公開

（事後評価・研究者用）

研　究　報　告　書

「圧縮線形代数：データ圧縮による省メモリ高速大規模行列演算」

研究期間：　2018 年 4 月～2020 年 3 月
研究者番号：　50212
研 究 者：　松井勇佑

1.　研究のねらい

本研究では「圧縮線形代数」という，高速かつ省メモリで演算を行える新しい学問領域を提案する．圧縮線形代数では，ベクトル・行列を圧縮し，省メモリでそれらを表現する．そして，圧縮した状態のままで距離計算といった数学的操作を高速に実現する．その実現のために，距離表を事前計算しておきテーブルルックアップで高速に計算を行う方式を提案する．

提案する圧縮線形代数により，人工知能問題を解く上で重要な大規模データ処理・大規模機械学習処理を従来よりずっと小規模な計算機環境で実現する．

2.　研究成果

（1）概要

近年の人工知能分野では，極めて大量のデータを扱う（例：ImageNet 機械学習における 1000 万枚以上の画像処理，また都市レベル三次元復元における 10 億個以上の点群処理）．こういった大量データを真面目に処理する場合，(1) データが多すぎてメモリに載らない，(2) ゆえに高速にデータを処理できない，という問題がある．特に，最も基礎的な処理である「探索（似たベクトルを探す）」ですら，大量のデータに対しては相当時間がかかる．これを解決するため，(1) データを圧縮してメモリに載せ，(2) 圧縮したまま探索を行う，直積量子化， およびその改良手法が，近年急速に発展してきている．直積量子化を用いると，10 億個の 128 次元ベクトルに対する探索処理を，32GB 程度の現実的なメモリ消費量で処理できる．これによって，大規模データに対する高速省メモリ探索処理が実現されている．

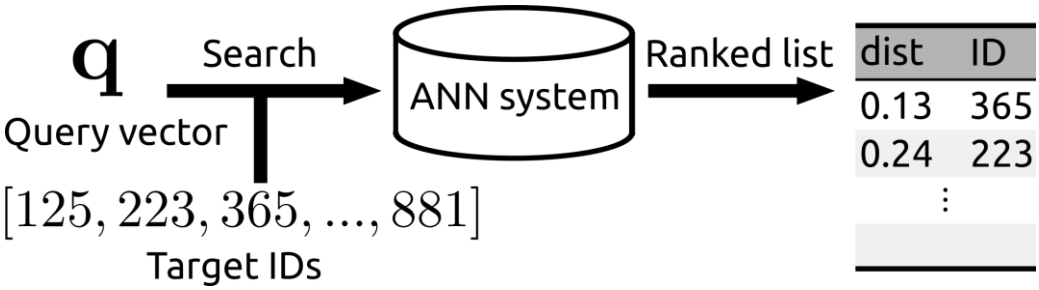
一方で，直積量子化ベースの手法は本質的には線形な探索処理を行うことしかできない．より一般的な数学的処理を扱うことが本研究の目的である．そのために，私は「部分に対する探索処理」，および「コードに対する意味の埋め込み」を実現した．これにより，通常の探索の枠を超えて，直積量子化コードに対する新たな数学的操作が実現され，大規模データ処理に対する新たな方法論が切り開かれた．加えて，コンピュータビジョン分野における最大の国内会議である MIRU にて，探索処理に関するチュートリアルを行った．また同様のチュートリアルをコンピュータビジョン分野世界最大の会議である CVPR において発表することが決定している．このように，探索分野に関する多くのアウトリーチ活動を行い，知見を広く社会に還元した．

（2）詳細

研究テーマ1 「部分に対する探索」

これまでの一般的な探索処理は，「全てのデータに対する探索」は高速であるが，「部分に

対する探索」は苦手だという直感に反する情勢にあった．私は現在広く用いられている転置インデクス型の探索処理に対し，そのような部分での探索を可能とした．これにより，例えば画像検索の文脈において，画像検索と単語検索を組み合わせるような複雑な検索が統一的に記述可能になる．この内容はマルチメディアのトップの国際会議である ACM Multimedia に採録され，オーラル発表された．その詳細を下図に記す．



ここでは簡単のため，画像検索を対象として説明する．画像検索の文脈では，画像は一本のベクトルとして表現される．事前に行われる前処理（インデクシング）として，データベース側の 1 番目から N 番目のベクトル（画像）は全て最近傍探索システムに登録される．オンラインの探索処理として，クエリのベクトル（画像）が与えられたとき，似ているベクトルを探す．これが画像検索である．

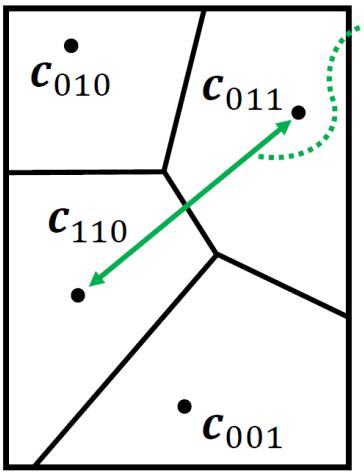
一方で、例えば「2019 年に撮影された画像の中から似ている画像を探したい」というような要望がある．の問題は，事前に撮影記録に対するフィルタリング処理などを行い画像番号を特定し，その番号集合のなかから似ているベクトルを探すという処理に相当する。すなわち、「2019 年に撮影された画像は 101 番目、245 番目、・・・4325 番目の画像である」という情報を得て，その指定されている番号の中から検索を行う．この問題は、上の図のように，検索時にクエリに加えて，ターゲットとなる画像番号の集合が与えられ，検索を行う，と立式できる．これを私は「部分に対する検索」と名付けた．

この問題は非常に大きな需要がありながら，近年の最近傍探索システムは対応していなかった．すなわち、既存の高速なシステムは全て（1～N 番目）のベクトルに対する探索には高速（そのために最適化されている）だが，部分に対する処理をサポートしていない場合がほとんどだからである．よって、この部分に対する検索問題は，従来はエンジニアリングによる力業で処理されてきた．私が提案した手法は、一つのデータ構造で二つの探索方式をサポートすることにより，「部分に対する検索」を実行可能にした．すなわち，与えられる画像番号集合の要素数が少ないときは単純な線形探索を行う．一方で，要素数が多いときは転置インデクス方式で処理を行い，後処理で絞り込む．このような二つの操作を，単一のデータ構造で実現可能とした．

研究テーマ2 「コードに対する意味の埋め込み」

直積量子化では，データ（ベクトル）はベクトル量子化により，最近傍のコードワードにアサインされる．これはすなわち，ベクトルに整数を割り振るという処理に相当する．これにより，ベクトルの全ての要素を保持するのではなく，整数一つを保持するだけでベクトルが（近似的に）表現できる．これはメモリ効率が良いデータ表現であると言える．

一方で，量子化されたあとの世界（ベクトルが整数で表された世界）では，「足し算」といった基本的な処理しか考慮されていない．私はここで，整数を二進表現したバイナリベクトルについて，重みつきハミング距離を考えることにより，これまでにない距離尺度の埋め込みが理論的に可能であることを示した．その概念図を下に示す．



$$d(\boldsymbol{c}_{011}, \boldsymbol{c}_{110}) = \|\boldsymbol{c}_{011} - \boldsymbol{c}_{110}\|_2$$

Optimize the index assignment s.t. $d \sim d_{\mathrm{H}}$

$$d_{\mathrm{H}}(011, 110) = \mathbf{1}^{\top}\left(\begin{bmatrix}0\\1\\1\end{bmatrix} \oplus \begin{bmatrix}1\\1\\0\end{bmatrix}\right)$$

Optimize $\boldsymbol{w}$ s.t. $d \sim d_{\mathrm{WH}}$

$$d_{\mathrm{WH}}(011, 110) = \boldsymbol{w}^{\top}\left(\begin{bmatrix}0\\1\\1\end{bmatrix} \oplus \begin{bmatrix}1\\1\\0\end{bmatrix}\right)$$

コードワードを c とあらわし，その添え字（整数）を二進数として表現するとする．このコードワード間の距離（緑色の矢印）は，コードワード間のユークリッド距離である．一方で，添え字の二進表現をバイナリベクトルだとみなし，そのハミング距離を考えることが出来る．このハミング距離が元の距離に近くなるように添え字の割り振りを変える方式を，Polysemous Codes [Douze, ECCV 16]と言う．この方式を用いることで，別の距離尺度をコード間に埋め込むことが出来る．

一方で，ハミング距離と元の距離間ではスケールやオフセットが違うため，必ずしも良い近似を達成することが出来ない．そこで私は，新たに重みベクトルを学習することで，ハミング距離を重み付きハミング距離に拡張した．これにより，L1 距離といった別の距離尺度をより効果的に埋め込めることを確認した．

提案方式は理論的かつ萌芽的な内容である．しかし，「量子化されたコードに新たな意味を与える」というこれまでにない研究分野の確立への第一歩を踏み出したと言える．この後続の研究が続いていってほしいと考えている．

アウトリーチ活動

私は ACT-I 期間中に以下のアウトリーチ活動を行った

- 産業技術研究所における招待講演
- MIRU（CV分野国内最大の会議）におけるチュートリアル講演
- 東北大学における招待講演
- CVPR（CV 分野世界最大の会議）におけるチュートリアル講演の決定（2020/6 開催予定）

これらの多くの招待講演やチュートリアル講演は，多くの人々に対し私の研究や関連する探索分野を広く人々に知ってもらうために貢献できたと考える．特に，MIRU におけるチュートリアル内容は SpeakerDeck にて公開され、9.9K 回の閲覧されている．

https://speakerdeck.com/matsui_528/jin-si-zui-jin-bang-tan-suo-falsezui-qian-xian

また，CVPR におけるチュートリアルはメルカリ社との合同のものであり，社会との接点を広げられていると思う．今後も，そういった情報発信を進めていきたい．

3. 今後の展開

探索問題は大規模データ処理の際に必要な実用的かつ理論的は裏付けをもつ分野である．今後は，実際にそういった技術をもつ会社とのコラボレーションや，コンペティションの開催といった形で，社会に技術を還元していきたい．

4. 自己評価

本研究計画期間においては，応用面として「部分探索の探索」，理論面として「コードへの埋め込み」を達成することが出来た．このどちらも，インクリメンタルな研究というよりは，「意味のある問題」に対して取り組み，解答を示せたと思う．その意味で，十分な達成があった．また，様々なアウトリーチを持つなど，社会への波及効果は大きくあったと考えている．

一方で，「コードへの埋め込み」は最終的に「新しい理論的枠組みの第一歩」を踏み出すにとどまった．これをベースとして，より多くの理論的発展を本研究期間中に行えなかった面が心残りである．本研究をベースとした後続研究をこれから進めていきたいし，また周りからそういった研究が出てくることで分野として発展していけばよいと思う．

5. 主な研究成果リスト

（1）論文（原著論文）発表

（2）特許出願

研究期間累積件数：0 件

（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）

1. Yusuke Matsui and Shin’ichi Satoh, “Revisiting Column-Wise Vector Quantization for Memory-Efficient Matrix Multiplication”, IEEE International Conference on Image Processing, 2018.
2. Yusuke Matsui, Ryota Hinami, and Shin’ichi Satoh, “Reconfigurable Inverted Index”, ACM International Conference of Multimedia, 2018
3. 松井勇佑, “billion-scale の近似最近傍探索”, 招待講演（産業技術総合研究所）, 2019
4. 松井勇佑, “近似最近傍探索の最前線”, 第 22 回画像の認識・理解シンポジウム(MIRU), チュートリアル, 2019
5. 松井勇佑, “近似最近傍探索の最前線”, 招待講演（東北大学, 2020
6. Yusuke Matsui, Takuma Yamaguchi, and Zheng Wang, “Image Retrieval in the Wild”, Computer Vision and Pattern Recognition, Tutorial, 2020 (accepted)