公開

研 究 終 了 報 告 書

「大規模で不完全なセンサデータに対する高速な最近傍探索」

研究期間: 2019 年 10 月～2023 年 3 月
研 究 者: 松井 勇佑

1. 研究のねらい

本研究の目的は、「エッジコンピュータ上での高速な情報処理基盤」を実現することにある。具体的に、「ARM アーキテクチャ上での高速な近似最近傍探索」の実現を目指す。

私が対象とするのは IoT 技術全般の中でセンサノードに相当するエッジコンピュータである。特に、ラズベリーパイや NVIDIA Jetson といった、ARM アーキテクチャによって構成されるコンピュータを考える。現在主流の x86 アーキテクチャに比べ、ARM は新しいアーキテクチャである。ARM は費用効率や熱効率の面で優れていると言われ、近年急速に普及が進んでいる。

近似最近傍探索とは「クエリベクトルが与えられたとき、ベクトル集合の中から似ているものを返す」という処理である。近似最近傍探索は計算機科学における基盤的な問題であり、速度・精度・メモリ消費の三者のトレードオフのもとに様々な分野で様々な方式が提案されている。近似最近傍探索は多くのアルゴリズムのビルディングブロックであるため、近似最近傍探索の性能を向上させることは、情報処理全体の高速化につながる。

本研究では、IoT のエッジである ARM アーキテクチャ上にて、高速な近似最近傍探索の実現を目指す。現在の非力な ARM デバイス上では、近似最近傍探索手法の速度は十分ではない。ARM デバイス上で高速な近傍探索を実現することにより、(1) 基盤的な処理である探索の性能向上、(2) その直接的な応用として、三次元画像処理の速度向上、を目指す。また、本提案は基盤的な処理を高速化することを目的とするため、提案する方式は広く実際に人々に使われ実際に役に立つことを目指す。

2. 研究成果

(1)概要

本研究で達成した内容について、(1)理論、(2)応用、(3)アウトリーチの3点に分けて述べる。

理論について、近似最近傍探索における理論的な研究を行った。特に、「研究テーマA」で述べる「ARM 上での高速探索」は、本研究において最も取り組みたかった「ARM 上での近傍探索」そのものを実現したものである。本内容は信号処理分野におけるトップ会議である ICASSP にて発表された。本内容は近似最近傍探索におけるデファクトスタンダートのライブライである faiss に取り込まれており、私が目指していた「人々みなの役立つもの」となった。「研究テーマB」で述べる「効率的なデータ削減」では、大量のデータがあるとき、前処理としてゴミデータを発見し捨てるというものである。そのような「ゴミ捨て」を行うことにより、前処理として問題規模を小さくすることが可能になる。これはエッジのような非力な計算機における有効なアプローチである。

応用について、特に計算が重い三次元化像処理の高速化を進めてきた。とくに「研究テーマ

B」で述べる「点群位置合わせ高速化」では、三次元処理の基本となる点群位置合わせについて高速化を達成したものである。本内容はコンピュータビジョン分野の第4位の国際学会であるBMVC にて発表された。エッジ上での三次元処理としては、例えば自動運転車に小型のコンピュータをとりつけ外界の情報を処理したいといった需要がある。そのような場合に、本提案の方式は効果を発揮する。

最後にアウトリーチについて、本研究に関して私はコンピュータビジョン分野トップ会議であるCVPR、およびマルチメディア分野トップ会議である ACM Multimedia にてけるチュートリアルを行った。とくに CVPR のチュートリアルは、公開したスライドのビュー数が 29000 件を超えるなど注目を浴びた。それに関連し、国内の様々な場所で招待講演も行った。本さきがけにより、私自身、探索の専門家としての立ち位置を固められたと考えている。

(2)詳細

研究テーマ A「ARM 上での高速な近傍探索」

近似最近傍探索とは、ベクトルの集合に対しクエリベクトルが与えられたときに、クエリに一番近いベクトルを集合中から探す操作を指す。近似最近傍探索は計算機科学における基本的な操作であり、データベース・画像処理・自然言語処理などの各分野にまたがって長く研究が続けられてきた。その中でも、近年注目を浴びている手法の一群が「4-bit Product Quantization; 4-bit PQ」というものである。これは、ベクトル量子化の一種である直積量子化(Product Quantization; PQ)を用いた探索方式に対し、SIMD 演算命令をフルに活用することで超高速な探索を実現するものである。この 4-bit PQ 方式は x86 アーキテクチャ上で現在最高速度を達成する方式として知られている。

我々の目的はこの 4-bit PQ を ARM アーキテクチャ上で高速に実現することである。もし ARM 上での高速探索が実現できると、ラズベリーパイを始めとしたマイコンなどでも高速探索が可能となる。

ARM 上での高速な 4-bit PQ を実現する上での技術的な課題は、4-bit PQ は x86 専用の SIMD 命令を多数使っているという点にある。よって、これを直接 ARM に用いることはできない。特に、x86 における AVX2 命令は 256bit の SIMD レジスタを前提としてアルゴリズムが設計されているが、ARM では一般的に 128 ビットのレジスタしか準備できない場合が多い。このギャップをどう埋めるかが課題となる。

これを解決するために、我々は 128 ビットレジスタを二つ(仮想的に)連結し、シャッフル命令をその両方に適用するという方式を提案した(図1)。左は x86 で用いられている 128 bit 用のものである。x86 はこれを2つ(合計 256 bit)同時に実行することができる。右は我々が提案する方式である。我々は仮想的に 256 bit の状況を作ることで、統一したアルゴリズムでの実行を可能とした。本内容は信号処理分野第一位の国際会議である ICASSP に採録された。

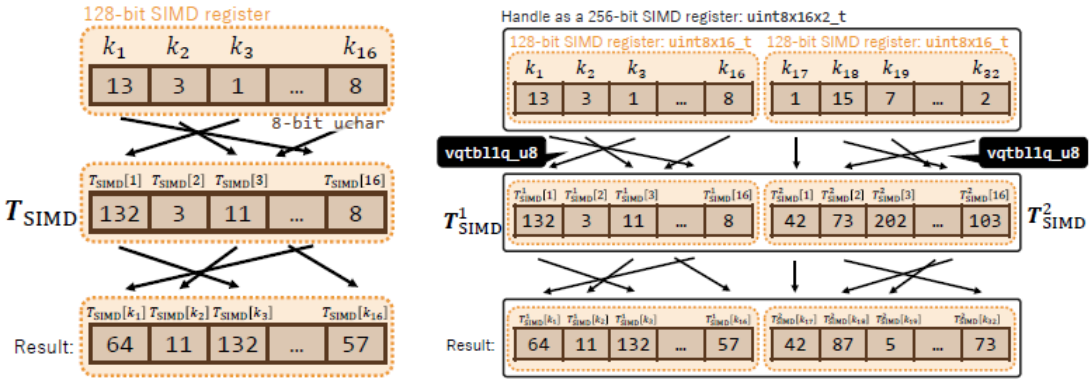

図 1　提案する SIMD 計算方式

提案方式は優れた性能を残しただけではなく、デファクトスタンダードなライブラリである faiss に実装が採用された。これは、faiss における ARM アーキテクチャ全般の性能を底上げするものだと言える。現在、faiss ライブラリを用いることで、世界中の誰でも今すぐに我々の方式を使うことができる。

提案方式は「研究のための研究」という枠ぐみを超えて、「世界中の ARM 使用者に対し高速な処理を提供する」ことを実現した。これはまさに私がさきがけで行いたかったことであり、それを実現することができた。

研究テーマ B「効率的なデータ削減」

エッジコンピュータの計算力は非力であるため、エッジ上での探索は難しい課題である。計算用サーバであれば豊富な計算機資源を用いて高速に実行できる処理も、エッジではそもそもデータをメモリ上に読み込むことすら難しい場合がある。「データがメモリに載るかどうか」はそのあとにどのような処理が可能になるかどうかに直結する重要な課題である。

よって、「探索精度を担保したまま、如何にしてデータ量を削減するか」は取り組むべき課題である。ここで説明のため、D 次元のデータが N 個あるとする。このとき、一つの数字が float (4 byte) で表されるとすると、最低でも 4DN バイトの領域が必要である。通常、データを減らしたいときは、次元削減をおこない D を小さくする。だが、その場合必ず探索問題の精度は低下する。そこで、私は「大量データをサブサンプリングして、N を減らす」という方式を提案した。これはすなわち、大量データ中から外れ値やゴミを取り除くような処理に相当する。もし最終精度に影響しないゴミデータを事前に取り除くことが出来れば、そのぶんだけメモリ消費量を削減することができる。それにより、より、大量データをエッジ上で扱いやすくすることができる。

我々は古典的な統計量である Hubness score を用いることで、そのようなサブサンプリングを実行できることを実験的に示した。Hubness score はデータの「人気度」を測る指標であり、この人気度が低いものを単純に取り除くことで、ゴミデータの除去を前処理として実現することが出来た。本内容はマルチメディア検索分野の会議である ICMR に採録された。

研究テーマ C「高速な点群位置合わせ」

近傍探索処理を直接的に用いる応用処理として「三次元点群処理」、特に「三次元点群位置合わせ」が挙げられる。私は三次元点群位置合わせ処理の高速化を達成した。

三次元点群位置合わせ処理の概要を図2に示す。三次元点群とは、三次元物体の形状を、三次元点の集合で表したものである。例えば図2のウサギの図にあたる。このような三次元点群は、例えばライダーカメラといった特殊なカメラで物体を撮影することで得られる。ここで、二つの三次元点群（図2のように、同じ物体を別の角度から見たものである場合が多い）が与えられたとき、それがピッタリ重なるように点群の位置を調整する（並進行列と回転行列を求める）処理を三次元点群位置合わせという。これは三次元情報処理を行う際の基盤的な技術である。例えば自動運転を考えるとき、車載カメラ上から外界を三次元点群として取得し、位置合わせを行って物体の検出を行う。

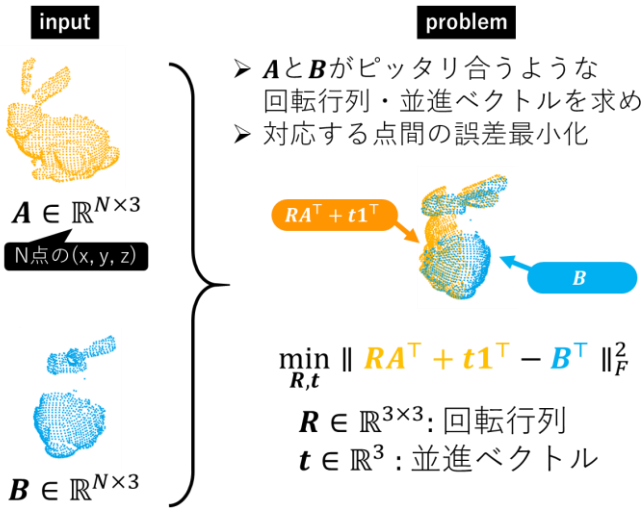

図 2　三次元点群位置合わせ

既存方式では、重い処理を反復的に実行し解を求めていた。これについて私は行列演算を展開し整理することで、計算に冗長な部分があることを発見した。その冗長な部分を取り除くことで、精度は落とさずに高速に解を求めることが可能となった。具体的に、最初の1回だけ時間をかけて重い計算を行い、後半の反復は軽量な処理を繰り返す方式を導入した。これにより、現在の最高精度の手法である RPM-Net を、精度を落とさず数倍高速化させることに成功した。本内容はコンピュータビジョン分野4番手にあたる国際会議 BMVC に採録された。

<u>アウトリーチ</u>

コンピュータビジョン分野最大の国際会議 CVPR、およびマルチメディア分野最大の国際会議である ACM Multimedia にて本研究の内容についてチュートリアルを行った。特に CVPR の内容については、公開しているスライドのビュー数が 29000、公開しているビデオのビュー数が 96000 を達成するなど、大きな話題を呼んだ。

3. 今後の展開

ARM 探索について： 本研究計画で達成した ARM 探索方式はデファクトライブラリ faiss に搭載されたため、既に世界中の人々に毎日ダウンロードされ使われている。例えば、M1 Mac にて faiss を用いる場合は、自動的に私のコードが利用されている。これのさらなる発展として、より発展

的な SIMD システムに対する探索が考えられる。私が取り組んだ ARM の SIMD は NEON という伝統的なものである。より発展的な ARM SIMD として、SVE と言った方式が存在する。そのような発展的なものについて、現在の近似最近傍業界は全くカバーできていない状況にある。その点を攻め、シェアを広げたい。これについては、次の1年以内になんらかのアクションをとれると考えている。また、提案方式が実際に研究の内容として用いられる様子を可視化していきたい。これについては、他の研究者とのコラボレーションをより広めていきたい。

三次元処理に関しては、まだ世界中の人々に使われる状態ではない。今後もより多くの高速化方式を提案し、「高速な三次元処理であれば松井のグループだ」と認知されるようになる必要がある。そのためには3年程度必要であると考えている。加えて、これについても、実際にマイコン上で切に高速三次元処理を希望している研究者とのコラボレーションが必須だと考えている。

さきがけが終了したのち、さきがけメンバーである天方さん、塩川さん、および OSX の研究者である西村さんとともに立ち上げたプロジェクトが「AIP 加速課題：超高速データサイエンス基盤」に採択された。これは本さきがけの内容をさらに発展させたものである。自分・天方さん・塩川さんにとっては、ACT-I、さきがけ、AIP 加速とステップアップしてきた内容になる。引き続き、高速化の研究をすすめていきたい。

4. 自己評価

最もやりたかった「高速 ARM 探索、およびそれをデファクトライブラリに搭載」を実現することが出来た。この点については多いに満足している。「研究のための研究」ではない、実際に使われるソフトウェアを構築するという点は強い想いを持っており、それを実現することが出来た。また、それは単に良いコードを書くということだけではなく、「目の付け所」が重要であると感じる。今回は、「ARM 上での探索」という、いつか必ず必要になる部分に誰よりも早く目を付けられたことがポイントであったと思う。今後もそのように、何が人々にとって重要なのかを意識していきたい。また、このような知見を他の研究者にフィードバックするなどして、なかなか日本の技術が情報産業のなかでシェアをとれない現状を変えていきたい。

一方で、論文の数という点では十分とは言えない。もっと多くの論文を執筆すべきであった。この点については反省が残る。また、コロナ禍の直撃、および自分自身研究室を立ち上げる、子供が出来る、といったライフイベントが重なり、国際的なコミュニケーションは思ったほどに取れなかった。特に、結局国際会議に現地参加することはさきがけ期間中に一度も無かったという状況にある。特にCVPRチュートリアルなどでは、多くのトップ研究者と直接触れあえるはずだったので、この点については後悔が残る。

5. 主な研究成果リスト

(1) 代表的な論文（原著論文）発表

研究期間累積件数：13件

1. Yusuke Matsui, Yoshiki Imaizumi, Naoya Miyamoto, Naoki Yoshifuji, “ARM 4-bit PQ: SIMD-based Acceleration for Approximate Nearest Neighbor Search on ARM”, International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2022.
ARM 上の高速な近似最近傍探索方式を提案した。既存の最速手法である 4-bit PQ は、

x86 における SIMD 演算である AVX2 を多用した方式である。よって、SIMD レジスタ幅が短い ARM においては、4-bit PQ を直接用いることはできない。本論文では、AVX2 のインストラクションを ARM のものでエミュレートすることで、アルゴリズム側に一切手を加えることなく 4-bit PQ を ARM にて実現した。提案方式はベースラインの PQ アルゴリズムに比べ同じ精度で 10 倍程度の高速化を達成した。本論文は近年の近似最近傍探索の分野で初めて ARM アーキテクチャについて言及したものである。

2. Kimihiro Tanaka, Yusuke Matsui, Shin'ichi Satoh, “Efficient Nearest Neighbor Search by Removing Anti-hub”, International Conference on Multimedia Retrieval (ICMR), 2021.

効率的なデータ削減の方式を提案した。最近傍探索の文脈では、データベースサイズが大きいほど問題の規模が大きくなり、処理が難しくなる。一方で、現実のデータベースの中には外れ値に相当する、最終精度に影響しないゴミベクトルが多数含まれる可能性がある。本論文では、ハブネスという統計値に注目し、ハブネスが低いベクトルを取り除くことで、探索の最終精度を損なうことなくゴミベクトルを除去できることを実験的に示した。

3. Yoichiro Hisadome, Yusuke Matsui, “Cascading Feature Extraction for Fast Point Cloud Registration”, British Machine Vision Conference (BMVC), 2021.

高速な三次元点群位置合わせ方式を提案した。三次元データを処理する際に、点群位置合わせとは最も基礎的で速度が要求される処理である。現在最高精度を達成する方式は、繰り返し処理のたびに特徴量抽出をし直す必要があり、計算が冗長であった。本論文では、処理に必要な行列演算を整理することで冗長な演算部分を発見し、それを取り除くことで精度を損なわずに位置合わせ速度を高めることが出来ることを発見した。

(2)特許出願

研究期間全出願件数:0 件(特許公開前のものも含む)

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

追記事項なし