遺伝子文化共進化による深層ニューラルネット構成の自動最適化

代表研究者　　篠崎　隆宏　　東京工業大学　工学院　准教授

1 研究の目的

深層ニューラルネット(DNN)が性能を発揮するためには、モデル構造や学習条件などの設定を調整し最適化する必要がある。この調整は一種のブラックボックス最適化問題であり、最適化の自動化には進化最適化手法の応用が考えられる。一般に進化最適化では、最適化対象となるメタパラメタを遺伝子表現し、遺伝子に対して最適化を行う。進化的最適化手法の DNN への応用では、学習に２重構造が存在するという他の進化最適化タスクには無い独特な特徴がある。すなわち、進化手法において遺伝子の分布に対して最適化が行われると同時に、遺伝子により指定された設定の下に作成された DNN においてバックプロパゲーション法により学習データからネットワーク結合重みの学習が行われる。学習におけるこの２重構造の存在は、人類の進化においても同様である。遺伝子の最適化は種の進化であり、個々の脳による学習が個体学習である。そして進化生物学の分野では、遺伝子文化共進化理論が人類の突出した知性を説明する理論として提唱されている。本研究の目的は、DNN の効率的で高度な最適化手法を実現することを目標に遺伝子文化共進化理論のアイデアを工学的に DNN の最適化問題に応用する手法を提案するとともに、その可能性を調査することである。

2 はじめに

ディープニューラルネットワーク（DNN）ベースのシステムは、従来のシステムの性能を大幅に向上させるとともに、従来は技術的に困難であった様々な人工知能タスクを実現しつつある。しかし、DNN の能力を最大限に発揮させるためには、ネットワーク構造や学習条件などのメタパラメタ調整が不可欠である。DNN の膨大な数のネットワーク結合重みパラメタはバックプロパゲーション法により学習データから効率的に推定することが可能であるが、ネットワーク構造や学習条件の解析的な最適化は不可能である。そのため、現状では人間の専門家が、試行錯誤に基づいて多大な労力を費やしながら DNN のメタパラメタ最適化を行っている。多くの場合において、メタパラメタの最適化が DNN を用いたシステムの開発において最も時間を必要とするプロセスであるとともに、システムの性能が職人的な最適化のスキルに依存してしまっている。

このような背景のもと、DNN のメタパラメタ最適化をベイズ的最適化や進化的アルゴリズムのようなブラックボックス最適化手法を応用することで自動化する研究も行われている[1]。これらの手法により、ランダムな初期モデルや人手により最適化された初期モデルよりも高い性能のモデルが得られることが示されている。しかしこれらの手法では、個体としての DNN の学習結果は適応度としての性能評価値を除き次世代には引き継がれず全て廃棄されてしまっている。DNN の学習と評価は多くの計算を費やして行われるものであり、明らかに非効率的である。

生物との類似を考えると、DNN 構成の最適化は遺伝子の進化による脳のデザインの指定に対応し、DNN の重みパラメタの最適化は脳を持つ個人による学習に対応している。人類の進化・発展では、遺伝子の進化に加えて、社会の中での個人の学習に基づく文化の進化が存在している。進化生物学の分野における遺伝子文化共進化理論[2]では文化の進化と遺伝子の進化の間に図１に示すように相互作用があることを指摘しており、人の卓越した知性が遺伝子進化と文

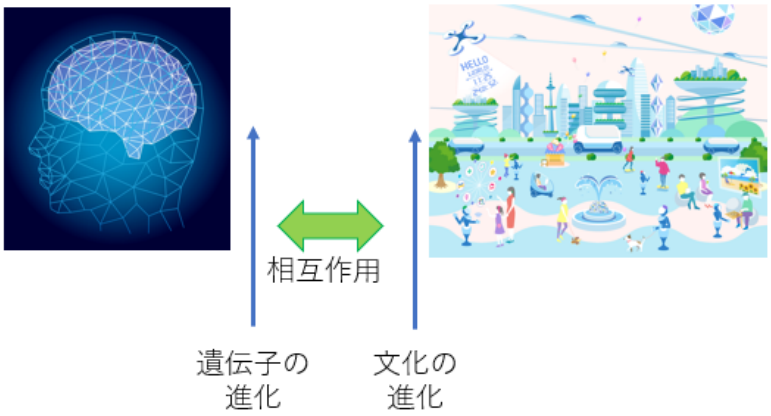

図 1 人類における遺伝子と文化の共進化

公益財団法人電気通信普及財団
研究調査助成報告書　第35号　2020年度

化進化の相乗効果の結果であると説明している。

本研究では、遺伝子文化共進化理論のアイデアをもとに従来の進化戦略最適化法を拡張する二重相続進化戦略（DI-ES）を提案する。提案法は各世代の個体の適合度スコアに基づいて遺伝子分布を更新することに加えて、祖先 DNN から子孫 DNN への付加的な情報伝達経路を導入する。提案法は DNN 一般に対して適用可能であるが、本研究では End-to-End 型の音声認識システムを最適化対象のシステムとして実験を行う。またベースとする進化手法としては共分散行列適応進化戦略（CMA-ES）を用いる。

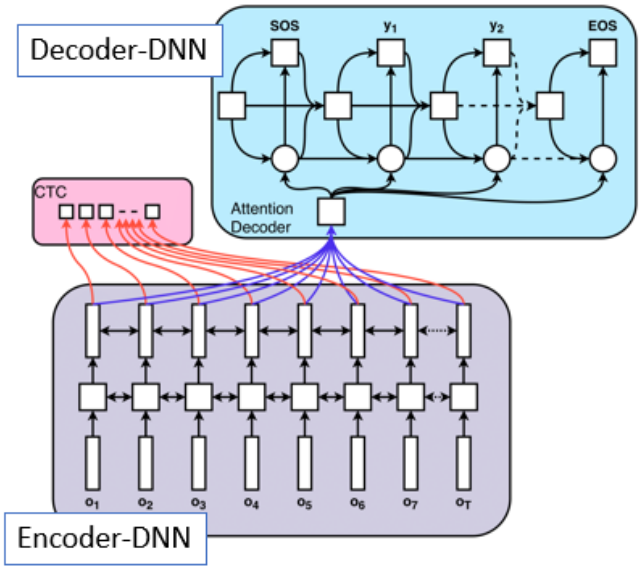

図 2 End-to-End 型音声認識システム

3 音声認識システム

3-1 認識の基本原理

現在の音声認識システムは統計的なモデル化に基づいている。典型的なシステムではまず入力音声信号に対してオーバーラップさせた短い時間窓毎に周波数分析を行う。分析窓幅は周波数パターンの時間変化を捉えるように音素の継続時間よりも短くかつ十分な周波数分解能を得られるように 25ms 程度にとることが多い。各分析窓においてフーリエ変換を行い、振幅スペクトルを求める。振幅スペクトルに対してさらに各種の処理を適用して音響特徴量を得る。振幅スペクトルを直接音響特徴量として用いる場合もある。いずれの場合も音響特徴量はベクトルとして表現される。分析窓を 10ms ずつシフトする場合、1 秒あたり 100 個の割合の音響特徴量ベクトルの時系列が得られる。

入力音声に対応した長さTの音響特徴量ベクトルの時系列を$\boldsymbol{O} = \langle o_1, o_2, \cdots, o_T \rangle$、長さ$N$の単語列またはテキストを$\boldsymbol{W} = \langle w_1, w_2, \cdots, w_N \rangle$とする。このとき、音声認識は式(1)に示すように$P(W|O)$を最大にする$\widehat{\boldsymbol{W}}$の探索として定式化される。作用素 argmax は、最大値を与える引数を返す演算である。

$$\widehat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}} \boldsymbol{P}(\boldsymbol{W}|\boldsymbol{O}). \tag{1}$$

条件付確率$\boldsymbol{P}(\boldsymbol{W}|\boldsymbol{O})$を計算機上で数値評価するためには、真の確率分布を近似する音声モデル$\boldsymbol{P}_{\boldsymbol{\theta}}(\boldsymbol{W}|\boldsymbol{O})$が必要である。ここで$\boldsymbol{\theta}$は確率モデルのパラメタ集合である。一発話の長さやその中に含まれる単語数は様々である。このため確率変数$\boldsymbol{O}$および$\boldsymbol{W}$は可変長の時系列であり、それら可変長時系列に対する条件付確率のモデル化が必要となる。また、$\boldsymbol{W}$の種類数は長さ$\boldsymbol{N}$に対して指数的に増加するので、探索においては効率的に計算を行う工夫も必要となる。

生成モデルのアプローチでは、ベイズの定理 $\boldsymbol{P}(\boldsymbol{W}|\boldsymbol{O}) = \frac{\boldsymbol{P}(\boldsymbol{O}|\boldsymbol{W})\boldsymbol{P}(\boldsymbol{W})}{\boldsymbol{P}(\boldsymbol{O})}$ および$\boldsymbol{P}(\boldsymbol{O})$が$\boldsymbol{W}$の最大化に対して定数であることを利用して式(2)に示す変形を行う。この方法は、DNN を用いたモデル化が一般化する前には音声認識技術の主流であった。

$$\underset{W}{\operatorname{argmax}}\, P(W|O) = \underset{W}{\operatorname{argmax}}\, P(O|W)P(W). \tag{2}$$

他方条件付確率$P(\boldsymbol{W}|\boldsymbol{O})$を直接モデル化する識別モデルのアプローチは、$\boldsymbol{O}$および$\boldsymbol{W}$が時系列データであることから、最近まで困難であった。しかし、深層学習の進展とともに$P(\boldsymbol{W}|\boldsymbol{O})$を直接モデル化するアプローチが End-to-End 音声認識として急速に発展し、実用されるようになった。図 1 の例に示すように、End-to-End 音声認識システムは全体が一つのニューラルネットで構成される。このため、同様にニューラルネットで実装された対話システムや画像処理との一体化など、様々な拡張が容易である。この利点は大きく、従来は不可能であった高度で柔軟な音声情報処理の研究が、言語処理や画像処理、信号処理など近隣分野と融合しながら展開されつつある。

3-2 モデル学習

音声モデル$P_{\boldsymbol{\theta}}(\boldsymbol{W}|\boldsymbol{O})$を用いて音声認識を行うためには、事前に$\boldsymbol{\theta}$中のパラメタを推定する必要がある。パラメタの推定は計算機中に認識対象の言語の知識を蓄えることであり、人が生まれた後に言語を学習することに対応する。現在一般に行われているのは、音声とその書き起こしテキストのペア$\langle \boldsymbol{O},\boldsymbol{W} \rangle$を大量に使用する教師あり学習である。

4 進化最適化

ブラックボックス関数を最適化する進化最適化手法としては、遺伝的アルゴリズムや進化戦略などがある。最適化対象のブラックボックス関数はいずれの手法においても入力に対して出力が求まりさえすればよく、勾配などは用いない。

4-1 遺伝的アルゴリズム

遺伝的アルゴリズムは、自然界での生物の進化をそのまま模倣したアルゴリズムである。遺伝的アルゴリズムでは解の候補を整数その他数値の並びで表現したものを遺伝子とみなし、遺伝子に対応した個体の評価を元に遺伝子集合を逐次的に更新する。遺伝子集合の更新は世代交代になぞらえられる。世代交代の方法には多くのバリエーションが存在するが、1)優良個体の選択、2)複数の親からの遺伝子の組み換え、3)遺伝子の突然変異、の組み合わせを基本とする。すなわち生存競争を生き抜いた個体が親となり子孫を残す仕組みである。優良個体の選択は適合度に基づいた選択圧をかけるためであるが、遺伝子プールに多様性を確保し局所最適解に陥ることを避けるためには最優秀ではない個体にもいくらかのチャンスを与えることが重要となる。また突然変異の確率は小さすぎると進化が遅く、逆に高すぎると遺伝子が無秩序化してしまうので適切な値にする必要がある。

遺伝的アルゴリズムは他の手法と比較して必ずしも進化効率は高くなく、また進化効率を高めようとすると直感や経験に頼った様々な工夫を行うことになる。一方遺伝子のサイズに対する制約は少なく、非常に大規模な遺伝子を扱うことも可能である。実際生物が遺伝的アルゴリズムにより進化したと捉えれば、遺伝子配列のサイズは例えば人でおよそ 30 億と巨大である。

4-2 進化戦略

進化戦略は遺伝的アルゴリズムと類似の手法であるが、歴史的に異なる経緯を持ち、遺伝子を実数ベクトルで表現するのが特徴である。中でも CMA-ES[3,4]は様々なタスクで効果的であることが示され多く用いられている手法である。CMA-ES と類似した手法として、自然進化戦略[5]がある。それぞれの手法には複数のバリエーションがあるものの、その後の研究で両者の定式化の中心的な部分は等価であることが示されている[6]。

遺伝的アルゴリズムが遺伝子の分布を直接個体の集合で表現するのに対して、CMA-ES では遺伝子が固定長の実数ベクトルであることを利用しガウス分布により表現する。具体的な解候補としての遺伝子は、ガウス分布からサンプルすることにより得る。CMA-ES ではブラックボックス関数の直接的な最適化に代えて、ガウス分布に基づくブラックボックス関数の値の期待値である期待適応度を目的関数としてガウス分布の平均ベクトルと共分散行列を最適化する。期待適応度を最大化することはガウス分布の確率密度が性能の良い遺伝子が集まる領域に集中するように推定することに対応し、そこからサンプルした遺伝子は高い確率で性能の良いものであると期待できることになる。

期待適応度の最大化には勾配法を用いることが考えられるが、そのためにはブラックボックス関数をブラックボックス関数として扱いながら期待適応度の勾配を評価する必要がある。このために log トリックとよばれる対数関数の微分の性質と期待値のサンプル近似を用いる。勾配評価のためのこの手法は、勾配方策法による強化学習手法で用いられているのと同じものである。そのため CMA-ES は 1 状態の強化学習とみなすこともでき、その場合はガウス分布が方策関数に対応し、遺伝子がアクションに対応する。進化計算としての観点からは、サンプル近似による勾配評価のためのサンプル集合が一つの世代の個体集合に相当し、勾配法の更新ステップが世代交代に対応する。

期待適応度の勾配をサンプル近似により推定することで、ブラックボックス最適化を行うという目的は達

成できる。しかし、この定式化では安定した最適化を行うには大きな繰り返し数が必要になるなどの欠点がある。CMA-ES のもう一つのポイントは、通常の勾配の代わりに自然勾配を用いることでできるだけ少ないサンプルの評価回数で効果的に最適化を行うことである。さらに実際の CMA-ES の応用では特定の評価尺度に対する依存性を下げるために、ブラックボックス関数の出力をそのまま使用するのではなく

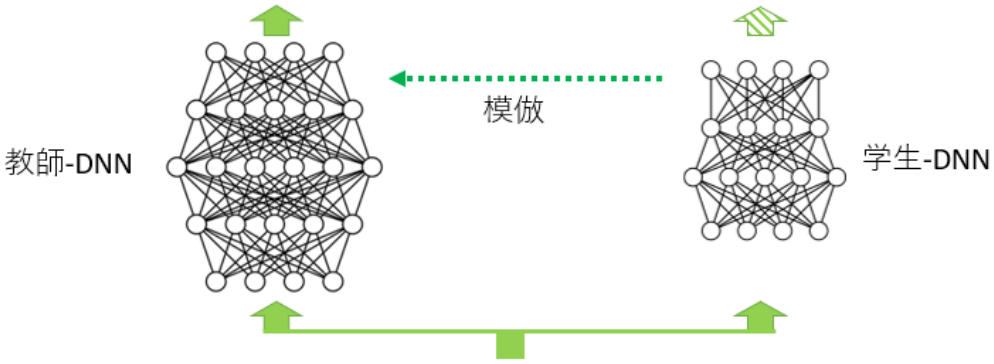

図 3　教師-学生学習

順位のみに依存する重みに直してから用いられることが多い。

CMA-ES では全共分散を使うため，遺伝子ベクトルの次元数の 2 乗に比例してガウス分布のパラメタが増加する。このため何らかの工夫をしない限り、大きな次元数(例えば 50 次元以上)の遺伝子の取扱いは難しくなる。

5 教師-学生（TS）学習

教師-学生（TS）学習 [7, 8, 9]は、図３に示すように教師役となる高性能な DNN からコンパクトで計算量の少ない学生 DNN に知識を伝達させることで、効率的で高性能な DNN を学習するための手法として広く用いられている。コンパクトな DNN を直接学習するよりも、計算量を多く必要とする大規模で高性能な DNN を先ず学習しそれを模倣するようにコンパクトな DNN を学習することで、経験的に多くの場合でより高い性能が得られることが知られている。ただし、教師 DNN からモデルサイズを削減しながら高い性能を達成する学生 DNN の設計に指針はなく、ここでも試行錯誤が必要となる。

6 提案手法

提案法は、図４に示すようにベースとなる既存進化最適化手法と、ニューラルネットの学習結果を親世代のニューラルネットから子世代のニューラルネットへと直接伝搬させる仕組みから構成される。このうち既存進化最適化手法としては、共分散行列適応進化戦略手法(CAM-ES)を用いた。最適化対象となるメタパラメタ集合を実数ベクトルで表現し、その分布を多変量ガウス分布でモデル化する。遺伝子が決まればメタパラメ

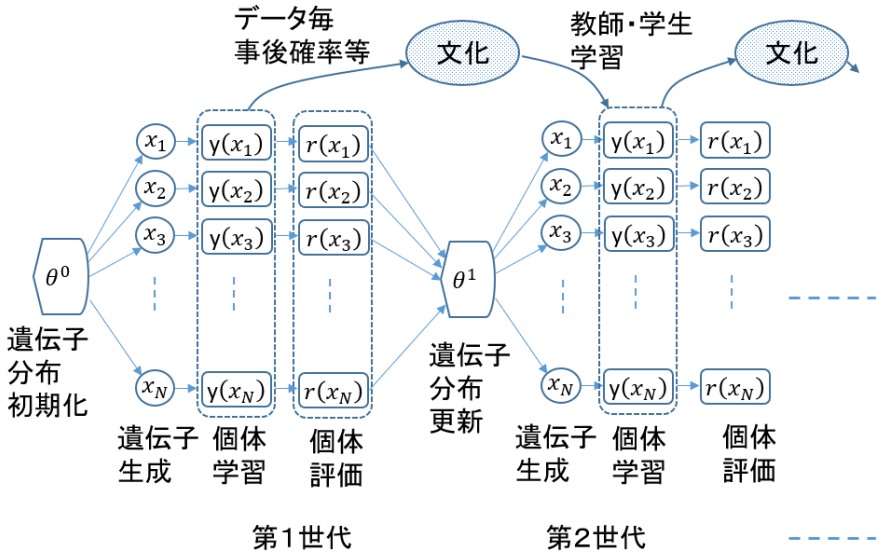

図 4　提案二重相続進化戦略（DI-ES）法

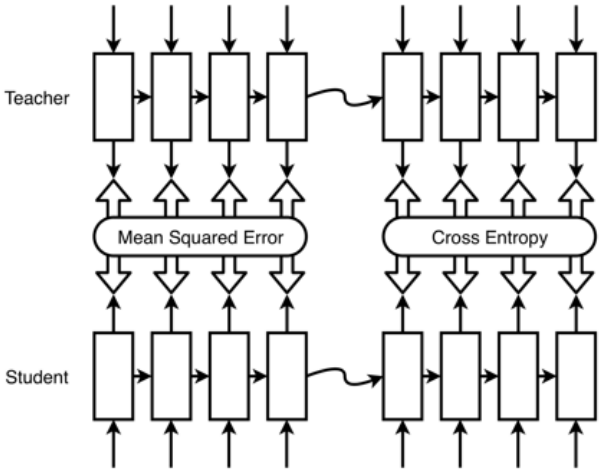

図 5　教師学生学習の実装

表 1 遺伝子の定義と初期設定

Type	Meta-parameters	Initial value
Learning	patience	3
	mtlalpha	0.5
Encoder	elayers	4
	eunits	320
	eprojs	320
Decoder	dlayer	1
	dunits	300
Attention	adim	320
	aconv-chans	10
	aconv-filts	100
TS learning	μ (TS weight)	0.3
	λ (End/Dec balance)	0.5
	T_T	20

タの値が決まり、それに従って DNN を学習・評価することで遺伝子を評価する。

ニューラルネットの学習結果を次世代へと伝搬する仕組みには、知識蒸留法を応用した。知識蒸留は、コンパクトで高性能なニューラルネットを学習するための手法として広く用いられている。計算量が多くても高い性能を実現できるニューラルネットを教師、少ない計算量で実現可能なニューラルネットを学生とし、学生が教師をまねるように学習させることが知識蒸留の基本的なアイデアである。具体的な実装としては、教師ネットワークの出力を学生ネットワークの学習時にソフトターゲットとして用いる方法や、教師ネットワークの隠れ層の出力を真似るように学生ネットワークの隠れ層を学習する方法などがある。出力層のユニット数は対象タスクのカテゴリ数で決まるが、隠れ層のユニット数は任意である。そのため、後者の場合は教師ネットワークと学生ネットワークで隠れ層のユニット数が異なってもよいように教師ネットワークの隠れ層出力にアダプタとなる変換層を適用する。本研究では、図 5 に示すように両方の手法及びその組み合わせを実装し、実験に用いた。各世代の学習において祖先世代から性能の良い個体を選抜し、それを教師として現世代の各個体の学習を行う。

提案アルゴリズムでは、祖先からの知識を有効に活用できる個体が有利となる。またそれにより高い性能を実現する個体が教師モデルとして選抜されることにより、より高い性能を実現する個体の出現が期待される。知識蒸留では教師ネットワークからの知識と学習ラベルからの知識のバランスをとるための重みを設定する必要があるが、提案法はその重みも最適化対象として遺伝子に含めることができる。

7 実験条件

評価実験は、End-to-End 型の音声認識システムを最適化対象として用いて行った。音声認識システムの実装には ESPnet ツールキット[10]を用い、an4 コーパスを認識評価タスクとして用いた。an4 の学習セットに含まれる話者数は 74 名であり、948 の発話が収録されている。発話長は平均 3 秒である。評価セットの話者は 10 名であり、130 の発話が含まれている。進化の起点となる初期個体には、ESPnet ツールキットで用いられている an4 認識システムの初期値を用いた。表 4 に遺伝子の定義と初期設定を示す。進化実験は CMA-ES と提案法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。進化の目的関数としては、文字誤り率とモデルサイズの重み和を用いた。

8 実験結果

図６に、ベースライン手法として従来の CMA-ES を用いた場合の進化の様子を示す。横軸が DNN のモデルパラメタ数、縦軸が音声認識における文字誤り率を示している。図では左下に行くほどコンパクトなモデルで高い認識性能を実現していることを意味している。

図７に、CMA-ES 法と提案する DI-ES 法を用いて 15 世代にわたる進化を行った後の、手法毎のパレートフロンティアを示す。個体数は 50 である。従来法の CMA-ES を用いた場合と提案法の遺伝子文化共進化法を用いた場合のどちらの場合でも、初期個体より大きく改善した性能を持つ個体を得ることができた。また CMA-ES を用いた場合と比較して提案法を用いた場合は進化性能が改善し、従来法よりも優れた個体を得ることができた。

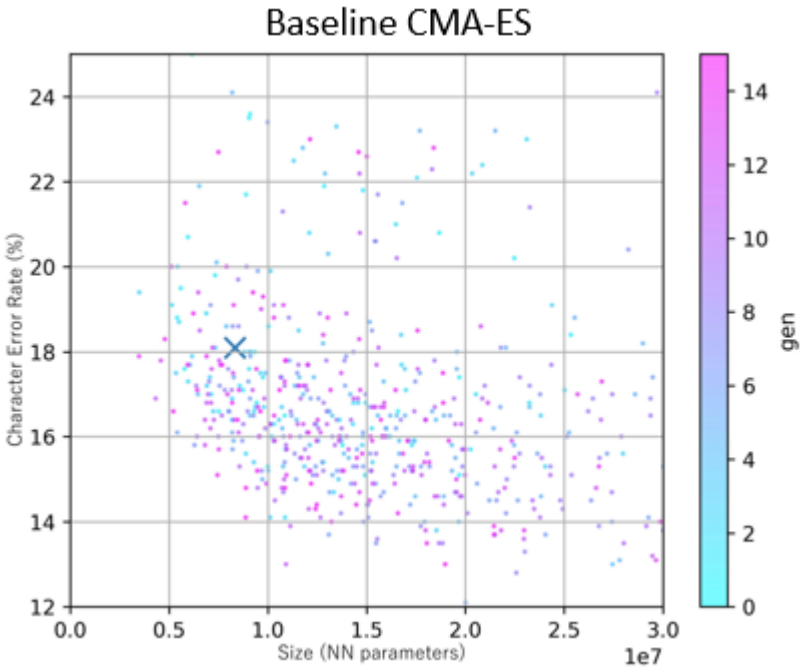

図 6 ベースライン CMA-ES 法による進化の様子

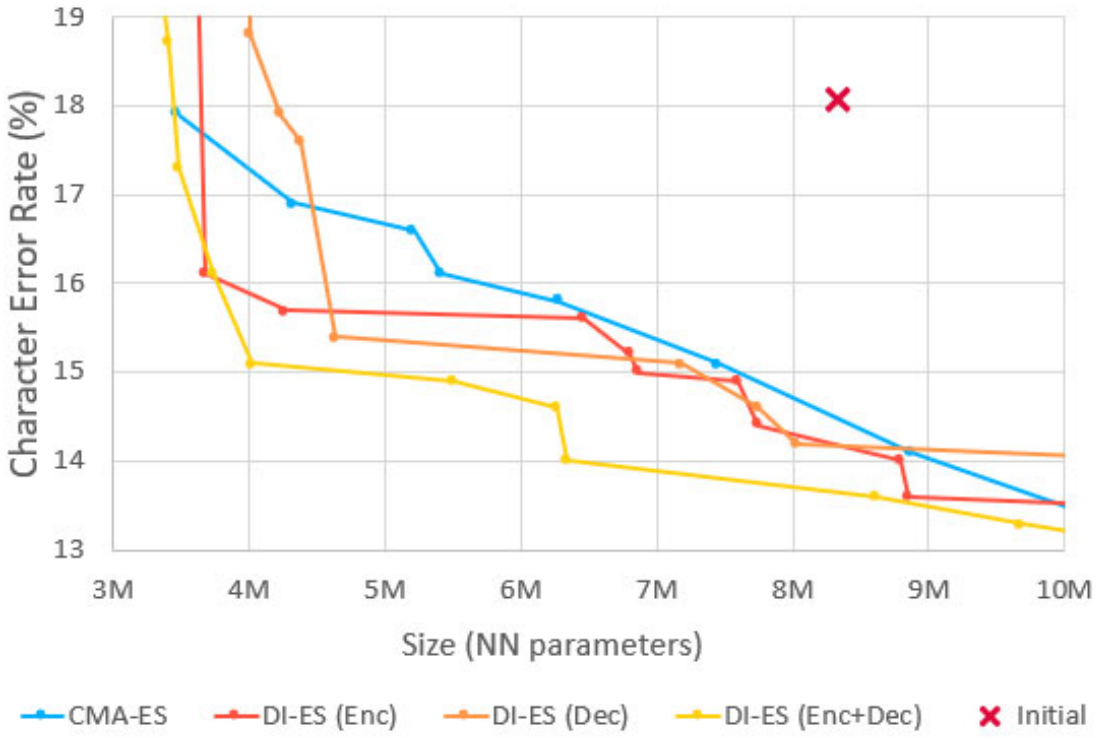

図 7　従来 CMA-ES 法と提案 DI-ES 法の比較

9 まとめと課題

今後の課題としては、より詳しい評価実験を行うことや教師個体の選択方法を工夫することなどが挙げられる。また提案法は、従来進化生物学における仮説でしかなかった遺伝子文化共進化理論に検証可能な実験モデルを与える可能性を持つものである。このため、工学および進化生物学の双方の観点から、世代間でより多様な情報を伝達させる手法の実現に取り組むことも有用と考えられる。

【参考文献】

[1] Takafumi Moriya, Tomohiro Tanaka, Takahiro Shinozaki, Shinji Watanabe, Kevin Duh, "Evolution-Strategy-Based Automation of System Development for High-Performance Speech Recognition," IEEE Transactions on Audio, Speech and Language Processing, Vol.27, No.1, pp77-88, 2019-1.

[2] J. Henrich and R. McElreath, “Dual-inheritance theory: the evolution of human cultural capacities and cultural evolution,” in Oxford handbook of evolutionary psychology, 2007.

[3] N. Hansen, S. D. Müller and P. Koumoutsakos, "Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES)," in Evolutionary Computation, vol. 11, no. 1, pp. 1-18, March 2003

[4] Nikolaus Hansen, Anne Auger, Raymond Ros, Steffen Finck, and Petr Pošík. 2010. Comparing results of 31 algorithms from the black-box optimization benchmarking BBOB-2009. In Proceedings of the 12th annual conference companion on Genetic and evolutionary computation (GECCO ’10)

[5] D. Wierstra, T. Schaul, J. Peters and J. Schmidhuber, "Natural Evolution Strategies," *2008 IEEE* Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. 3381-3387

[6] Youhei Akimoto, Yuichi Nagata, Isao Ono, and Shigenobu Kobayashi. 2010. Bidirectional relation between CMA evolution strategies and natural evolution strategies. In Proceedings of the 11th international conference on Parallel problem solving from nature: Part I (PPSN’10)

[7] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD ’06). Association for Computing Machinery, New York, NY, USA, 535–541.

[8] Geoffrey Hinton, Oriol Vinyals and Jeff Dean. Distilling the Knowledge in a Neural Network. arxiv:1503.02531

[9] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “Fitnets: Hints for thin deep nets,” in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[10] 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," Proc. Interspeech'18, pp. 2207-2211 (2018)

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
Dual Inheritance Evolution Strategy for Deep Neural Network Optimization	Proc. IEEE Congress on Evolution Computation (CEC) 2020	2020.7 (accepted)
二重相続進化戦略による音声認識システムの最適化	日本音響学会 2020 年春季研究発表会講演論文集	2020.3
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	音声言語情報処理研究会(SLP)	2020.2