Vol.2018-SLP-123 No.9
2018/7/27

End-to-End 音声認識システムの強化学習の検討

ポン イリュウ[1]　柴田 駿人[1]　篠崎 隆宏[1]

概要：現在，音声認識システムは高い性能が得られるようになり応用が進みつつあるが，システム開発において教師あり学習のための大量のラベル付きデータが必要で開発コストが非常に大きい問題がある．この問題に対する解決策として，強化学習を応用し多数のユーザーからのフィードバックをラベルの代わりとして活用することが考えられる．しかし一般的な構造の End-to-End モデルを実際に強化学習を用いて学習しようとすると，学習が進まず収束しない場合が多い．そこで本研究では，モデル構造や目的関数の設定などについて様々な検討を行い，強化学習を効果的に進めるための条件を明らかにする．

キーワード：音声認識，ゼロリソース，End-to-End，強化学習，方策勾配法

1. はじめに

近年，教師あり学習によって学習した音声認識システムは高い認識性能を実現している．しかしシステムの学習には大量のラベル付きデータが必要であり，大きなコストがかかる問題がある．そのため，ラベル付きデータを用いない教師なし学習法の応用によるゼロリソース音声認識が模索されている [1]．しかし教師なし学習のみを用いて単語綴りや意味理解を含めた高度な学習を実現することは，非常に難しい目標である．そこでもう一つの解決策として，強化学習を用いることが考えられる．

音声認識システムにおける強化学習の応用としては，これまでに方策勾配法 [2], [3] を用いて単語誤り率を直接的に最適化する手法 [4], [5], [6] や，多数のユーザからのフィードバックを利用するオンラインモデル適応法 [7] が提案されている．しかしこれらはいずれも教師あり学習によって学習したモデルを用いている．教師あり学習による初期化をせずに強化学習を行うのは困難であると言われており [8]，完全な教師なし学習の条件での強化学習を用いた研究は行われていない．その為それがどの程度困難なのかを含め，不明な点が多い．そこで本研究では報酬値のみを用いた方策勾配法による強化学習の学習実験を行うとともに，学習を効果的に進めるためのネットワーク構造や報酬値の与え方について検討を行う．

[1] 東京工業大学
Tokyo Institute of Technology, Kanagawa, Japan
www.ts.ip.titech.ac.jp

2. 方策勾配法

方策勾配法は強化学習の手法の 1 つである．方策勾配法における定式化では，一般にシステムの方策関数 f が状態 s を受け取り行動 a の確率分布 $\pi_\theta(a|s)$ を出力する．ここで θ は方策関数のパラメタである。行動は $\pi_\theta(a|s)$ からサンプリングされ，実行される．音声認識の場合，状態 s は音響特徴量系列，行動 a は認識仮説である．そして，システムは行動に対して報酬 $r_s(a)$ を受け取る．

学習の目的は報酬の期待値 $\mathbb{E}\left[r_{\boldsymbol{s}}\left(a\right)\right] = \sum_a \pi_{\boldsymbol{\theta}}\left(a|\boldsymbol{s}\right) r_{\boldsymbol{s}}\left(a\right)$ を θ に関して最大化することである．最大化は勾配法を用いて行うが，直接にこの期待値の勾配を求めることは困難である．そこで自然進化戦略 [9], [10] と同様に，log トリック $\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}\left(a|\boldsymbol{s}\right) = \frac{\nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a|\boldsymbol{s})}{\pi_{\boldsymbol{\theta}}(a|\boldsymbol{s})}$ による式変形を行う．

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathbb{E}\left[r_{\boldsymbol{s}}\left(a\right)|\,\boldsymbol{\theta}\right] &= \nabla_{\boldsymbol{\theta}} \sum_a \pi_{\boldsymbol{\theta}}\left(a|\boldsymbol{s}\right) r_{\boldsymbol{s}}\left(a\right) \\ &= \sum_a \pi_{\boldsymbol{\theta}}\left(a|\boldsymbol{s}\right) \left(\frac{\nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}\left(a|\boldsymbol{s}\right)}{\pi_{\boldsymbol{\theta}}\left(a|\boldsymbol{s}\right)}\right) r_{\boldsymbol{s}}\left(a\right) \\ &= \mathbb{E}\left[r_{\boldsymbol{s}}\left(a\right) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}\left(a|\boldsymbol{s}\right)\right]. \qquad (1)\end{aligned}$$

式 (1) は，$r_{\boldsymbol{s}}\left(a\right) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}\left(a|\boldsymbol{s}\right)$ が期待報酬の勾配の不偏推定量であることを示している．従って，パラメタ θ の勾配法による最適化は式 (2) により実現できる．

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \epsilon r_{\boldsymbol{s}}\left(a\right) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}\left(a|\boldsymbol{s}\right). \qquad (2)$$

ここで，$\epsilon(>0)$ は学習率である．方策関数がニューラルネットワークの場合，このアルゴリズムは REINFORCE [11] と呼ばれる．

ⓒ 2018 Information Processing Society of Japan

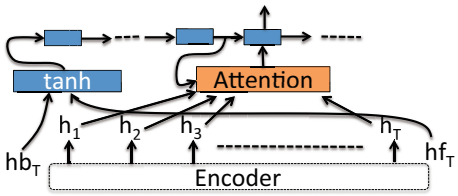

図 1 Attention モデル.

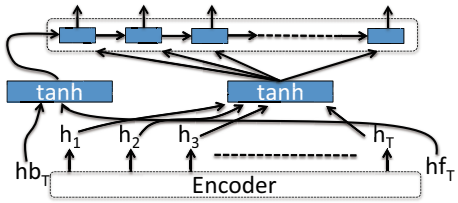

図 2 Spoke(in,out) モデル.

3. モデル構造の検討

本研究では End-to-End モデル [12], [13] として，図 1 および図 2 に示す 2 つの encoder-decoder 型ネットワークを検討した．2 つのモデルにおいて encoder と decoder の構造は共通で，encoder は順方向逆方向ともに隠れ層のユニット数が 128 の 5 層の BiLSTM，decoder は隠れ層のユニット数が 256 の 1 層の LSTM である．また encoder の出力層における順方向文脈ベクトル hf_T と逆方向文脈ベクトル hb_T を連結し，線形変換と tanh 変換を適用したベクトルを decoder の初期状態としている．メモリセルについても同様に連結し，線形変換と tanh 変換を適用したベクトルを decoder のメモリセルの初期値としている．2 つのモデルの違いは，encoder におけるフレームごとの隠れ層出力の利用法である．図 1 の方のモデルは一般的な attention 機構 [14] を用いたネットワークであり，以下では “**attention**” モデルと呼ぶ．図 2 のモデルは我々が提案する構造で attention 機構を用いず，代わりに隠れ層出力をすべて連結し，それに線形変換と tanh 変換を適用した 512 次元のベクトルを decoder ネットワークの各フレームに入力している．線形変換の次元を揃えてすべの発話で共有するために，このモデルでは入力の最大長を仮定し（本論文の実験では 570 フレーム），不足分はゼロ詰めしている．

4. 報酬の与え方の検討

本研究では強化学習の初期研究として，正解精度 (Acc) をもとにした報酬について検討する．Acc の評価には書き起こしテキストを必要とするため，実際の教師なし学習条件では Acc を報酬として用いることはできない．しかし Acc は音声認識において最も基本的な評価尺度であり理想的な報酬が与えられた場合とみなすことができ，モデル構造や報酬の与え方について他の報酬を用いる場合にも有用な知見が得られると期待されるためである．

しかし予備実験において，Acc をそのまま報酬として用いると学習が全く進まず，システムは常に一単語（一数字）のみからなる認識仮説を出力をする問題があることが明らかになった．これはランダムに初期化されたシステムでは出力もランダムで正解精度が 0% を下回っており，挿入エラーを防ぐ単純な戦略として常に最小限の長さの認識仮説を出力するように学習が行われるためと考えられる．そこでこの知見に基づいて，以下の報酬値について検討する．

4.1 クリップ正解精度 ($ClpAcc$)

基本となる Acc は式 (3) により定義される．

$$Acc = \frac{N_{ref} - E}{N_{ref}} = 1 - Err. \tag{3}$$

ここで N_{ref} は正解文の長さであり，E は認識結果のエラー数，Err は誤り率である．エラー数 E は，挿入エラー数 (I)，削除エラー数 (D) および置換エラー数 (S) の和である．負の Acc が学習に悪影響を与えないようにすることを目的として，式 (4) に示す Acc の 0 以下の値をクリップした $ClpAcc$ を定義する．

$$ClpAcc = \max\{Acc, 0\}. \tag{4}$$

4.2 対称正解精度 ($SymAcc$)

出力の長さの問題を考慮する別の方法として，式 (5) に示す対称正解精度 ($SymAcc$) を定義する．

$$SymAcc = \max\left\{\frac{N_{ref} - E}{2N_{ref}} + \frac{N_{hyp} - E}{2N_{hyp}}, 0\right\} \tag{5}$$

$$= \max\left\{1 - \frac{1 - Acc}{2}\left(1 + \frac{N_{ref}}{N_{hyp}}\right), 0\right\}. \tag{6}$$

ここで，$N_{hyp} = N_{ref} + I - D$ は認識仮説の長さである．式 (6) に示すように，出力の長さが正解文より短い場合には報酬値が小さくなることから，システムが長い出力をするように学習を進める効果が期待できる．

4.3 長さペナルティ付き正解精度 ($LPAcc$)

出力の長さを考慮したもう 1 つの方法として，式 (7) によって長さペナルティ付き正解精度 ($LPAcc$) を定義する．$LPAcc$ では出力の長さと正解文の長さの差を用いて明示的にペナルティを与える．α はペナルティ係数であり，予備実験に基づいて 0.3 とした．

$$LPAcc = \max\{Acc - \alpha\,|N_{ref} - N_{hyp}|, 0\}. \tag{7}$$

4.4 平均報酬クリップ対称正解精度 ($SymAccRMC$)

方策勾配法では reinforcement baseline を導入し分散を減少させることが学習の改善に有用であることが知られている [11], [15]．同様の効果を期待して，式 (8) に示すように先行サンプルの平均性能 m で報酬値をクリップした平均報酬クリップ対称正解精度 ($SymAccRMC$) を定義する．

表 1 学習データにおける 1 発話中の数字数の分布

# digits	1	2	3	4	5	6	7
frequency	2464	1232	1232	1332	1132	0	1231

表 2 教師あり学習によって 2 つのモデルから得られた DER.

	attention	spoke(in,out)
DER	1.3%	6.7%

$$SymAccRMC = \begin{cases} SymAcc & (SymAcc \geq m) \\ 0 & (otherwise) \end{cases}. \quad (8)$$

ここで m は直前の 8.5k サンプルの Acc の平均とした．これは，直前の平均 Acc を基準として高い報酬が得られたサンプルだけを学習に使用することと等価である．

5. 実験

5.1 実験条件

本研究では英語の連続数字音声コーパスである TIDIGITS [*1] 中の成人話者による発話データを用いた．学習データは男性話者 55 名と女性話者 57 名が発話した計 8623 発話からなる 4.2 時間の音声である．同様に，テストデータは男性話者 56 名と女性話者 57 名が発話した 8700 サンプルからなる 4.3 時間の音声である．表 1 に，学習データにおける一発話当たりの数字数の分布を示す．音響特徴量は Kaldi toolkit [16] を用いて作成した 13 次元 MFCC 特徴量である．実験は TensorFlow [17] を用いて行った．ミニバッチサイズは 64 である．

比較のため，学習サンプルに対してラベルをランダムに推定し，一致した場合のみ教師あり学習を行うという学習戦略をベースラインとして用いた．すなわち，あてずっぽうの答えがたまたまあっていたら，それを正解ラベルとして用いる戦略である．ラベルのランダム推定は学習データ中の発話ラベルの分布が既知として，その分布から発話単位のラベルをサンプルすることで行った．学習データに対して正しい推定が行われる確率は 0.77% である．

5.2 実験結果

表 2 に，教師あり学習を行った場合の 2 つのモデルの Digit Error Rate (DER)[*2]を示す．spoke(in,out) よりも attention の方が精度が高い．パラメタの更新には ADAM [18] を用いた．

図 3 に，ベースライン学習戦略と強化学習の比較を示す．初期モデルのパラメタは乱数により初期化している．この図では，強化学習の報酬に $SymAcc$ を用いた場合について attention と spoke(in,out) の 2 つのモデルの DER を示している．予備実験により勾配法の適用において ADAM では学習が全く進まなかったため，SGD を用いている．学習率は 0.0005 とした．ベースライン学習戦略において attention モデルを用いた場合，学習がほとんど進まない．特に初めのうちは常に長さ 1 の同一の認識仮説ばかりが出力された．$1.2 * 10^7$ サンプルを学習した後 DER はわずかに改善したが，出力の長さは 1 のままであった．これは表 1 に示されるように長さ 1 の学習データが多く，また attention 機構によって誤ったアライメントが行われることが原因として考えられる．強化学習では attention を用いると同様に全く学習が進まない一方で，spoke(in,out) では単調かつベースライン戦略よりも大幅に DER が減少している．これは attention と異なりアライメントを考慮しないことが有利に働いていると考えられる．

attention の問題点をさらに検証するため，補足実験として教師あり学習によってモデルの初期化をした後，強化学習を行った．教師あり学習による学習で DER 12.0% が得られた後で強化学習に切り替えて学習を行ったところ DER 10.8% に改善し，初期モデルがある程度の認識性能を持つ場合には attension モデルにおいても強化学習が有

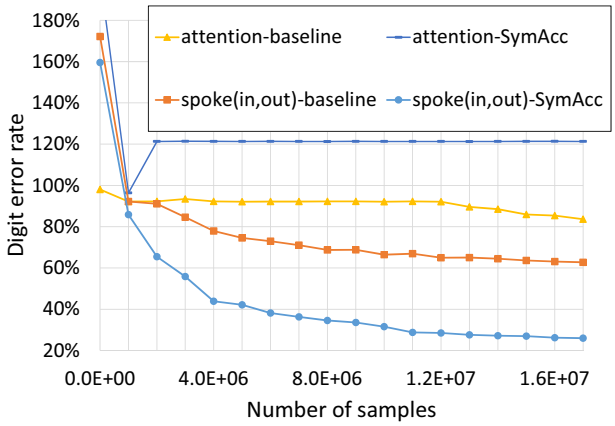

図 3 報酬に $SymAcc$ を用いた強化学習におけるモデル構造の違いがテストデータの DER に与える影響．“baseline” はベースライン学習戦略．横軸は学習サンプル数を表す．

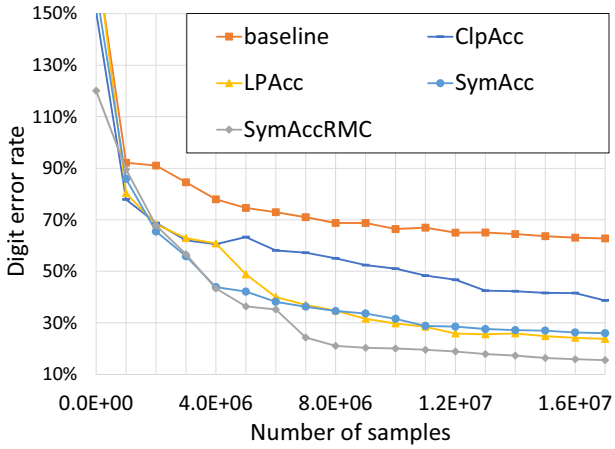

図 4 モデルに spoke(in,out) を用いた場合の，ベースライン学習戦略と様々な報酬値を用いた強化学習のテストデータの DER の比較．横軸は学習サンプル数を表す．

[*1] https://catalog.ldc.upenn.edu/ldc93s10
[*2] DER の定義は数字を単位とする以外 Word Error Rate (WER) と同じである

ⓒ 2018 Information Processing Society of Japan

効に進むことが確認できた．

図 4 に，報酬の違いによる DER の比較を示す．モデルには spoke(in,out) を用いている．全ての強化学習において，ベースラインより高い認識性能が得られた．提案した報酬定義のうち $SymAcc$ と $LPAcc$ はほぼ同等の性能で，$ClpAcc$ より高い性能が得られた．ただし $SymAcc$ は $LPAcc$ とは異なりペナルティ係数がなく，チューニングの必要がないという利点がある．最良の結果が得られたのは $SymAccRMC$ で，$1.7 * 10^7$ サンプルを使って学習した時点で，15.5% の DER が得られた．

6. まとめ

本研究は End-to-End 音声認識システムの方策勾配法による強化学習について検討した．教師あり学習による初期化をしない場合，強化学習は困難であるものの，モデル構造や報酬の設計を工夫することにより学習が進むことを示した．今後の課題としてはより少ない学習サンプル数での効率的な学習を実現することや，より難しいタスクにおける学習を行うことなどが挙げられる．

謝辞 本研究は JSPS 科研費 17K20001 の助成を受けたものです．

参考文献

[1] Versteegh, M., Thiollière, R., Schatz, T., Cao, X. N., Anguera, X., Jansen, A. and Dupoux, E.: The Zero Resource Speech Challenge 2015, *Proc. Interspeech*, p. 3169 3173 (2015).

[2] Sutton, R. S., McAllester, D., Singh, S. and Mansour, Y.: Policy Gradient Methods for Reinforcement Learning with Function Approximation, *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pp. 1057–1063 (1999).

[3] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D. and Kavukcuoglu, K.: Asynchronous Methods for Deep Reinforcement Learning, *Proceedings of The 33rd International Conference on Machine Learning* (Balcan, M. F. and Weinberger, K. Q., eds.), Proceedings of Machine Learning Research, Vol. 48, PMLR, pp. 1928–1937 (2016).

[4] Shannon, M.: Optimizing Expected Word Error Rate via Sampling for Speech Recognition, *Proc. Interspeech 2017*, pp. 3537–3541 (2017).

[5] Zhou, Y., Xiong, C. and Socher, R.: Improving End-to-End Speech Recognition with Policy Learning, *Proc. ICASSP*, pp. 5819–5823 (2018).

[6] Tjandra, A., Sakti, S. and Nakamura, S.: Sequence-to-Sequence ASR Optimization via Reinforcement Learning, *Proc. ICASSP*, pp. 5829–5833 (2018).

[7] Kato, T. and Shinozaki, T.: Reinforcement Learning of Speech Recognition System Based on Policy Gradient and Hypothesis Selection, *Proc. ICASSP*, pp. 5759–5763 (2018).

[8] Graves, A. and Jaitly, N.: Towards End-To-End Speech Recognition with Recurrent Neural Networks, *International Conference on Machine Learning*, pp. 1764–1772 (2014).

[9] Wierstra, D., Schaul, T., Glasmachers, T., Sun, Y., Peters, J. and Schmidhuber, J.: Natural Evolution Strategies, *J. Mach. Learn. Res.*, Vol. 15, No. 1, pp. 949–980 (2014).

[10] Hansen, N., Müller, S. D. and Koumoutsakos, P.: Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES), *Evolutionary Computation*, Vol. 11, No. 1, pp. 1–18 (2003).

[11] Williams, R. J.: Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning, *Machine learning*, Vol. 8, No. 3-4, pp. 229–256 (1992).

[12] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to Sequence Learning with Neural Networks, *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 3104–3112 (2014).

[13] Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H. and Bengio, Y.: Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation, *EMNLP*, pp. 1724–1734 (2014).

[14] Bahdanau, D., Cho, K. and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, *arXiv preprint arXiv:1409.0473* (2014).

[15] Weaver, L. and Tao, N.: The Optimal Reward Baseline for Gradient-Based Reinforcement Learning, *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., pp. 538–545 (2001).

[16] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlıcek, P., Qian, Y., Schwarz, P., Silovskı, J., Stemmer, G. and Veselı, K.: The Kaldi Speech Recognition Toolkit, *Proc. ASRU* (2011).

[17] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y. and Zheng, X.: TensorFlow: A System for Large-Scale Machine Learning, *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, Berkeley, CA, USA, USENIX Association, pp. 265–283 (2016).

[18] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *arXiv preprint arXiv:1412.6980* (2014).

ⓒ 2018 Information Processing Society of Japan