5D-04

信号教師モデルから記号生徒モデルへの知識蒸留を用いた演奏音符列を対象とした楽器分類

澤田 隼 †

東京理科大学 †

1 はじめに

音楽情報処理の分野では，楽器音を分類する研究が多数存在する．音楽音響信号を入力として楽譜を出力する自動採譜に関する研究が盛んに行われているが，完全な楽譜にするためには何の楽器なのかを認識する必要がある．また，楽器の編成が楽曲の特徴を決める重要な要素である点に注目して楽曲を探索する試みもされている [1].

従来の楽器分類の研究は，主に音響信号を対象とした楽器の分類に焦点がおかれている．また，音響信号を対象とした楽器分類でも，単音を対象とした分類や，ある程度の長さの演奏データを対象とした分類，楽曲全体を対象とした分類に分けられる．音響信号を対象とした楽器分類モデルは，主に音色などの特徴を学習し，分類される．

一方で，各楽器の持つ特徴的な音高の動きなどを捉え，演奏音符列を対象とした楽器分類のタスクも存在する．Kevin らは楽譜に書かれた音符列を自然言語のテキスト分類タスクとして扱い演奏音符列を対象とした楽器分類を実現した [2]．本研究でも同様に楽器の楽譜や MIDI などの演奏音符列を対象とした楽器分類を目指す．似た構造を持つ楽器や似た音色を持つ楽器は，同じような演奏がされる（特徴的な音高の動きをする）と仮定すると，楽器間の類似度に関する情報は有用であると考えられる．そこで，音響信号を対象として学習したモデルがとらえる楽器間の関係性に関する知識を，演奏音符列のみを対象とするモデルに転移学習することを考える．本稿では，実際の楽曲の演奏された楽器音の音響信号を対象として学習した教師モデルから，演奏音符列を対象とする生徒モデルに知識蒸留を行い，音符列のみを入力とした楽器分類について提案する．

2 演奏音符列の楽器分類

2.1 問題設定：楽器分類問題

ここでは，本研究で扱う楽器分類問題について述べる．今回対象とする楽器分類は，楽譜の中の 2 小節の音符列を取り出して，何の楽器の譜面であるかを当てる問題と考えることができる．

Musical Instrument Classification for Musical Note Sequence Based on Knowledge Distillation from Audio Teacher Model to Symbolic Student Model
Shun Sawada†, Tokyo University of Science

2.2 提案手法：信号教師モデルから記号生徒モデルへの知識蒸留

知識蒸留（Knowledge Distillation）とは，一般的には複雑な教師モデルの学んだ知識を，軽量な生徒モデルの学習に利用するもので，単純に生徒モデルだけで学習するよりも，良い精度を得ることが知られている．

蒸留する知識として様々な手法があるが，教師の出力をソフトターゲットとして，生徒の出力の分布がこれと近くなるように学習する方法を採用した．

ハードラベル（one-hot など）の学習データから得られる情報は，入力とその正解の情報でしかなく，それ以外のクラスは不正解という情報しかない．一方で，学習済みの教師モデルから得られる情報は，合計すると 1 になる確率の分布と見なすことができ，クラスごとに予測結果が出力される．正解ではないクラスも相対的な予測結果からクラス間の類似性などの情報を知ることができるため，生徒モデルの学習に利用する知識となる．知識蒸留は，単に軽量なモデルで学習できるようになるだけではなく正則化の効果も期待できる．順序を持つ分類タスクでは Label distribution learning と呼ばれる学習方法を利用することによって有効であることが示されている．ラベルがハードラベル（one-hot）ではなく，あるラベルを中心にその出力が正規分布のようになっている．楽器間の類似性や順序関係がわかればその手法を用いることができるが，自明ではない．そこで信号教師モデルの出力を楽器間の類似性と見立てて，生徒モデルがその出力を模倣するように学習する．

本手法では，楽器音の音響信号を対象として学習した教師モデル（信号教師モデル）から，音符列を対象として学習する生徒モデル（記号生徒モデル）に知識蒸留を行い，音符列のみを入力とした楽器分類方法について提案する．本研究で注目しているのはモデルの複雑さではなく，入力が異なる 2 つのモデルの知識を蒸留する点である．

以下に使用したモデルの概要を示す（**表 1**）．楽器分類で有用であることが知られている Convolutional Neural Network（CNN）を用いた．教師と生徒の出力の分布間の損失として KL Divergence を利用した．

2.2.1 前処理

音響信号の前処理は次のように行った．まず，サンプリング周波数 22050 で読み込み，それぞれのトラックを MIDI の 2 小節のセグメントに分割したと対応する箇所を分割する．この際，各トラックの 2 小節のサンプル数は各楽曲のテンポによって異なるため，今回は最大のサンプル数（一番テンポの遅い曲）に合わせて時間軸方向にゼロ埋めを行った．次にメルスペクトログラムに変換した．フーリエ変換の窓幅とシフト幅は 512 とし，メ

表 1　使用したモデル

Layer	Output Shape	
	Teacher	Student
Input	128 × 528 × 1	128 × 16 × 1
Conv2D	64 × 264 × 256	64 × 8 × 256
LeakyReLU	64 × 264 × 256	64 × 8 × 256
MaxPooling2D	64 × 264 × 256	64 × 8 × 256
Conv2D	32 × 132 × 512	32 × 4 × 512
Flatten	2162,688	65,536
Dense	8	8

ルフィルタバンクのビン数は 128 とした．入力次元は (128,528) となる．

MIDI の前処理は次のように行った．最小分解能を 8 分音符とし，それぞれのトラックを 2 小節のセグメントに分割した．データセットの MIDI にはベロシティが含まれているが，想定するシステムは楽譜や音符列の入力であるため，必ずしもベロシティがあるとは限らない．今回はピッチイベントが有れば 1，無ければ 0 の 2 値をとるベクトルに変換した．入力次元は（128 × 16）となる．

3　実験結果

音響教師モデルを用いた演奏音符列の生徒モデルへの知識蒸留が有効であるかを調べるため，次の 3 つのモデルを用意した．

- 音響分類モデル（Audio Instrument Classification: AIC）：音響信号を入力とするモデル
- 記号分類モデル（Note-sequence Instrument Classification: NIC）：演奏音符列を入力とするモデル
- 音響蒸留記号分類モデル（Knowledge Distillation Instrument Classification: KDIC）：AIC モデルから NIC モデルへ知識蒸留を行うモデル

データセットとして演奏音符列と音響信号のペアが必要であるため，音源分離などのタスクで用いられるデータセット Synthesized Lakh (Slakh) Dataset を用いた [3]．これは Lakh MIDI Dataset を FluidSynth で wav ファイルに合成して作られたデータセットである．

楽器分類の正解データは Slakh Dataset の分類の中から次の 8 つを用いた（Guitar, Piano, Bass, Strings, Organ, Brass, Pipe, Reed）．

表 2 に各モデルとその分類結果を示す．音響信号を入力として学習させたモデル AIC の F 値は 0.72 となり，演奏音符列のみを入力として学習させたモデル NIC の F 値は 0.61 となった．演奏音符列のみを入力として，モデル AIC の出力を教師データとして学習させたモデル KDIC の F 値は 0.72 となった．

演奏音符列のみを入力として学習したモデル（NIC）よりも，音響信号を入力として学習したモデルから演奏音符列を対象とする生徒モデルに知識蒸留を行ったモデル（KDIC）のほうが高い精度で分類でき，音響信号のみを対象として学習したモデル（KDIC）と同等の精度で分類することが可能になった．音響信号を入力として学習したモデルを教師モデルとし，演奏音符列を入力とする生徒モデルに知識蒸留を行うことによって，音響信号を対象として学習したモデルがとらえる楽器間の関係性に関する知識を，演奏音符列のみを対象とするモデルにが学習できたことが示唆された．

表 2　結果

モデル	入力		教師	F 値
	音響信号	音符列		
AIC	✓		ハード	0.76
NIC		✓	ハード	0.61
KDIC		✓	ソフト（AIC）	0.72

4　おわりに

本稿では，実際の楽曲の演奏された楽器音の音響信号を対象として学習した教師モデルから，演奏音符列を対象とする生徒モデルに知識蒸留を行い，音符列のみを入力とした楽器分類について提案をした．

演奏音符列のみを入力として学習したモデルよりも，音響信号を入力として学習したモデルから演奏音符列を対象とする生徒モデルに知識蒸留を行ったモデルのほうが高い精度で分類でき，音響信号のみを対象として学習したモデルと同等の精度で分類することが可能になった．音響信号を入力として学習したモデルを教師モデルとし，演奏音符列を入力とする生徒モデルに知識蒸留を行うことによって，音響信号を対象として学習したモデルがとらえる楽器間の関係性に関する知識を，演奏音符列のみを対象とするモデルにが学習できたことが示唆された．これにより，楽器分類にとどまらず，音符列のみを用いた旋律の楽器らしさの指標への応用などが期待される．

今後の課題はさらにデータや使用するモデルを増やし，より詳しい検証を行う．分類楽器の種類も増やし，楽器の階層性や類似性に関する学習が行われているのかを検証する．また，音響教師モデルから知識蒸留を行った演奏音符列生徒モデルの出力が，未知の楽器に対してのヒントとして使用することができるのか，さらに，旋律の楽器らしさの指標として自動作曲システムへの応用に興味がある．

参考文献

[1] 高橋 卓見, 深山 覚, and 後藤 真孝. Instrudive：楽器編成の自動認識に基づく楽曲探索システム. **情報処理学会論文誌**, 61(4):777–788, apr 2020.

[2] Kevin Ji, Daniel Yang, and TJ Tsai. Instrument classification of solo sheet music images. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 546–550, 2021.

[3] Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux. Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019.