リカレントニューラルネットワークを用いた音楽データの予測と生成

武田 敦志 †

† 東北学院大学教養学部情報科学科

1　はじめに

近年、ディープラーニングと呼ばれる多層ニューラルネットワークを用いた機械学習の研究が盛んに行われており、リカレントニューラルネットワーク（RNN）を活用することにより音楽データや文章データの予測や生成が可能となりつつある。特に、Long-Short Term Memory（LSTM）[1] や Gated Recurrent Unit（GRU）[2] を導入した RNN を用いることにより、LSTM や GRU を導入していない RNN よりも正確に音楽データを予測することが可能となる [3]。しかし、最新の RNN 技術を用いたとしても、音楽データの予測の精度は不十分であり、音楽データをより正確に予測する手法が必要である。

そこで、本稿では、LSTM や GRU よりも正確に音楽データを予測することができる Double Track Recurrent Unit（DTRU）を提案する。DTRU は、正確な予測を行うため、RNN の各層が維持する状態値と次層に伝達する出力値を別々に計算するという特徴を持つ。従来手法である LSTM や GRU では、状態値を次の層に出力値として伝達するため、状態値に入出力の関係を含める必要があった。一方、提案手法である DTRU では、状態値と出力値を別々に計算するため、RNN の各層の状態値には各層の状態のみを反映させればよい。そのため、DTRU を導入した RNN を用いることにより、従来よりも正確に音楽データを予測できると考えられる。

本稿では、DTRU の構成を示し、その計算方法を説明する。また、音楽データの予測タスクの実験結果より、DTRU が従来手法よりも正確に音楽データを予測できることを示す。さらに、DTRU を用いることにより、音楽データを生成できることを示す。

2　Double Track Recurrent Unit（DTRU）

DTRU では、RNN の各層の状態値と出力値を正確に計算するため、これらの値を別々に計算する構造となっている。図 1 に DTRU の構造を示す。DTRU では、時刻 t の入力値 $\boldsymbol{x}(t)$ と前回の状態値 $\boldsymbol{h}(t-1)$ から、新しい状態値 $\boldsymbol{h}(t)$ と出力値 $\boldsymbol{y}(t)$ を計算する。また、DTRU では、状態値の制御ゲートである Join Gate と Update Gate が

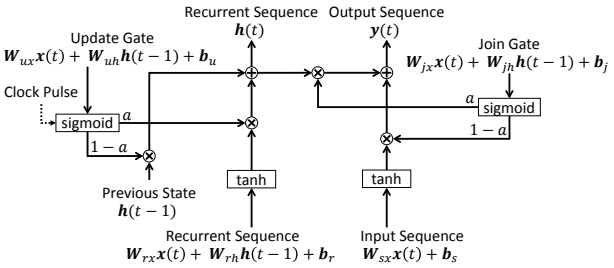

図 1: DTRU の構造

あり、入力値と前回の状態値からこれらゲートの値を計算する。ここで、Join Gate の値は出力値に反映される状態値の割合であり、Join Gate の値が大きくなると状態値をより反映させた出力値となる。一方、Update Gate の値は状態値を更新する割合であり、Update Gate の値が大きくなると状態値をより新しい値に更新する。

時刻 t における状態値 $\boldsymbol{h}(t)$ と出力値 $\boldsymbol{y}(t)$ は

$$\begin{aligned}\boldsymbol{h}(t) &= \boldsymbol{r}(t)\odot\boldsymbol{u}(t)+\boldsymbol{h}(t-1)\odot(1-\boldsymbol{u}(t))\\ \boldsymbol{y}(t) &= \boldsymbol{h}(t)\odot\boldsymbol{j}(t)+\boldsymbol{i}(t)\odot(1-\boldsymbol{j}(t))\end{aligned}$$

となる。ここで、$\boldsymbol{r}(t)$ は前回の状態値を含む入力値であり、$\boldsymbol{i}(t)$ は前回の状態値を含まない入力値である。また、$\boldsymbol{u}(t)$ は Update Gate の値であり、$\boldsymbol{j}(t)$ は Join Gate の値である。$\boldsymbol{r}(t)$、$\boldsymbol{i}(t)$、$\boldsymbol{u}(t)$、$\boldsymbol{j}(t)$ はそれぞれ

$$\begin{aligned}\boldsymbol{r}(t) &= \boldsymbol{W}_{rx}\boldsymbol{x}(t)+\boldsymbol{W}_{rh}\boldsymbol{h}(t-1)+\boldsymbol{b}_r\\ \boldsymbol{i}(t) &= \boldsymbol{W}_{ix}\boldsymbol{x}(t)+\boldsymbol{b}_i\\ \boldsymbol{u}(t) &= \boldsymbol{W}_{ux}\boldsymbol{x}(t)+\boldsymbol{W}_{uh}\boldsymbol{h}(t-1)+\boldsymbol{b}_u\\ \boldsymbol{j}(t) &= \boldsymbol{W}_{jx}\boldsymbol{x}(t)+\boldsymbol{W}_{jh}\boldsymbol{h}(t-1)+\boldsymbol{b}_j\end{aligned}$$

となる。ここで、$\boldsymbol{W}_{rx}$、$\boldsymbol{W}_{rh}$、$\boldsymbol{W}_{ix}$、$\boldsymbol{W}_{ux}$、$\boldsymbol{W}_{uh}$、$\boldsymbol{W}_{jx}$、$\boldsymbol{W}_{jh}$ は結合の重みであり、$\boldsymbol{b}_r$、$\boldsymbol{b}_i$、$\boldsymbol{b}_u$、$\boldsymbol{b}_j$ はバイアスである。

DTRU では Update Gate の値によって状態値の更新割合が決定される。そこで、1 か 0 の値 (Clock Pulse) を Update Gate の値に掛けることにより、Clock Pulse が 1 のときにのみ状態値の更新が行われるようにできる。この仕組みを導入することにより、Clockwork RNN[4] などと同様に階層的なデータの学習が可能となる。

3　実験と評価

DTRU を用いた RNN を実装し、音楽データの予測タスクを用いて性能評価を行った。ここでは、音楽データとして piano-midi.de[5] を用い、64 分の 1 拍子ごとに次に鳴る音の音階と長さを予測する RNN を構築した。構

A new Recurrent Neural Network for Predicting and Representing Music Data

†Atsushi TAKEDA

†Department of Information Science, Tohoku Gakuin University

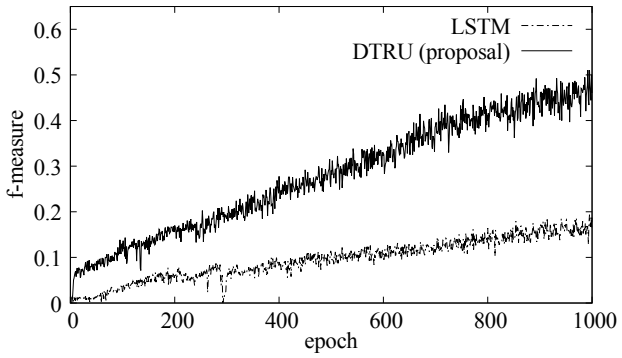

図 2: 訓練データに対する F 値

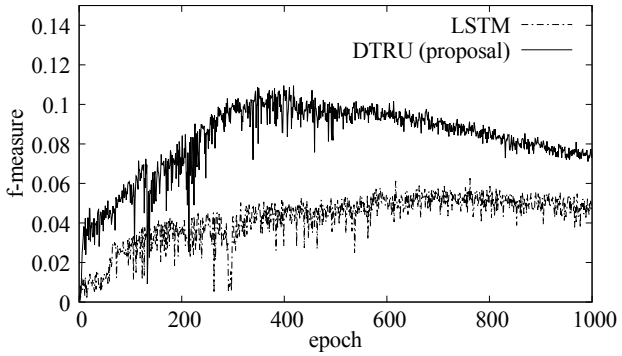

図 3: テストデータに対する F 値

	LSTM	DTRU (提案手法)
訓練データ	0.193	**0.511**
テストデータ	0.063	**0.110**

表 1: LSTM と DTRU の F 値の最高値

築した RNN は Encoder-Decoder 方式 [2] のニューラルネットワークであり、Encoder と Decoder は中間層が 2 層で各層が 400 ユニットの RNN で構成されている。入力は 92 次元ベクトル (音階:87+リズム:5) であり、出力は 2785 次元ベクトル (音階:87× 長さ:32+停止フラグ:1) である。また、Encoder から Decoder に渡される特徴値は 200 次元ベクトルとした。パラメータの学習は Back Propagation Through Time(履歴の長さ=128) を用いて行い、最適化手法として Adam[6](α = 0.0001、β_1 = 0.9、β_2 = 0.999) を用いた。上記の条件で音楽データの予測タスクを行い、DTRU を導入した RNN の訓練データとテストデータに対する F 値 (f-measure) を計測した。さらに、比較のための従来手法として、LSTM を導入した RNN についても同様の計測を行った。

図 2 に訓練データに対する F 値を、図 3 にテストデータに対する F 値を示す。また、表 1 に訓練データとテストデータに対する F 値の最高値を示す。訓練データ・テストデータともに、DTRU の F 値は LSTM の F 値よりも高い値となっている。以上より、DTRU を用いた RNN が LSTM を用いた RNN よりも正確に音楽データを予測できることを確認した。

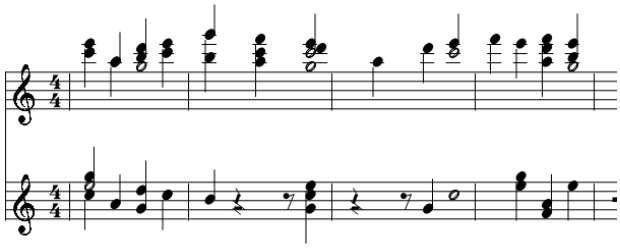

図 4: 生成された音楽データの一部

また、上記の性能評価で実装した RNN を用いて音楽データの生成を行った。ここでは、JSB Chorales[5] を用いてパラメータの学習を行い、任意の特徴ベクトルから音楽データを生成した。図 4 に生成された音楽データの一部を示す。実際に使われる和音 (最初の 4 分音符は c-major) で構成されており、RNN が音階の関係性を考慮して音楽を生成できることを確認した。

4 まとめ

本稿では、LSTM や GRU よりも正確に音楽データを予測することができる Double Track Recurrent Unit（DTRU）を提案した。また、実験結果より、DTRU が従来手法よりも正確に音楽データを予測できることと、DTRU が音楽データを生成できることを確認した。

参考文献

[1] Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural Computation*, 12(10):2451–2471, 2000.

[2] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

[3] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2342–2350, 2015.

[4] Jan Koutník, Klaus Greff, Faustino Gomez, and Jüergen Schmidhuber. A clockwork rnn. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1863–1871, 2014.

[5] Nicolas Boulanger-lewandowski, Yoshua Bengio, and Pascal Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1159–1166, 2012.

[6] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *the 3rd International Conference on Learning Representations (ICLR 2015)*, 2015. arXiv:1412.6980.