aware and Interpretable Music Representation Learning, *ICASSP*.
[46] Chen, T.-P. and Su, L.: Harmony Transformer: Incorporating Chord Segmentation into Harmony Recognition, *ISMIR*.
[47] Ba, J. L., Kiros, J. R. and Hinton, G. E.: Layer Normalization, *arXiv* (2016).
[48] Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, *ACM-SIGKDD*.
[49] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical and Jazz Music Databases, *ISMIR*.
[50] Hennequin, R., Khlif, A., Voituret, F. and Moussallam, M.: Spleeter: A Fast and Efficient Music Source Separation Tool with Pre-Trained Models, *Journal of Open Source Software*.
[51] Böck, S., Korzeniowski, F., Schlüter, J., Krebs, F. and Widmer, G.: Madmom: A New Python Audio and Music Signal Processing Library, *ACM international conference on Multimedia*.
[52] Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., Ellis, D. P. and Raffel, C. C.: mir_eval: A Transparent Implementation of Common MIR Metrics, *ISMIR*.
[53] Vogl, R., Widmer, G. and Knees, P.: Towards Multi-Instrument Drum Transcription, *DAFx*.
[54] Loshchilov, I. and Hutter, F.: Decoupled Weight Decay Regularization, *ICLR* (2019).
[55] He, K., Zhang, X., Ren, S. and Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, *ICCV*.