# Deep Learning for Audio Signal Processing

（2019）Hendrik Purwins*, Bo Li*, Tuomas Virtanen*, Jan Schlüter*, Shuo-yiin Chang, Tara Sainath

http://arxiv.org/pdf/1905.00078v2.pdf

## どんなもの？

**[Problem Categorization]**
- (Multi-label)Sequence classification : 単一（または複数の）ラベルを予測する問題
- Sequence regression：一定の値（テンポ、次に演奏される音）などを予測する問題

**[Audio Features]**
- 主な特徴量に（振幅対数変換、離散コサイン変換を施した）**メル周波数ケプストラム係数(MFCCs)**を使用。

**[Models]**
- **CNN**：特徴量がスペクトル表現の場合は**1d時間畳み込み、2d時間周波数畳み込み**を採用することが一般的。生の波形を入力する場合は、**時間領域1d畳み込み**が適用される。
- **RNN**：周波数相関をモデル化すべく**周波数LSTM(F-LSTM)、時間-周波数LSTM(TF-LSTM)**を導入。
- **Sequence-to-Sequence**：シーケンス（音の反復進行）をシーケンスに直接変換するモデル。ディープラーニングを活用し、入力信号をターゲットシーケンスに直接マッピングするend-to-endのシステム構築に対する関心が高まっている。
- **Generative Adversarial Networks（GANs）**：低次元のランダム潜在ベクトルから特定のデータセットに関する現実味のあるサンプルを生成する教師なし学習モデル。音源分離、楽器変換、音声強調等でノイズ除去のために使用されてきたが、画像分野に比べて利用は限定的。
- **Loss Functions**：微分可能な**動的タイムワープ距離、アースムーバーの距離**の利用可能性（わずかに非線形に歪んだ信号が似ているように聞こえるという事実を説明する上で有用？）
- **Phase modeling**：スペクトルの対数変換で失われる**位相スペクトル**は音声合成に必要。位相は、Griffin-Limアルゴリズムを使用してマグニチュードスペクトルから推定できる。

**[Data / Evaluation]**
- ラベル付けされた学習用データセットが画像分野に比べて限定的なので、データ補完すべく**Data generataion、Data augmentation、transefer learning**等が用いられている。
- 評価指標：(話者分類) **word error rate**、(シーン分類) **accuracy**、（音声合成）**mean opinion score**

## 技術の手法や肝は？

**<Analysis>**
- Speech：音声認識の社会実装開始は最近。ガウス混合モデル/隠れマルコフモデルが主流だった。
- Music : low-level（基本周波数、リズム推定等）、high-level（楽器検出、音楽の類似性の推定等）。
- Environmental Sounds : 主に音響シーン分類、音響イベント検出、タグ付けを行う。
- Localization and Tracking : マルチチャンネルオーディオを用いた音源の定位と追跡を行う。

**<Synthesis and Transformation>**
- Source Separation : 複数の音源から、個々の音源に対応する信号を抽出する
- Audio Enhancement : ノイズを減らすことで音声品質の改善を目的として、音声を強調する。
- Generative Model：学習した音響特性に基づいて音を合成し、リアルな音のサンプルを生成する。

## 議論はある？

**[Features / Models]**
- 音響分析においてメル周波数ケプストラム係数を利用することが本当に最も望ましいのか？
- どのような状況下で、生の波形を使用すべきか？どのモデルを用いるべきか？

**[Data Requirements]**
- 訓練済のモデルは、新しいタスク（語彙のない単語、新しい音楽様式等）に適応できるか？
- 画像分野のImageNetのように、カテゴリを跨ってラベル付けされたデータセットがない。

**[Computational Complexity]**
- 携帯電話や補聴器など、計算リソースに厳しい制限があるアプリでは小型モデルが必要。

**[Interpretability and Adaptability]**
- レイヤパラメータと実際のタスクとの間の関係性についての解釈が困難。

2019/06/07

# WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

（2016）A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu

https://arxiv.org/pdf/1609.03499.pdf

## どんなもの？

- DeepMind社(Googleが買収)が開発した**"WaveNet"**（生の音声波形を生成するDeep Neural Network）

- WaveNetは、確率的・自己回帰型モデル。causal filterとdilated convolutionを組み合わせることで、**受容野(＝次の信号を予測するために過去のデータをどの程度見るか？)**をレイヤーの深さと共に指数関数的に増加させることが可能。**毎秒数万サンプル**の音声データを訓練できる。

- Text to Speechでは、最良のparametric modelで生成した音声よりも**自然な響き**と評価。

## どうやって有効だと検証した？

Text to Speechにおいて、英語・北京語共に**mean opinion score**が他のモデルを超えた。

=> <mean opinion scoreとは？>「1 : Bad, 2 : Poor, 3 : Fair, 4 : Good, 5 : Excellent」の5段階評価の平均値

(以下、平均スコア ± 標準偏差)
- **WaveNet** : 英語 = **4.21 ± 0.081** / 北京語 = **4.08 ± 0.085**
- LSTM-RNN parametric : 英語 = **3.67 ± 0.098** / 北京語 = **3.79 ± 0.084**
- HMM-driven concatenative : 英語 = **3.86 ± 0.137** / 北京語 = **3.47 ± 0.108**

## 技術の手法や肝は？

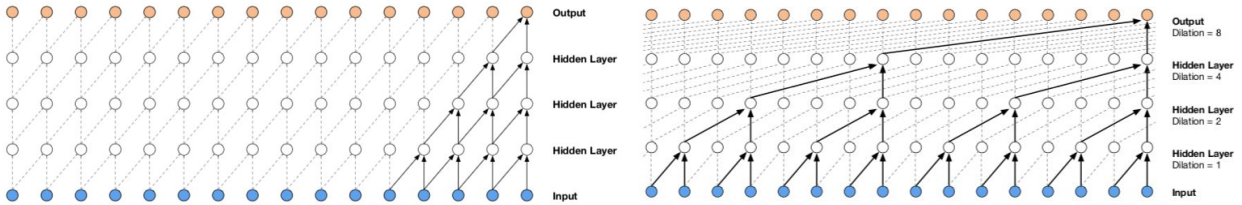
- 波形x = {x 1、...、x T}の同時確率を、条件付き確率の積で表現（各音声サンプルxtは、前のすべてのタイムステップにおけるサンプルに基づいて調整される）。

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

- causal convolutionの課題は、受容野の増大に"多数の層"や"大きなフィルタ"を必要とすること。Wavenetでは、dilated convolutionを用いて課題を克服している。（dilated convolutionの要諦は、**"出力層に近い層ほど離れた要素同士を畳み込む"**ことで、より多くの入力値の情報を出力層に反映すること。

- 先行研究ではデータの値が連続的である（＝画像のピクセル強度、音声の数値）場合でも、softmax分布がよりうまく機能する傾向があることが示されており、WaveNetでは、softmaxを採用している。

（左図）Causal Convolution （右図）Dilated Convolution




## 先行研究と比べて何がすごい？

- Stacked dilated convolutionにより、ネットワーク全体の入力解像度と計算効率を維持しながら、非常に大きな受容野を持つこと（＝少ない畳み込み層で過去の入力情報を多く取り込むこと）に成功した。

- モデルに変数を追加する（＝条件付き確率p(x)の式に状態変数を追加する）ことで、様々な特性を有する音声を生成できる。例えば、multi-speaker（＝複数の話者が存在）の設定では、スピーカーのIDをモデルの変数に追加することでスピーカーを選択できるようになる。

## 次に読むべき論文は？

- 1. Gonzalvo, Xavi, Tazari, Siamak, Chan, Chun-an, et al. "Recent advances in Google real-time HMM-driven unit selection synthesizer. "In *Inter- speech*, 2016.
- 2. Zen, et al. "Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices." In *Interspeech*, 2016.
- 3. van den Oord,et al. "Conditional image generation with PixelCNN decoders." *CoRR*, (2016)

2019/06/07

# Conditional Image Generation with PixelCNN Decoders

（2016）Zen, Heiga, Agiomyrgiannakis, Yannis, Egberts, Niels, Henderson, Fergus, and Szczepaniak, Prze- mysław.

https://arxiv.org/pdf/1606.05328.pdf

## どんなもの？

- PixelRNNアーキテクチャの基本的な考え方は、autoregressive connectionsを利用して、イメージをピクセルごとにモデル化し、**画像の同時分布を条件付き確率の積に分解**する。
=> ImageNetのクラスラベルを用いて、現実味のある多彩な画像（異なる動物、物体、風景、構造）を生成することができる。

- 本稿ではPixel CNNにGated Convlution Layerを導入し（**Gated PixelCNN**）、計算効率を向上させた。

## どうやって有効だと検証した？

2つのDatasetについてNLL testを実施（対数尤度を測定）。結果を先行研究と比較した。

1. CIFAR-10 dataset
=> Gated PixelCNNはPixelCNNよりも0.11(bit/dim)優れており、生成されるサンプルの視覚品質が高い。

2. ImageNet
=> 半分以下の学習時間（32 GPUを使用して60時間）で、PixelRNNと同等の精度を達成。

## 技術の手法や肝は？

- Conditional Pixel CNN：latent vector(“h”)に依存する画像の条件付き分布p(x | h)をモデル化した。

$$p(\mathbf{x}|\mathbf{h}) = \prod_{i=1}^{n^2} p(x_i|x_1, ..., x_{i-1}, \mathbf{h}).$$

- また、活性化関数をゲート付き活性化関数に変更すると共に、latent vector(“h”)を加味した。

$$\mathbf{y} = \tanh(W_{k,f} * \mathbf{x} + V_{k,f}^T \mathbf{h}) \odot \sigma(W_{k,g} * \mathbf{x} + V_{k,g}^T \mathbf{h}),$$

- 例えば、hがクラスを指定するOne-Hot Encodingの場合、全てのレイヤでクラスに依存するバイアスを追加することを意味する。

- これにより、特定の動物や物を出現させるように指定することができ、さらに異なる位置や姿勢、背景で出現させることができる。

## 先行研究と比べて何がすごい？

- LSTMは、反復的な接続（reccurent connections）により前の層のピクセルの近傍全体にアクセスできる利点がある一方で、pixelCNNに利用可能な近傍の領域が畳み込みスタックの深さと共に線形に増大する課題を抱えている。

- 課題を解消すべくpixelCNN内の畳み込み層間の（線形の）活性化ユニットを、**ゲート付き活性化ユニット**に置き換え、精度を維持しつつ計算時間を短縮した。

## 次に読むべき論文は？

- 1. Lucas Theis and Matthias Bethge. Generative image modeling using spatial LSTMs. In *Advances in Neural Information Processing Systems*, 2015.

2019/06/07

# Generative Image Modeling Using Spatial LSTMs

（2015）Lucas Theis and Matthias Bethge.

https://arxiv.org/pdf/1506.03478

## どんなもの？

- 画像生成分野では、pixel間の強い**統計的依存性**が自然な画像の分布のモデル化を困難にしている。

- Recurrent Neural Network（RNN）が長期依存性の解消に成功し、画像生成への応用が期待される。

- RNNにおけるLong Short-term model（LSTM）を空間的に応用したモデル**（RIDE）**を紹介する。

## どうやって有効だと検証した？

- van Haterenのデータセットから抽出された画像パッチと大きな画像の平均対数尤度率を比較した。

- RIDEはすでにパッチのMCGSMより優れている。

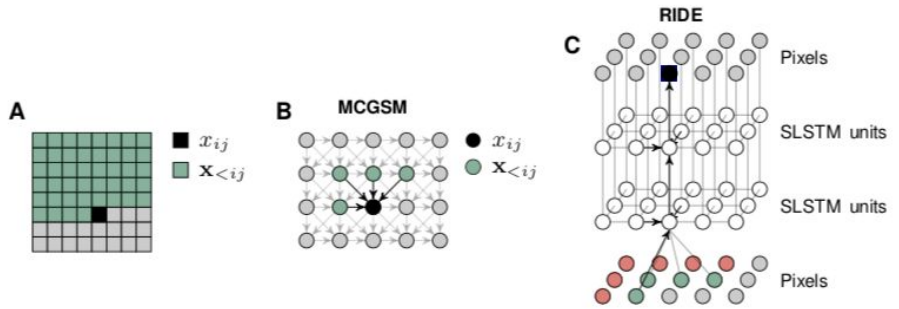
- 尤度とは”モデルの当てはまりの良さをあらわす統計量”であり、最尤法はそれを最大にするようなパラメータを探そうとする推定方法。平均対数尤度が高いということは、それだけモデルの説明力があると言える。

| Model | 256 dim. [bit/px] | ∞ dim. [bit/px] |
|---|---|---|
| GRBM [13] | 0.992 | - |
| ICA [1, 48] | 1.072 | - |
| GSM | 1.349 | - |
| ISA [7, 16] | 1.441 | - |
| MoGSM, 32 comp. [40] | 1.526 | - |
| MCGSM, 32 comp. | 1.615 | 1.759 |
| RIDE, 1 layer, 64 hid. | **1.650** | 1.816 |
| RIDE, 1 layer, 128 hid. | - | 1.830 |
| RIDE, 2 layers, 64 hid. | - | 1.829 |
| RIDE, 2 layers, 128 hid. | - | 1.839 |
| EoRIDE, 2 layers, 128 hid. | - | **1.859** |

## 技術の手法や肝は？

- （下図A）ピクセル（黒色）の予測が左上の緑の領域のどのピクセルにも依存するように、画像の分布を因数分解した。
=> 因果性のある近傍を持つmixtures of conditional Gaussian scale mixtures （MCGSM）を図示すると（下図B）のようになる。

- （下図C）は2層のspatial LSTMを用いたリカレント画像モデルの視覚化したもの。feedforward connectionでは、ピクセルの予測はその近傍（緑色）に直接依存しており、spatial LSTMの層を深くすることで広範囲の領域（赤色）の情報へのアクセスを可能にした。




## 先行研究と比べて何がすごい？

- 回帰的ネットワークは、recurrent image density estimator (RIDE)が予測のためにはるかに大きな領域のピクセルを使用し、MCSGMを適用する前にピクセルを非線形に変換することを可能にした。

- 空間LSTM（SLSTM）のレイヤーを積み重ねることで、モデルの表現力をさらに高めることができる。

2019/06/07

# Long Short-term Memory

(1997) S. Hochreiter and J. Schmidhuber

https://www.bioinf.jku.at/publications/older/2604.pdf

## どんなもの？

- Recurrent Nurarl Networkの再帰的バックプロパゲーションによる学習は、非常に時間を要する。

- 本稿では、**Long Short-Term Memory（LSTM）**と呼ばれる、新しい効率的な学習方法を紹介。
=> 神経科学における短期記憶（short-term memory）・長期記憶(Long-term memory）から着想した。

- 応用分野には、（1）時系列予測、（2）作曲、（3）音声処理が含まれる。

## どうやって有効だと検証した？

- RNNの評価でベンチマークとされる**"Embedded Reber Grammar(ERG)"**問題（入力した文字列に対して、特定のパターンの文字列を出力）、ノイズを加えたシーケンス等でモデルのパフォーマンスを測定。

## 技術の手法や肝は？

- RNNの学習方法は、**BPTT(Back-Propagation Through Time)法**と**RTRL(Real-Time Recurrent Learning)法**の2つが主流。両手法ともComplete Gradientベースのアルゴリズム。

=> これらのアルゴリズムは多くの場合、逆方向に（時間を遡る方向に）誤差を伝播させた時に勾配が「爆発」または「消滅」する問題を抱えていた。
=> 経路の情報（時間的に過去の情報、あるいは文脈など）が予測に重要な場合、これを回避したい。

- LSTMでは、勾配が爆発しないように、前のセルからの出力を短期記憶、長期記憶に分けると共に、重みの更新で不要な情報を捨象するゲートが組み込まれている。

- 具体的にLSTMは以下3つのゲートを有する。

1. 忘却ゲート(forget gate)
2. 入力ゲート(input gate)
3. 出力ゲート(output gate)

=> シグモイド関数σが、流れてくる信号の量を調節するゲート（重み）の役割を果たす。
（if σ = "0：閉鎖、0.5：半開、1：全開"）

- メモリセル（右図のbox）で入力値を保持する
=> 1step遅れで値を再入力（Scj = Scj + g*y_in）

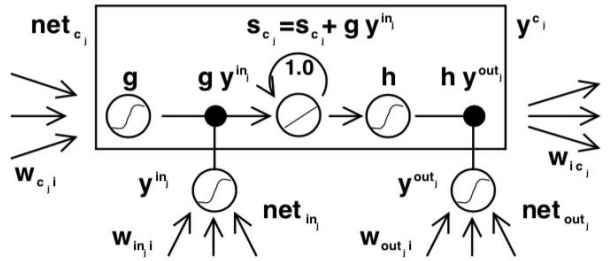



Figure 1: *Architecture of memory cell $c_j$ (the box) and its gate units $in_j, out_j$. The self-recurrent connection (with weight 1.0) indicates feedback with a delay of 1 time step. It builds the basis of the "constant error carrousel" CEC. The gate units open and close access to CEC. See text and appendix A.1 for details.*

## 議論はある？

- LSTMにおける効率的な**"記憶打ち切り型逆伝搬（truncuated backpropagation）"**では、strongly delayed XOR問題の解決が容易ではない。(strongly delayed XOR問題のゴールは、ノイズの多いシーケンス内で発生した2つの（遠く離れた）入力値に対して、XORを計算すること)

## 先行研究と比べて何がすごい？

- 従来のRNNでは学習できなかった**長期依存(long-term dependencies)**を学習可能にした。
=> メモリセル内でのconstant error backpropagationにより、時間的な遅れが考慮される。

- 長い時間差を考慮する必要がある問題で、様々な特徴量（ノイズ、分散表現、連続値）を処理可能。

- 適切な短いタイムラグ訓練の見本に頼ることなく、入力シーケンス内の特定の要素の2つ以上の遠く離れた発生を区別することを学習する。

- （学習率、入力ゲートバイアス、出力ゲートバイアスなど）幅広いパラメータの組合せで機能する。

2019/06/07