様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月　 ９ 日現在

機関番号：３２６８９
研究種目：挑戦的研究（萌芽）
研究期間：2018～2019
課題番号：１８Ｋ１９７８６

研究課題名（和文）深層学習フレームワークでの利用を目指した完全準同型暗号による行列計算に関する研究

研究課題名（英文）A Study of Matrix Multiply by Homomorphic Encryption for Utilizing in Deep Learning Frameworks

研究代表者
木村　啓二（Kimura, Keiji）

早稲田大学・理工学術院・教授

研究者番号：５０３１８７７１

交付決定額（研究期間全体）：（直接経費）　　4,900,000円

研究成果の概要（和文）：本研究では、準同型暗号方式により暗号化された値による行列積計算の高速化手法の検討を行う探索型の研究である。得られた高速化方式に基づき、深層学習フレームワークでの利用を目指す。本研究により、汎用プロセッサのSIMD演算化で対象となる行列積処理を構成する重要な計算部分2カ所をそれぞれ最大5.53倍、3.73倍高速化することができた。また、アクセラレータ利用時に重要となるデータ転送ユニットを開発した。さらに、最小限の演算量で必要な計算を可能とするように、演算精度と計算速度の検討・評価を行い、畳み込み層5層削減した並列処理で認識精度およそ8ポイント、認識速度でおよそ54%の向上が確認できた。

研究成果の学術的意義や社会的意義
準同型暗号により暗号化したまま計算可能なことで、秘密を保ったままクラウドなどの第三者環境にデータを提供し安全に計算処理を行うことができるようになったが、その計算コストが極めて大きいことが問題となっていた。本研究により得られた成果により、準同型暗号による行列積の処理を高速化可能となる。行列積は深層学習処理の主たる計算要素であるため、秘密を保ったままにしてクラウドで深層学習処理（主に推論処理を想定）を行い、結果を安全に利用者に返すことが可能となる。

研究成果の概要（英文）：This research aims at accelerating matrix-multiply in homomorphic encryption toward utilizing it in deep learning frameworks. Through the research, we obtained 5.53x and 3.73x speedups in maximum for two important computational parts in the target encrypted matrix-multiply process. In addition, we have developed a data transfer unit, which can quickly provide required data to accelerator hardware units. We also investigated and evaluated the relationship between the precision of computations and calculation time to reduce the calculation cost while keeping the appropriate precision. As a result, we obtained 8 points accuracy improvement and 54% speedup for image recognition at the same time by parallel inference with eight smaller neural networks.

研究分野：計算機システム

キーワード：秘密計算　準同型暗号　高速化　マルチコア　アクセラレータ　FPGA

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

現在、クラウド上に世界中のあらゆるデータが集約され深層学習（ディープラーニング）に用いられている。集約されるデータに個人の医療情報や社会保障情報などの秘密情報が含まれる場合、これら秘密情報の漏洩が大きな社会的損失を引き起こす。

通常、クラウドサービスのユーザがデータをクラウドに送信する場合、これらのデータはアプリケーションにより暗号化されている。しかしながらクラウド側ではこれらの情報を通常の暗号化されていないデータ（平文）に復号してから処理を行わなければならず、このときに情報漏洩の危険がある。また、クラウド側の深層学習による推論結果をユーザが受け取るとき、その中に他のユーザの情報が意図せずに混入してしまうかもしれない。

一方、情報を暗号化したまま演算処理を可能とする準同型暗号あるいは完全準同型暗号の研究が活発に進められている。本技術をクラウド環境で動作する深層学習処理に適用することで、情報漏洩の危険を回避可能である。本技術を用いることにより、演算処理が暗号化されたデータにより行われるため、ユーザが他ユーザの情報が混入した情報を受け取ることはない。
しかしながら完全準同型暗号による演算処理は極めて大きな時間がかかることが知られている。すなわち、たとえクラウドに多くの計算資源を集めても実用的な時間で処理できない状況となっている。

２．研究の目的

本研究の目的は、クラウド環境で秘密情報を用いた深層学習を行う状況を想定し、これらの秘密情報を暗号化したまま高速学習・推論処理するシステム構築に向けた、準同型暗号処理による行列計算を高速化するハードウェア・ソフトウェアの検討である。

３．研究の方法

本研究では、まず深層学習処理を高速化するにあたり最も重要な処理である行列積に着目する。その上で、目的である準同型暗号による行列積高速化を、アプリケーションからハードウェアまで含めたシステム全体と捉えたときに、いかにして処理を最適化できるか、という観点から高速化手法を探索する。すなわち、準同型暗号処理による要素演算のみ、行列積のみ、ハードウェアの高速化機構のみ、といった要素を個別に着目することをせずに、深層学習で用いる行列積に要求される要件、それを準同型暗号化して処理した場合のハードウェアの挙動、及びこれを高速化するためのハードウェアの高速化機構、を有機的に捉える方針をとる。

例えば、行列積は特にスーパーコンピュータ上の計算で広く用いられている処理であり、このような場合では浮動小数点による演算を行い、その精度は単精度（32 ビット）や倍精度（64 ビット）、場合によっては四倍精度（128 ビット）とする必要がある。しかしながら深層学習ではこのような高い演算精度は不要であり、学習処理の場合は半精度（16 ビット）がよく使われ、さらに学習データを用いた推論処理では整数演算で十分であるという報告もある。このような深層学習特有の要求演算精度や、行列積そのもののメモリアクセス特性などを、各種条件を変えた実験により調査しデータに基づき検討を行う。

このような検討に際し、本研究では、IBM の開発した HElib あるいは Microsoft の Seal といった既存の準同型暗号ライブラリを用いて実験評価を行い、演算加速方式の予備評価としてマルチコアサーバ、GPGPU、あるいは FPGA を用いる。

４．研究成果

(1) 評価プラットホームの選定並びに演算精度の検討

対象となる行列積の浮動小数点演算に関しては、現在の準同型暗号計算が整数演算を対象としており、実用的な浮動小数点演算処理を実装するのは研究期間内では困難と判断し、本研究では整数演算により処理することとした。整数演算を行う場合の必要精度の検討に関しては、まず深層学習における演算ビット数削減手法について調査を行った。これらの先行研究として、ニューラ ルネットワークの重みを二値化する BNN や eBNN、及び畳み込み演算まで二値化する XNOR-Net があり、これら先行研究における知見は本研究においても有用である。さらに、研究協力者であるエジプト日本科学技術大学(E-JUST)の El-Mahdy 教授と深層学習におけるパラメータの精度による演算量変化が実行性能にもたらす影響に関する議論と評価を進めた。深層学習フレームワーク Caffe 用公開モデル集の調査では、本研究で評価を行う行列サイズを検討するため、広く公開されている深層学習モデル集である Caffe Model Zoo に収録されている 52 のモデルに対して、それらで使用されている層の数や行列積のサイズを調査した。

(2) 準同型暗号ライブラリ HElib による行列積計算高速化の検討

(1)の検討により本研究が実験対象とすることにした準同型暗号ライブラリ HElib による行列積計算高速化の検討を行った。HElib は公開鍵暗号方式による準同型暗号ライブラリであり、平文を公開鍵により暗号化し、暗号化文を秘密鍵により復号化する。行列積の評価には、このライブラリに付属する行列積のテストプログラムである Test_matmul.cpp を用いて行った。本プログラムを、行列サイズに関連するパラメータを変更しつつ実行時間を測定した結果、GenKeySWmatrix 関数に約 50-80%、DoTest 関数に約 10-30%の時間がかかっていることがわかった。ここで、GenKeySWmatrix 関数では key-switching に用いられる行列の生成が行なわれてい

る。key-switching とは、暗号化した行列要素の格納位置を移動する automorph と呼ばれる処理に伴い変化する秘密鍵を元に戻すための処理である。生成した key-switching 用の行列は公開鍵に埋め込まれる。この行列生成処理は計算前に一回行われれば良い。また、DoTest 関数では行列・ベクトルの encode（多項式化）・暗号化、演算、復号化が行われている。さらに、HElib が備えているマルチスレッド機能を用いて評価を行った結果、スレッド数増加に応じた性能向上が確認され、8 スレッド実行時に 1 スレッド実行時の約 5 倍の性能向上を得ることができた。

上記の測定結果を基に、GenKeyMatrix 関数、及び DoTest 関数における暗号文の加算・乗算部の高速化を行った。前述の通り、HElib のマルチスレッド機能によりマルチコアを用いた性能向上は既に実現されているため、SIMD 命令による性能向上を目指した。

GenKeyMatrix 関数では、Horner 法による多項式計算の高速化を行っており、これに実行時間を要する。Horner 法では、以下のような多項式の変形を行うことにより、乗算の回数を n 回に抑えることができる。

$$a_n \times x^n + a_{n-1} \times x^{n-1} + a_{n-2} \times x^{n-2} + \cdots + a_1 \times x + a_0$$
$$= \left(\cdots\left((a_n \times x + a_{n-1}) \times x + a_{n-2}\right) \times x + \cdots + a_1\right) \times x + a_0$$

上記の変形後の式を各繰り返しでxを乗ずるループにより実現するが、この実装ではループ繰り返し間に依存が生じてしまい SIMD 化できない。そのため、Horner 法適用後の式を以下のように分割した：

$$A_{0-3} = a_n \times x^3 + a_{n-1} \times x^2 + a_{n-2} \times x^1 + a_{n-3} \times x^0$$
$$A_{0-6} = A_{0-3} \times x^3 + a_{n-4} \times x^2 + a_{n-5} \times x^1 + a_{n-6} \times x^0$$
$$A_{0-9} = A_{0-6} \times x^3 + a_{n-7} \times x^2 + a_{n-8} \times x^1 + a_{n-9} \times x^0$$
$$\cdots$$

このように分割することにより各A_{0-m}の計算に対して SIMD 化が適用できる。また、オリジナルの HElib の実装では演算に倍精度浮動小数点を利用していたが、対象となる深層学習推論で利用する場合この精度は過剰であると考え、本研究では単精度浮動小数点で演算を行いその分 SIMD 幅を拡大した。

DoTest 関数における暗号文の加算・乗算部の高速化に関しては、まず加算の SIMD 化に関しては自然に適用可能である。本処理における乗算は、より正確には$a \times b(mod\ n)$であり、HElib 中の実装では演算コストの高い剰余演算を直接行わず、シフト演算を組み合わせることにより実現している。本研究では、これらの SIMD 化を permutation 処理を組み合わせることにより実現した。

上記の各演算に対する高速化処理を評価した。評価には Intel Xeon W-2145(3.7GHz)を搭載したサーバを用いた。またコンパイルには Intel C++ コンパイラ（Parallel Studio XE2018）を用いた。まず GenKeyMatirx 関数の Horner 法に関しては SIMD 化と演算精度の単精度化により約 3.4 倍の性能向上を得ることができた。また、DoTest 関数の加算部に対しては行列サイズに関するパラメータにより 3.64 倍から 5.53 倍の性能向上を、乗算部に関しては 2.74 倍から 3.73 倍の性能向上をそれぞれ得ることができた。

(3) ハードウェアアクセラレータによる準同型暗号による行列計算の検討

①アクセラレータ用データ転送機構の開発

演算処理をアクセラレータにより高速化する場合、アクセラレータそのものだけではなくアクセラレータにデータを供給する機構が重要になる。本研究では、従来のデータ転送機構では効率的なデータ転送が困難な、疎行列演算時に頻出する A[B[i]]のような間接参照のデータを効率よく転送する機構の開発も行った。

開発したデータ転送機構のブロック図を図 1 に示す。本機構は二つのデータ転送器（DMAC1, DMAC2）をカスケード接続する構成となっていることを特徴とする。本データ転送機構では DMAC1 の Index Read Port により B[i]のようなインデックス配列をメモリから読み出す。これに、予め設定した転送対象配列のベースアドレス（&A[0]）を Address Calculation で加算することによりデータのアドレス（&A[B[i]]）を求める。得られたアドレスにより DMAC2 の Data Read Port からデータを読み出す。読み出されたデータは転送先のアクセラレータ用メモリに連続的に配置されるため、複雑なメモリアクセス機構を必要とせずにアクセラレータが効率よくデータを読み出すことができる。さらに、Data Read Port にキャッシュを付加し、データの空間的局所性を利用する構成に関しても検討した。

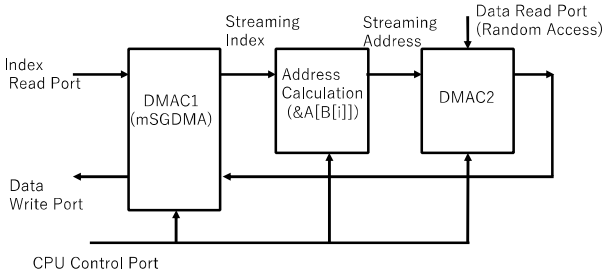

図 1: 提案データ転送機構のブロック図

提案データ転送機構を含むベクトルマルチコアアーキテクチャを Intel Arrai10 FPGA 評価ボードに実装し評価した。実装したベクトルマルチコアを図 2 に示す。本マルチコアは CPU コアとベクトルアクセラレータコアを持つ Processing Element(PE)を 4 つ搭載し、各 PE がそれらコアと共に提案するデータ転送機構を持つ。本データ転送機構はマルチコア外の DDR メモリから PE 内のローカルデータメモリ（LDM）にデータを転送し、ベクト

ルアクセラレータが転送されたデータを用いて演算処理を行う。実装したマルチコアでは、各 PE は 50MHz で動作し、またメモリバスには 10MHz, 30MHz, 及び 50Mhz を設定可能である。本報告では、30MHz を設定したときの評価結果を掲載する。

本データ転送機構を持つベクトルマルチコアアーキテクチャを疎行列・ベクトル積（SpMV）を用いて評価した。評価には、Florida Sparse Matrix Collection のデータセットを利用した。評価結果を図 3 に示す。図は各データセットを 1PE から 4PE で実行した場合の CPU 転送（CPU transfer)、キャッシュ無し提案データ転送機構（CDMAC w/o Cache)、及びキャッシュあり提案データ転送機構（CDMAC w/ Cache）それぞれの性能を MFLOPS で示したものである。評価の結果、psmigr_2 で 1PE の場合にキャッシュありデータ転送機構が CPU 転送に対して 18.0 倍、4PE 使用時は 12.8 倍、及び 4PE 使用時に平均 9.9 倍の性能向上を得ることができた。また、キャッシュの有無に関しては、4PE 使用時に平均 2.5 倍の性能向上を得ることができた。

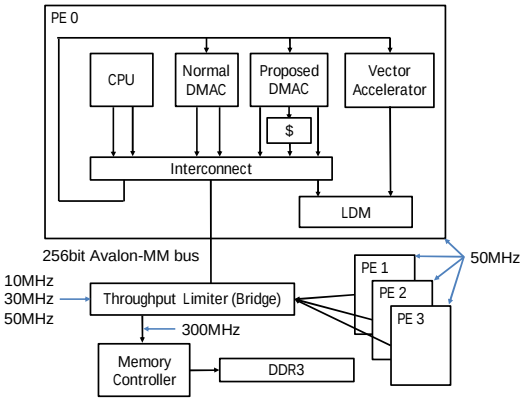

図 2: 提案データ転送機構を備えたベクトルマルチコア

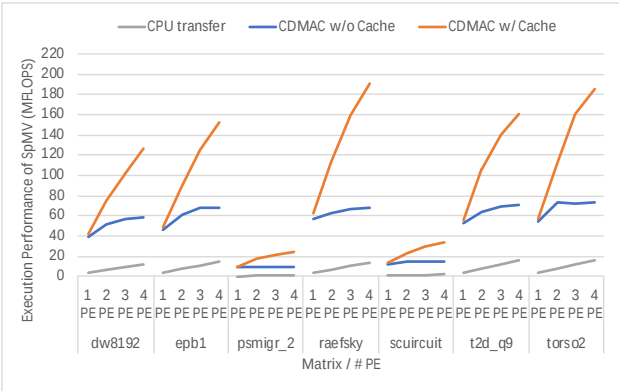

図 3: 提案データ転送機構の評価結果

②アクセラレータによる推論精度と行列サイズ（ネットワークサイズ）のトレードオフに関する検討

準同型暗号を用いた深層学習の高速化に寄与することを目的として、特に推論精度を保ちつつニューラルネットワークのサイズ（つまりは行列積における配列サイズ）を縮小する方法について検討・評価を行なった。

深層学習において認識精度を向上させるためには，ニューラルネットワークの層数を増やし、あるいはネットワークの構造を複雑にすることにより、より規模の大きいネットワークを利用する方法が取られることが多い。しかしながら，こういった方法では，深層学習処理の大部分を占める行列・ベクトル演算における行列・ベクトルのサイズが大きくなり、準同型暗号を用いた深層学習処理を行う際，より不利な状況を招いてしまう。そこで本研究では、小規模なニューラルネットワークを複数並列に用いる手法を提案・検討した。

提案手法では、複数のニューラルネットワーク間で認識対象を分担することで、認識精度を保ちつつ、個々のニューラルネットワークの規模を縮小することを目的としている。画像認識で広く用いられる VGG16 と Inception Resnet-V2 を参考に畳み込み層を 5 層に縮小した小規模かつ単純な構造を持つ畳み込みニューラルネットワークを作成し、Digilent ZYBO Zynq-7020 上に実装して評価を行なった。

開発環境の高位合成機能を利用して FPGA への実装を行い、作成したニューラルネットワーク 8 つを並列に用いて推論を行った結果、1 つの VGG16 を同様に FPGA 上に実装したと比較して、認識精度でおよそ 8 ポイント、認識速度でおよそ 54 パーセントの向上がみられた。認識精度を維持あるいは向上させつつ、個々のニューラルネットワークの規模、つまりは内部で演算に用いられる行列・ベクトルのサイズを縮小することにより、準同型暗号を用いた深層学習処理の高速化につながるものと考えられる。なお、本手法では複数のニューラルネットワーク

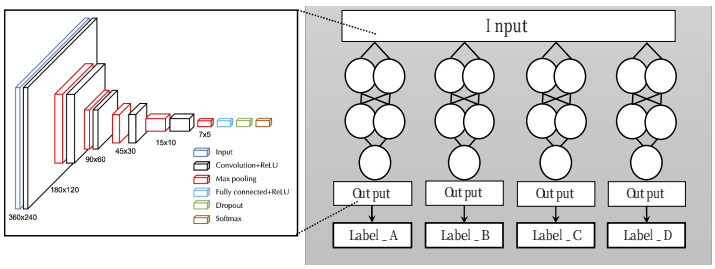

図 4: 複数ネットワークを並列に用いた推論

表 1: 並列推論による認識精度・推論時間の向上

	VGG16		提案ネットワーク	
並列ネットワーク数	1	2	1	8
認識精度[%]	74.2	78.6	71.5	82.4
推論速度[FPS]	12.4	13.3	13.4	19.2

を複数同時に用いるため、全体の演算量としては増加する可能性があるが、個々のニューラルネットワークは相互に依存関係になく、並列にこれらを用いることができるため、演算時間の増加を防ぐことが可能となる。

以上の成果により準同型暗号による行列計算を高速化する要素技術を得ることができた。今後は、これらの手法を利用して準同型暗号による行列計算全体の評価、及びこれを利用した深層学習の推論処理を実現し性能評価を行うことが目標となる。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　2件／うちオープンアクセス　2件）

１．著者名 OKI Yoshitake、ABE Yuto、YAMAMOTO Kazuki、YAMAMOTO Kohei、SHIRAKAWA Tomoya、YOSHIDA Akimasa、KIMURA Keiji、KASAHARA Hironori	４．巻 E103.C
２．論文標題 Local Memory Mapping of Multicore Processors on an Automatic Parallelizing Compiler	５．発行年 2020年
３．雑誌名 IEICE Transactions on Electronics	６．最初と最後の頁 98～109
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transele.2019LHP0010	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 ADHI Boma A.、KASHIMATA Tomoya、TAKAHASHI Ken、KIMURA Keiji、KASAHARA Hironori	４．巻 E103.C
２．論文標題 Compiler Software Coherent Control for Embedded High Performance Multicore	５．発行年 2020年
３．雑誌名 IEICE Transactions on Electronics	６．最初と最後の頁 85～97
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transele.2019LHP0008	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計15件（うち招待講演　1件／うち国際学会　7件）

１．発表者名 柏俣 智哉 、北村 俊明、木村 啓二、笠原 博徳
２．発表標題 DMAのカスケード接続による間接ロードの高速化
３．学会等名 情報処理学会　第226回 システム・アーキテクチャ研究発表会
４．発表年 2019年

１．発表者名 Tomoya Kashimata, Toshiaki Kitamura, Keiji Kimura, Hironori Kasahara
２．発表標題 Cascaded DMA Controller for Speedup of Indirect Memory Access in Irregular Applications
３．学会等名 9th Workshop on Irregular Applications: Architectures and Algorithms（国際学会）
４．発表年 2019年

１．発表者名
Keiji Kimura

２．発表標題
Cascaded DMAC Enabling Efficient Data Transfer for Indirect Memory Access Applications

３．学会等名
4th International Symposium on Research and Education of Computational Science (RECS)（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
牧田 哲也，宍戸 哲平，和田 康孝，木村 啓二

２．発表標題
準同型暗号による行列積の高速化の検討

３．学会等名
電子情報通信学会技術報告，CPSY2019-96,DC2019-102(2020-02) (ETNET2020)

４．発表年
2020年

１．発表者名
山本 一貴，藤田 一輝，柏俣 智哉，高橋 健，Boma A. Adhi，北村 俊明，川島 慧大，納富 昭，森 裕司，木村 啓二，笠原 博徳

２．発表標題
マルチターゲット自動並列化コンパイラにおけるアクセラレータコスト推定手法の検討

３．学会等名
情報処理学会研究会，Vol.2019-ARC-240 No.25，Vol.2019-SLDM-191 No.25，Vol.2019-EMB-53 No.25 (ETNET2020)

４．発表年
2020年

１．発表者名
田處 雄大，見神 広紀，細見 岳生，木村 啓二，笠原 博徳

２．発表標題
OSCAR自動並列化コンパイラとNECベクトル化コンパイラの協調によるベクトル・パーソナルスパコン上での自動ベクトル並列化

３．学会等名
情報処理学会研究会，Vol.2019-ARC-240 No.26，Vol.2019-SLDM-191 No.26，Vol.2019-EMB-53 No.26 (ETNET2020)

４．発表年
2020年

１．発表者名
青戸 武蔵，和田 康孝，三ツ木 萌

２．発表標題
FPGA上でのCNNパラメータ動的更新手法の性能評価

３．学会等名
情報処理学会 第82回全国大会

４．発表年
2020年

１．発表者名
Musashi Aoto, Moe Mitsugi, Takumi Momose, and Yasutaka Wada

２．発表標題
Towards the Improvement of Training Efficiency and Image Recognition Accuracy for an FPGA Controlled Mini-Car by Offloading Neural Network Training

３．学会等名
Proc. of The 2019 International Conference on Field-Programmable Technology (FPT2019), FPGA Design Competition（国際学会）

４．発表年
2019年

１．発表者名
青戸 武蔵，比留川 翔哉，和田 康孝，丸山 一貴

２．発表標題
単機能なニューラルネットワークを複数用いた高速・高精度な画像認識のFPGAによる実現

３．学会等名
RECONF2019-31

４．発表年
2019年

１．発表者名
Musashi Aoto, Shoya Hirukawa, Yasutaka Wada, Kazutaka Maruyama

２．発表標題
An FPGA based Autonomous Driving Car Design using Multiple Simple Neural Networks for Decision Making

３．学会等名
The Tenth International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies (HEART2019) FPGA Design Contest（国際学会）

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	和田　康孝 (Wada Yasutaka) (40434310)	明星大学・情報学部・准教授 (32685)	
研究協力者	エルマハディ　アムド (El-Mahdy Ahmed)	Ｅ－ＪＵＳＴ・Department of Computer Science and Engineering・Professor	