Vol.2020-SLP-131 No.11
2020/2/14

二重相続進化戦略による End-to-End 音声認識システムの最適化

木村 友祐[1]　日野 健人[1]　董 越[1]　篠崎 隆宏[1]

概要：音声認識システムにおけるハイパーパラメタ最適化法として，進化戦略と知識蒸留を組み合わせる手法を提案する．従来の進化戦略ではエラー率の情報のみが次世代へ受け継がれるのに対し，知識蒸留を組み合わせることで世代間で伝えられる情報量を多くすることが出来る．この手法を End-to-End 型の音声認識システムに適用し，従来法より精度が高くコンパクトなモデルが作成できることを示す．

キーワード：End-to-End 音声認識システム 知識蒸留 進化戦略 二重相続理論 構造最適化

1. はじめに

従来の進化戦略による End-to-End 音声認識システムの最適化法では個体の評価スコアのみが次世代に伝えられる．本研究では生物学における二重相続理論をもとに，進化の過程に知識蒸留を組み込む手法を提案する．評価スコアとともに学習した知識を次世代へ継承することで従来法よりも効率的にコンパクトで高性能なモデルが探索できることを示す．

2. 進化戦略による音声認識システムの最適化

2.1 共分散行列適応進化戦略 (CMA-ES)

ハイパーパラメタの最適化手法の一種に共分散行列適応進化戦略 (CMA-ES) [1] [2] [3] [4] がある．この手法では最適化したいハイパーパラメタ群を遺伝子とし，遺伝子の確率分布を多変量ガウス分布で表す．

ガウス分布からサンプリングにより遺伝子集合を求め，各遺伝子に対応したハイパーパラメタで個体の評価スコアを計算し，ガウス分布を更新する．このプロセスを 1 世代とし，世代数を重ねることでパラメタの最適化を行う．

End-to-End 音声認識システムの最適化では，ニューラルネットワークの構造や学習条件などが最適化したいハイパーパラメタとなる．また，認識エラー率とともにネットワークサイズの小さいモデルが望ましいシステムであるため，本研究では進化の評価スコアを式 1 の目的関数により定義する．ここで CER は，文字エラー率, $Size$ はモデルサイズを表す．

$$Score = CER + \frac{Size}{Size_{init}} \tag{1}$$

2.2 提案手法

進化戦略を用いた最適化においては，遺伝子分布の更新を介して各個体の評価スコアのみが次世代へと伝えられる．音声認識システムなどの大規模なニューラルネットを進化により最適化を行う場合，各世代でニューラルネットは多くの計算コストをかけて学習を行い知識を学習するが，評価スコアのみが次世代へ伝えられ，その他の多くの学習した知識は次世代へ伝えられない．

我々は生物・社会学分野における二重相続理論を応用し，遺伝子と共にこの知識を次世代へ継承する二重相続進化戦略 (Dual Inheritance CMA-ES:DI-ES) 法を提案する．二重相続理論は人類が際立って高度な知識を有していることを説明する理論であり，親から子への文化の伝達と遺伝子の伝達が相互に知能の発達の促進に働いたとするものである．この提案手法により世代間で伝達する情報を増やし，また進化と共にその情報を有効に活用する個体が増加することが期待される．

知識蒸留の手法として複数の手法があるが本研究では図 1 に示す二種類の手法を組み込み，比較を行った．まず各世代において祖先世代から教師となる個体を選び，教師と生徒のエンコーダの出力同士の平均二乗誤差 $Loss_{MSE}$ と，デコーダの出力同士の交差エントロピー誤差 $Loss_{CE}$ を計算する [5].

これらの誤差を式 2 に示すようにもとの誤差 $Loss_{base}$ と

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

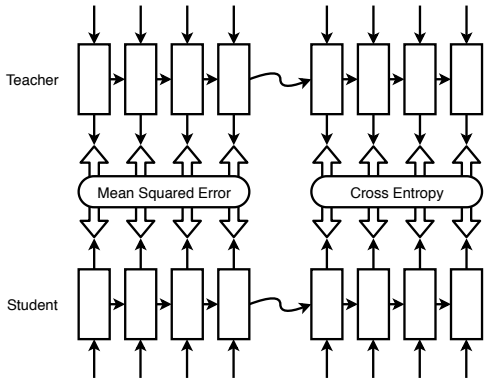

図 1 追加した損失関数

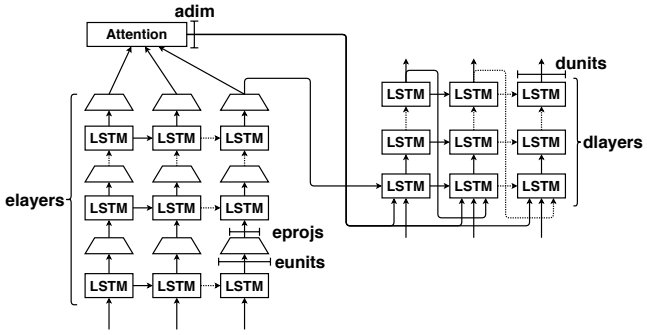

図 2 ESPnet の構造
CTC の部分は最適化対象としておらず，図では省略している．

の重み和 $Loss$ を損失関数として用いる．ここで $Loss_{KD}$ は $Loss_{MSE}$ または $Loss_{CE}$ を表す．

$$Loss = \mu \cdot Loss_{KD} + (1-\mu) \cdot Loss_{base} \quad (2)$$

この時の重み μ もハイパーパラメタであり，遺伝子に含めることで進化最適化の対象とする．

3. 実験条件

実験には ESPnet ツールキット [6] および an4 レシピを用いた．an4 の学習セットの話者は 74 名である．948 の発話が記録され，平均 3 秒，合計約 50 分の音声が記録されている．評価セットの話者は 10 名である．130 の発話が記録され，合計は約 6 分の音声が記録されている．

図 2 に使用した ESPnet の構造および進化最適化対象のパラメタを示す．最適化対象のハイパーパラメタにはモデル構造の他各種学習パラメタも含まれる．表 1 に最適化対象とした全ハイパーパラメタを示す．ただし，知識蒸留のグループのパラメタは従来法の CMA-ES では用いず，提案法でのみ用いる．

進化の起点となる初期個体には，an4 レシピの設定を用いた．提案法における教師個体は，全祖先世代の個体の中から初期個体よりもサイズの大きい個体のなかで認識エラー率の最も小さいものを選択した．第一世代の教師は初期個体とした．

表 1 最適化対象のハイパーパラメタ

種類	ハイパーパラメタ	初期値
一般	patience	3
	mtlalpha	0.5
エンコーダ	elayers	4
	eunits	320
	eprojs	320
デコーダ	dlayers	1
	dunits	300
Attentions	adim	320
	aconv-chans	10
	aconv-filts	100
知識蒸留	μ	0.3

表 2 全モデルサイズにおける最小エラー率

個体数	開発セット		
	CMA-ES	DI-ES($Loss_{MSE}$)	DI-ES($Loss_{CE}$)
15	11.4	12.6	12.5
25	12.3	12.5	12.0
50	12.1	11.3	12.2
個体数	評価セット		
	CMA-ES	DI-ES($Loss_{MSE}$)	DI-ES($Loss_{CE}$)
15	5.1	7.3	5.4
25	7.4	5.7	6.7
50	5.6	4.6	4.8

進化実験は CMA-ES と DI-ES のそれぞれについて，個体数 15,25,50 の 3 種類の条件で行った．各個体の最大 epoch 数は 40 であり，第 15 世代まで進化を進めた．

4. 実験結果

図 3, 4 に進化実験の結果を示す．図は各進化条件においてモデルサイズ (ニューラルネットの重みパラメタ数)S 以下において開発セットにおける最小の文字誤り率 CER を求めた結果を S を動かしながらプロットしたものである．いずれの進化手法・条件においても，ESPnet の初期個体よりも大幅に優れた結果が得られている．

従来法と提案法を比較すると，従来法は個体数を増やしても結果に大きな変化は見られないが，提案法では個体数を増やすと従来法よりサイズが小さくエラー率も低い個体が増加した．特に $Loss_{CE}$ を追加したものはサイズの小さな範囲でその傾向がみられる．

また，$Loss_{MSE}$ を追加したものは表 2 より最小エラー率が個体数とともに小さくなり，個体数 50 で開発セットにおける最小のエラー率となった．

今回の実験では，個体数が 15 の場合に従来法の方が提案法より優れた結果を示している．これは，従来法と提案法の最適化対象のパラメータ数の差が個体数 15 に対し比較的大きく，進化の過程に大きく影響を与えたためと考えられる．

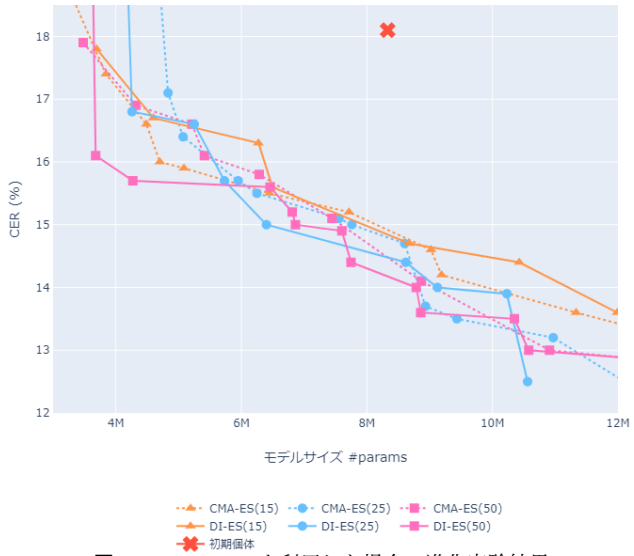

図 3　$Loss_{MSE}$ を利用した場合の進化実験結果

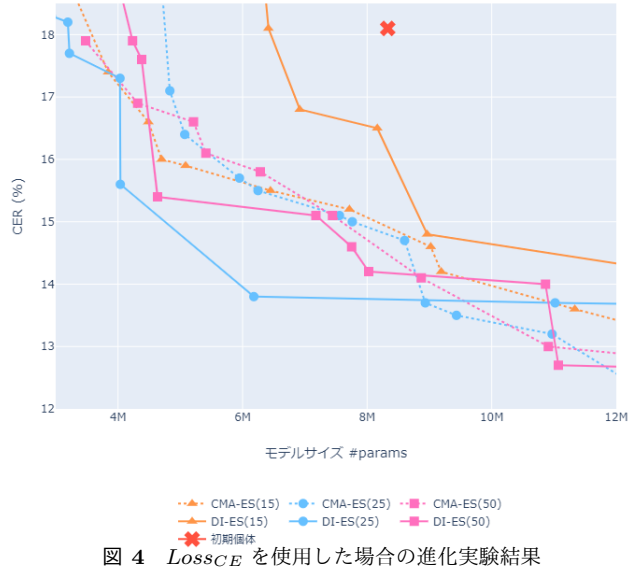

図 4　$Loss_{CE}$ を使用した場合の進化実験結果

5. まとめ

本研究では進化戦略法のニューラルネットへの適応において各個体の評価結果に加えて個体学習の結果に基づく知識を文化として子孫世代に伝達させる二重相続進化戦略法を提案し，音声認識システムを最適化対象として評価実験を行った．

提案法は，音声認識システム以外にもニューラルネットを用いた幅広いシステムに応用可能である．また，CMA-ES 以外の進化戦略法や，遺伝的アルゴリズムなどへの拡張も容易である．

従来法に対し $Loss_{CE}$ と $Loss_{MSE}$ を個別に追加し，それぞれにおいて従来法に対する優位性を示すことが出来た．以上から，$Loss_{CE}$ と $Loss_{MSE}$ を組み合わせることで，両方の特性を持った進化が可能になると考えられる．今後の課題としては，この組み合わせた手法の検証および，より詳しい評価実験を行うこと，教師個体の選択方法を工夫することなどが挙げられる．

参考文献

[1] N. Hansen, S. D. Mu ¨ ller, and P. Koumoutsakos, “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.

[2] Y. Akimoto, Y. Nagata, I. Ono, and S. Kobayashi, “Bidirectional relation between CMA evolution strategies and natural evolution strategies,” in Proc. Parallel Problem Solving from Nature (PPSN), 2010, pp. 154 - 163.

[3] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber, “Natural evolution strategies,” J. Mach. Learn. Res., vol. 15, no. 1, pp. 949 - 980, 2014.

[4] N. Hansen, A. Auger, R. Ros, S. Finck, and P. Po ˇ s ´ ık, “Comparing results of 31 algorithms from the black-box optimization benchmarking bbob-2009,” in Proc. the 12th annual conference companion on Genetic and evolutionary computation (GECCO), 2010, pp. 1689 - 1696.

[5] Adriana Romero, et al.”FitNets: Hints for Thin Deep Nets”, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015).

[6] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, “Espnet: End-to-end speech processing toolkit,”in Interspeech, 2018, pp. 2207 - 2211.