jh230045

機械学習ソフトウェアへのソフトウェア自動チューニング技術の適用（２）

田中輝雄（工学院大学）

概要

ソフトウェア自動チューニング(AT)におけるユーザプログラムの性能を決定する性能パラメタの最適組合せ探索に関する研究である．我々は，膨大な学習時間を要する機械学習ソフトウェアのハイパーパラメタの最適化に対し，スーパーコンピュータの多数 GPU を有効に用いた多重実行を制御する AT ツールの開発を進めている．単なる並列化では，アルゴリズムの特性により，多重実行の並列度を十分使えることができず，スーパーコンピュータの能力をフルに引き出せなかった．そのため，スーパーコンピュータの運用上の多重度に合わせた数のジョブを実行するようにアルゴリズムの仕様を変更することで，スーパーコンピュータの稼働率をほぼ 100%まで引き上げることができた．そのときのアルゴリズムの仕様の変更による性能の低下は評価したアプリケーションではほとんどないことを確認した。さらに，名古屋大学/不老で開発した AT ツールの並列制御環境をほぼそのまま九州大学/ITO に移植し動作可能なことを確認し，開発ツールの移植性の容易さを確かめた．

1. 共同研究に関する情報

(1) 共同利用・共同研究を実施している拠点名

東京大学 情報基盤センター
名古屋大学 情報基盤センター
九州大学 情報基盤研究開発センター

(2) 課題分野

大規模計算科学課題分野

(3) 共同研究分野

超大規模数値計算系応用分野

(4) 参加研究者の役割分担

田中 輝雄（工学院大学・情報学部）
・研究統括，自動チューニング手法
大島 聡史（名古屋大学・情報基盤センタ）
・GPU・メニコア向けアルゴリズム
藤井 昭宏（工学院大学・情報学部）
・並列数値計算アルゴリズム
矢島 雄河(工学院大学・大学院情報学専攻)
・自動チューニングソフトウェア開発，評価
加藤 由花（東京女子大学・現代教養学部）
・機械学習プログラム・データ提供，検証
浮田 宗伯(工学院大学・大学院情報学専攻)
・機械学習プログラム・データ提供，検証
片桐 孝洋(工学院大学・大学院情報学専攻)
・自動チューニング手法,GPU 向けアルゴリズム

2. 研究の目的と意義

我々は，多様な計算機環境での高性能化を実現する技術として，対象とするプログラム性能を決定する複数の性能パラメタを，キャッシュサイズやデータ通信性能等の計算機の特性に，自動的に最適化するソフトウェア自動チューニング(以下，AT と記載）の研究を進め，その手法を提案，実証している[文献 a].

現在，機械学習プログラムをターゲットに研究を進めており，その機械学習プログラムの

ハイパーパラメタに対して AT を実施し，予測モデルの精度向上および実行時間の短縮を図る．機械学習では１回の学習に 30 分から数時間かかる．ハイパーパラメタの組合せパタンは数千から数万通りであり，我々の開発した AT 機構でもチューニングに数日かかる．これに対し，並列化により，ハイパーパラメタの組合せ探索時間の短縮を目指す．

2022 年度までに，スーパーコンピュータの持つ複数 GPU を多重実行させ，探索時間の短縮を実現した．

2023 年度は以下を目的として，さらに研究を進める．

（目的 1）初期パラメタの探索

探索の初期パラメタの組合せは，それまでのデフォルト値やそれぞれのパラメタの取り得る値の中央値などを用いていた．これを大域的なランダム探索により設定する．

（目的 2）並列化効率の向上

並列化効率を高めるために，センター運用上の稼働状況により動的に決まる同時実行可能なジョブ数に，アルゴリズムを変更することで，プログラム上で同時に実行する多重ジョブ数を合わせる．

（目的 3）他 AT ツールとの比較

機械学習で用いられている他のチューニングツールとの比較を行う．

（目的 4）移行性の確認

開発している AT ツールは，名古屋大学の不老 Type II で開発を行ったが，不老 Type II のシステム状況に特化しているわけではない．それを示すために，東京大学情報基盤センターの Wisteria/BDEC-01(Odyssey)，九州大学情報基盤研究開発センターの ITO など GPU を持つスーパーコンピュータ上に移植を行い，開発した AT ツールを稼働させる．

今年度は各目的に対し検討するための対象アプリケーションとして，東京女子大学の加藤グループによる「人を回避しながら動くロボットの制御」に向けた人移動予測 AI(以降，歩行者経路予測アプリ)を用いる．対象とするアプリは Physical world と Cyber World(クラウド)から構成される．Physical world ではロボットが予測モデルを使い周囲の人移動を予測し動作する．Physical world で得たセンシング情報を Cyber World に送り，Cyber World では予測モデルを機械学習により随時更新する．

本研究は，スーパーコンピュータを効果的に利用することにより，応用研究の専門家が時間をかけて行ってきた(学習)モデルの最適化を自動化し，応用研究者がモデル自体の研究開発に注力できるようにすることに意義がある．

[a]T.Tanaka, R.Otsuka, A.Fujii, T.Katagiri, T.Imamura, Implementation of d-Spline-based Incremental Performance Parameter Estimation Method with ppOpen-AT, Scientific Programming, Vol.22, pp.299-307, 2014.

[b]R.Akabane and Y.Kato, Pedestrian Trajectory Prediction Using Pre-trained Machine Learning Model for Human-Following Mobile Robot, IEEE International Conference on Big Data Workshop (IoTDA 2020), pp.3453-3458, 2020.

[c]M.Haris, G.Shakhnarovich, N.Ukita, Deep Back-Projection Networks for Single Image Super-Resolution, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.43, no.12, pp.4323-4337, 2021.

3. 当拠点の公募型共同研究として実施した意義

本研究課題で対象とする機械学習は膨大な学習時間を要する．AT では，この計算量に対してハイパーパラメタを変更しながら，応用プログラムを何度も実行する必要がある．また，機械学習においては，その計算の特性から GPU が有効である．そのために，多数の GPU を有するスーパーコンピュータである名古屋大学情報基盤センターの不老 Type II(以下，不老 Type II)を用いる．本研究では，この不老 Type II の持つ大規模 GPU 並列環境を有効に利用する手段

を提案し，実証する．また，開発した AT ツールの移行性を示すために，東京大学情報基盤センターの Wisteria/BDEC-01(Odyssey)，九州大学情報基盤研究開発センターの ITO など GPU を持つスーパーコンピュータを用いる．

さらに，本研究では応用研究の専門家（ここでは，ロボット工学者あるいは高解像画像処理技術者）と高性能計算学者との学際研究として推進し，HPC 及びスーパーコンピュータの利用技術を実証することで，広く開発技術を提供，普及させていく．

4. 前年度までに得られた研究成果の概要

我々の開発する AT ツールを機械学習のパイパーパラメタ選択のために並列化した探索の概要を図１に示す．まず，初期ハイパーパラメタの値の組合せを指定し，それを中心に AT ツールは複数の方向の直線上のパイパーパラメタの値の組合せを探索する．それぞれに対する機械学習プログラムを実行し，AT ツールは次に実行すべき複数のパイパーパラメタの値の組合せを決定する．これを最良の値の組合せと判別するまで繰り返す．このパイパーパラメタの値の組合せごとの機械学習プログラムの複数同時実行をスーパーコンピュータの多重 CPU にマッピングし，いわゆるパラメトリックスタディとしての並列処理を行う．図２に名古屋大学不老 Type II を用いた実験例の稼働状況を示す．

ここで，用いたアプリケーションは豊田工業大学の浮田グループによる機械学習を用いた超解像プログラムである[文献 c]．図２に名古屋大学 不老 Type II を用いた実験例の稼働状況を示す．このハイパーパラメタの組合せは 13310 通り(=10x11x11x11)であり，AT のよる機械学習プログラムの実行回数は 363 回(全パタンの 2.7%)となり実行時間は 110 時間を要した．平均的な体重ジョブ実行の並列度は 25 程度となった．

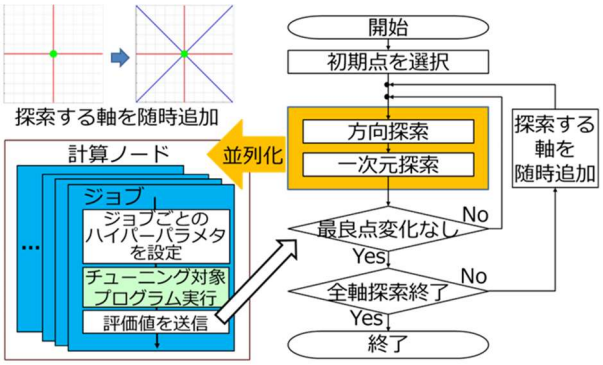

図１ ATによる探索の概要と並列化

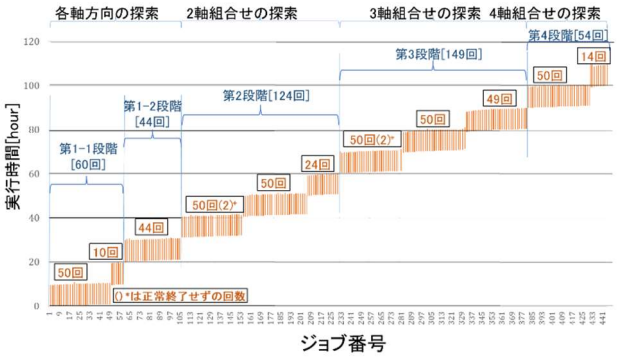

図2 並列実行状況

5. 今年度の研究成果の詳細

前述した今年度の各目的に対しての研究成果を示す．今回，適用するアプリは，歩行者経路予測アプリでの予測モデルに対する機械学習時のハイパーパラメタチューニングである．経路予測のためのデータは，“ETH Walking Pedestrians Dataset”から“Hotel”のデータを用いる．また， ハイパーパラメタは，表１に示す５種類で，それぞれ５通りの値を取り得るとし，5^5=3125 パタンから推定する．

表1 ハイパーパラメター覧

ハイパーパラメタ	特徴
Rnn size	LSTM (Long-Short Term Memory) において短期記憶の役割を果たす hidden state の大きさ
Grad clip	勾配が大きくならないように修正するための閾値
Learning rate	1回の学習でニューラルネットワーク内の重みやバイアスを更新する量の調整値
Dropout	過学習を抑えるために，学習時に特定のレイヤーの出力を0に落とす割合
Lamba	過学習を防ぐためのL2正則化における正則化パラメタ

まず，（目的 1）初期パラメタの探索および（目的 2） 並列化効率の向上に対する研究成果について説明する．表２に，(a)比較のベースとして昨年 2022 年度に開発した手法の並列版，

(b)目的１の初期点探索追加版，(c)目的２の今年度の改良版の結果について整理する．また，(a)の並列実行状況を図３に，同じく(b)および(c)の並列実行状況を図４および図５に示す．

表2 実行結果のまとめ

	2022年度 並列版	初期点探索 追加版	改良版
総ジョブ数	227	246	285
総実行時間（時間）	3h55m	3h23m	2h09m
性能FDE（m）	1.10	1.07	1.12
実行 ブロック数	10.3	8.5	3.5
1 実行ブロックあたりの平均並列数	22.9	28.7	60

※それぞれ5回の実行の平均値
※もとのユーザプログラムでのFDE値は1.85[m]

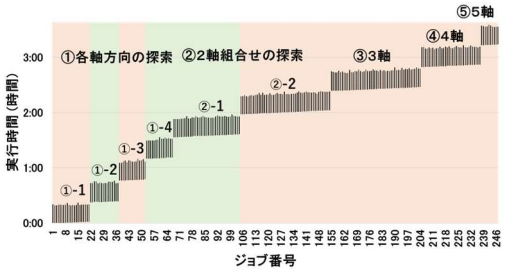

図3 2022年度に開発した手法による並列実行状況

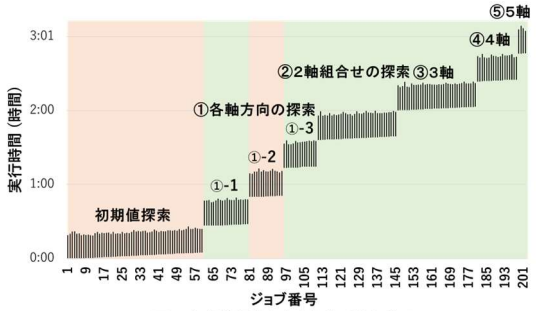

図4 初期値探索を含めた並列実行状況

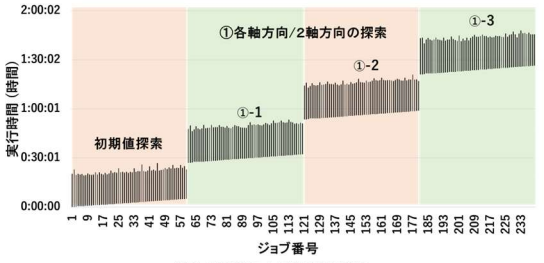

図5 改良版による並列実行状況

図３を用いて，図３から図５の図の見方を説明する．横軸は，ハイパーパラメタの組合せを変えて実行された各ジョブを１からナンバーリングしている．縦軸は実行時間である．各棒がひとつずつのジョブの実行状況を示している．

ここでは，５つのハイパーパラメタの取り得る値からなる５次元パラメタ空間 において，

まずは各軸と平行な直線上を探索する．その直線上の最小推定値の場所が変化している間は，各軸と平行となる方向を調べる．最小推定値の場所が変化しなくなれば，２軸の性能パラメタを組み合わせた斜め方向も探索範囲に追加する．これを随時組み合わせる軸の数を増やしていく．ここでは，１軸方向のみのとき４度最小推定値が変化し，２軸の組合せの方向を加えたとき２度最小推定値が変化していることがわかる（背景の色は最小推定値の場所が変化している位置を示している）．

次に初期値探索を組み込んだときの図である図４を説明する．最初の 60 回のジョブ実行は初期値を決めるために，ランダムな探索を行っている．したがって，図３に比べると，ジョブ実行回数が増えているが，初期値の決定後の探索回数が減少したので，全体の増加が抑えられている．

目的２に対して，並列化効率を高めるために，センター運用上で稼働状況により動的に決まる同時実行可能数とプログラム上で同時に起動する並列化数を合わせる．コンピュータシステムの運用時の状況に合わせて，これまでの手法では，システムの稼働状況で変化する最大の同時実行ジョブ数までジョブを投入する．もし，実行したい複数のジョブ実行数がそれを越えるときは，ジョブが終了を待ち，さらにジョブを投入する．したがって，運用上の最大同時実行可能ジョブ数と実行したい複数のパラメタの組合せ数が一致しないため，並列化効率は落ちる．図１では第 1-1 段階のときは 60 回の探索を必要としたが，このときの稼働環境での最大並列度は 50 のため，実質の実行時間は，最初の 50 個と次の 10 個に分けられている．これまでは，60 個のジョブをすべて終わってから，その中から最良値を探索したため，並列化効率がよくなかった．そこで，ここでは AT は 50 個

が終わった時点で評価を行い，新しい最良点が見つかったらすぐにその点を中心に探索を進める．このことにより，センター運用での最大同時実行数を稼働ジョブと一致させることができ，並列化効率を最大化することができる．この考え方は，目的(2)の初期値の設定のためのランダム探索にも適用できる

ここで示した３つの評価実験の結果を表２にまとめる．表２の数値はそれぞれを５回ずつ実行し，その平均値である．性能は FDE 値を用いる．FDE：Final Displacement Error は評価尺度であり，実際の歩行者到着地点と予測器の予測地点の誤差を表す．もとのユーザプログラムでの FDE 値は 1.85m であった．性能 FDE の値は初期点探索追加版では改善されており，改良版では悪くなっている．ただ、数%以内の差であり，もとのプログラムからはそれぞれ大幅に削減されている．

実行ブロック数は，並列に実行するジョブの集まりであり，これが並列の単位となる。よって，最下段の１実行ブロック当たりの平均並列数は総ジョブ数を実行プロック実行ブロック数で割れば良い．それぞれの値を見ると，総ジョブ数は少しずつ増えているが，総実行時間は大幅に減少している．

１実行プロック当たりの平均並列数をみると，改良版では，アルゴリズム自体がその並列度を実行するプログラミング環境に合わせて，大幅に向上していることがわかる

目的３として，我々が開発した AT ツールと広くパラメタ探索ツールとして用いられている Optuna と比較する．比較対象とするアプリケーションは目的２と同じく歩行者経路予測プログラムである．Optuna では検索回数の指定が必要となる．比較のため，Optuna の検索回数を我々の開発した AT ウールの検索回数と同じ 205 回に設定した．

表3 実行結果(上位5位)

	Optuna		開発したATツール	
Rank	Order	FDE(m)	Order	FDE(m)
1	92	1.06	67	1.14
2	111	1.11	161	1.16
3	100	1.12	118	1.16
4	155	1.13	167	1.16
5	154	1.13	59	1.17

表３は Optuna と開発した AT ツールを使って得られた実行結果のトップ 5 を示す．表中の Rank は，FDE の順番に並んでいる．Order は最小値を見つけたジョブ番号，FDE 値を示す．普段ユーザが用いるハイパーパラメタの組合せでの FDE 値は 1.86 であった．これをもとに考慮すると， Optuna は 0.43(=1-1.06/1.86)の相対的改善を達成したのに対し，開発した AT ツールは 0.39(=1.0-1.14/1.86)となった．これは，Optuna のパフォーマンスが開発した AT ツールより 4 ポイントほど優れていることを示している．ただし，開発した AT ツールは 67 回目の実行で，92 回目に実行した Optuna よりも早くトップ値を発見している．このように互いに優劣があり，開発した AT ツールはハイパーパラメタのチューニングにおいて Optuna と同等の効率性を示していると言える．AT ツールは効果的な実行時間値を迅速に収束させることができるが，その後，その値が設定された終了条件を満たすかどうかを検証するためにかなりの時間を費やす．この改善が重要となる．

目的 4 として，開発した AT ツールは，名古屋大学の不老 Type II 上で開発を進めたが，不老 Type II に特化しているわけではない．それを示すために，九州大学情報基盤研究開発センターの ITO など GPU を持つスーパーコンピュータ上で実験を行う．結論として，名大 不老で開発したツールをそのまま，ほとんど修正なしに，九州大学情報基盤研究開発センターの ITO 上で実行することができた．

6. 進捗状況の自己評価と今後の展望

本年度は，目的１，２に関して，大幅に並列化効率を向上させ，システム運用上の 100%の利用率を達成できた.

目的３のツール比較により，我々の開発した AT ツールが競争力あることがわかったが，さらに改善が必要なことも明確になった.

目的４により，移植性も問題ないとこがわかったので，さらに広く使える環境を増やし，実績を作っていきたい.

7. 研究業績

(1) 学術論文 （査読あり）

なし

(2) 国際会議プロシーディングス （査読あり）

なし

(3) 国際会議発表（査読なし）

[1] Yuga Yajima, Akihiro Fujii, Teruo Tanaka, Job level parallel search in software auto-tuning, HPCAsia2024 (Poster), 2024.3

[2] Teruo Tanaka, Acceleration Techniques for Software Auto-Tuning to Hyperparameters on Machine Learning Software, 2024 Conference on Advanced Topics and Auto Tuning in High-Performance Scientific Computing (ATAT on HPSC), 2024.3.

(4) 国内会議発表（査読なし）

[3] 田中 輝雄，機械学習ソフトウェアへのソフトウェア自動チューニング技術の適用，第 30 回 AT 研究会オープンアカデミックセッション (ATOS30)，2023.11.

(5) 公開したライブラリなど

現在，β版の開発済みであり，一部のユーザに試用をお願いしている.

(6) その他（特許，プレスリリース，著書等）

なし