論文 / 著書情報
Article / Book Information

論題(和文)	活性化関数のパラメータ制御を用いた LSTM による音声認識
Title(English)	Speech Recognition with Deep LSTM Networks via parametered activation function of LSTM's ” switch ” Gates
著者(和文)	松山祐輔, Ryan Price, 篠田浩一
Authors(English)	Yusuke Matsuyama, Ryan Price, Koichi Shinoda
出典(和文)	日本音響学会2015年秋季研究発表会講演論文集, , , pp. 1-2
Citation(English)	Proc. of ASJ (Acoustical Society of Japan) September 2015, , , pp. 1-2
発行日 / Pub. date	2015, 9

活性化関数のパラメータ制御を用いたLSTMによる音声認識*

◎松山祐輔, Ryan Price , 篠田浩一 (東工大)

1 はじめに

Long Short Term Memory (LSTM) は、ニューラルネットワークを構成するアーキテクチャの一つである。通常の Recurrent Neural Network (RNN) と比べ、より時系列間の情報を長く記憶できるように設計されており、時系列データのモデリングにより適している。LSTM を用いた音声認識 [1] は、これまでの Deep Neural Network (DNN) を用いた結果 [2] よりも大幅に精度が向上することが報告されている。

本研究では、LSTM を制御している 4 つのノードのうち、制御スイッチとして機能している 3 つのノードの活性化関数に注目した。この関数にその形を決めるパラメータを付与し、他の重みパラメータと同様に学習させることにより、TIMIT コーパスにおける音素識別率が改善したことを報告する。

2 LSTM のスイッチノードにおける活性化関数

Long Short Term Memory[3] は、系列モデルの学習における強力な学習モデルである。Fig. 1 にその構造を示す。

LSTM は Constant Error Carousel (CEC) と入力ノード、及びこれらを制御する 3 つのノードにより成り立っている。このうち Input, Output, Forget ゲートは、これまで一般的にスイッチのようなものと考えられてきた。例えば、Output ゲートは「セルが記憶しているを出力するかどうか」のスイッチとして表現される。以下これらをまとめてスイッチノードと呼ぶ。しかし、活性化関数はシグモイド関数が用いられることから、その出力は 0 や 1 の間の連続的な値をとることとなり、「スイッチ」の直感に反している。

そこで、これらスイッチノードの活性化関数を以下のように設定することを提案する。

$$\sigma(x, a) = \frac{1}{1 + e^{-ax}} \tag{1}$$

ここで a は一般的に Gain と呼ばれる値である。Gain 値が大きいほど勾配が急な曲線となり、ノードの出力が 0 か 1 に偏りやすくなる。我々は、この Gain 値についても、他の重みパラメータと同様に学習させることを提案する。

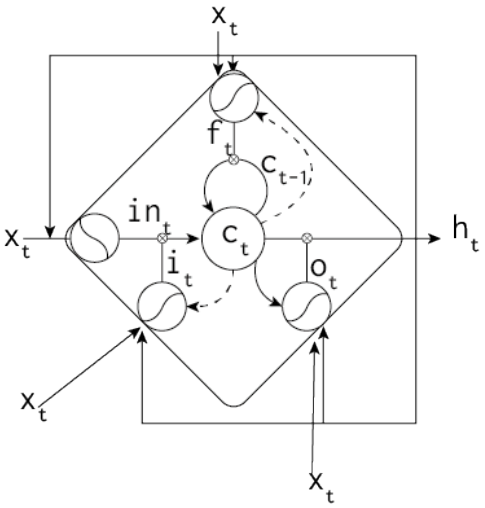

Fig. 1 Architecture of LSTM. i_t, o_t, f_t corresponds to output of Input, Output, Forget Gate.

3 評価実験

3.1 実験条件

本研究ではデータセットとして、TIMIT コーパス [4] を用いる。このうち、462 話者 3696 発声分を訓練データ (TrainSet) に、別の 50 話者 400 発声分を早期停止の判断のために (DevSet)、Core テストセット 24 話者 192 発声分をスコアの計算に用いた (TestSet)。言語モデルとしては、音素 2-Gram を用いた。

強制アラインメントを得るための GMM-HMM システムの構築には kaldi[5] を用いた。Alex らの研究 [1] と同じ条件となるよう、学習時は 61 音素で学習し、スコアの計算時に 39 音素へのマッピングを行う。この Monophone-GMM-HMM のテスト音素誤り率は 31.9% となった。GMM-HMM の訓練には MFCC12 次元と対数パワー、それらの一次および二次微分合わせて 39 次元の入力を用いた。また、LSTM の訓練には filterbank 特徴量を用いた。入力の次元数は、40 次元+エネルギーの合わせて 41 次元と、その一次および二次微分あわせて 123 次元となる。すべてのデータは、平均が 0, 分散が 1 となるように正規化した。

ネットワークには、2 層の Bi-directional LSTM を用いた。各層は 500 個 (前向き後ろ向き各 250 個) の LSTM ブロックをもち、重みパラメータの数は約 234 万となる。提案手法の LSTM ネットワークのパラメータ数は、既存の LSTM ネットワークと比べ、ネット

*Speech Recognition with Deep LSTM Networks via parametered activation function of LSTM's "switch" Gates. by MATSUYAMA Yusuke,Ryan Price,and SHINODA Koichi(Tokyo Institute of Technology)

Table 1　Average PER

	L2	testFER	testPER
LSTM	0.0001	32.044	21.754
LSTM with Gain	0.001	31.752	**21.581**

Table 2　Best PER of All experiments

	L2	testFER	testPER
LSTM	0.0001	31.801	21.382
LSTM with Gain	0.001	31.167	**21.178**

ワークがもつ LSTM 素子の数 1000 にスイッチゲートの数である 3 を掛けた 3000 個増える。これらのうち重みパラメータについては、$[-0.1, 0.1]$ の、バイアスパラメータおよびゲイン値は $[0.9, 1.1]$ の一様分布からランダムに生成した。出力ラベルは 183 個である (61 音素 ×HMM3 状態)。クロスエントロピー誤算関数の最小化には、ADAM 法 [6] を用いた。L2-weight decay のパラメータを $\{0, 0.0001, 0.001\}$ と変更して実験した。初期パラメータによって結果が変わることを考慮し、同じ条件の実験を各 5 回行った。

評価尺度として、音素誤り率 (Phone Erro Rate, PER) および、フレームごとの分類誤り率 (Frame Error Rate, FER) を用いる。

3.2　結果と考察

ネットワークの 2 層目における、各ゲイン値の学習結果を Fig. 2 に示す。また、2 層目の Output ゲートの出力の比較を Fig. 3 に示す。これらのことから、主に Output,Input ゲートにおいて、活性化関数のゲイン値が正の方向に学習され、出力がより 0 や 1 に近づくようになっていることがわかる。

Table. 1 にテスト音素誤り率の平均値がもっとも小さかった L2 の値における結果、Table. 2 には全実験の中でもっともテスト音素誤り率が小さかったものの結果を示す。活性化関数にパラメータを付与することにより、テスト音素誤り率が平均で 0.17 ポイント削減できたことがわかる。

4　おわりに

LSTM のスイッチノードの活性化関数のパラメータを学習させることにより、音素誤り率を削減できることが示された。また、Output および Input ゲートについては、より 0 や 1 に近い値をとるほうが良いのだろうということが考察された。今後の課題としては、これらの知見を元に、さらに効果的な活性化関数を考案することが考えられる。

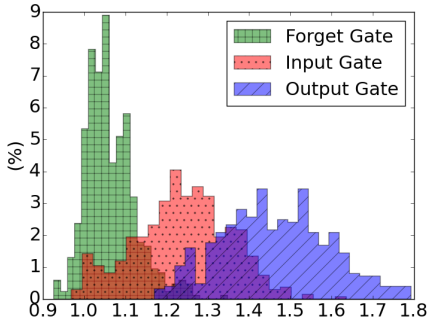

Fig. 2　Distribution of Gain a after Learning.

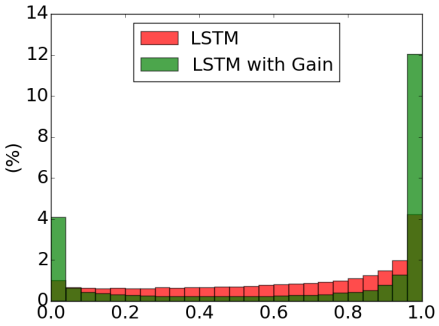

Fig. 3　Comparison of Output value of Output gate.

参考文献

[1] Alex Graves, Navdeep Jaitly and Abdel-rahman Mohamed, “Hybrid Speech Recognition With Deep Bidirectional LSTM”, ASRU, 2013.

[2] Alex Graves, Abdel Rahman Mohamed and Geoffrey Hinton, “Speech Recognition With Deep Recurrent Neural Networks”, ICASSP, 2013.

[3] Sepp Hochreiter and Jurgen Schmidhuber, “Long Short-Term Memory”, *Neural Computation*, pp.1735-1780,1997.

[4] Carla Lopes, Fernando Perdigao, “Phone Recognition on the TIMIT Database”, Speech Technologies, Chapter14 ,2011.

[5] Povey *et al.* “The Kaldi Speech Recognition Toolkit”, ASRU 2011.

[6] Diederik P. Kingma ,Jimmy Lei Ba “ADAM:A Method For Stochastic Optimization”, ICLR 2015.