1bit LLM の時代は来るのか，来ないのか，どっちなんだい？

徳永拓之 | LeapMind（株）

1bit LLM の時代が来る？

2024 年 2 月，The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits[1] というタイトルの論文が arXiv 上で公開され，にわかに話題となりました．“1.58 Bits”という表現はあまりなじみがありませんが，$\log_2(3) = 1.58\ldots$ ということで，パラメーターを三値にした場合の情報量を示しているようです．この論文（以下 b1.58 論文とする）は，同じ著者グループによる文献 2）を少し拡張したもので，大規模言語モデル（LLM）の効率化についての研究です．

本稿の前半ではこれらの論文の主張を解説し，後半ではその主張の妥当性について検討します．

なお，これらの 2 本の論文は，本稿執筆時点では，査読を経たものではありませんのでご注意ください．

ニューラルネットの量子化について

Transformer をはじめとする LLM によって自然言語処理（NLP）タスクの解析精度は劇的に向上しましたが，同時に，計算コストも飛躍的に増大しました．その計算コストを削減するための手段として，ニューラルネットワークの量子化には大きな期待が寄せられています．

情報理論における量子化とは連続的な量を離散的な値で近似することですが，ニューラルネットワークの分野においては，量子化とは，十分な精度で表現されていた離散的な値を，より少ないビット数，たとえば 8bit で表現することを指します．

離散的な値を使ってどうやって学習するのか，その詳細に立ち入ると必要なページ数が 3 倍以上に増えてしまうので，ここでは概略にだけ触れておきます．

現代のニューラルネットワーク量子化手法を大きく分けると，学習済みのモデルを量子化する方法と，量子化と学習を同時に進める方法があり，今回紹介する論文は後者に属します．

量子化しつつ学習を行う場合の最も大きな問題は，量子化関数の勾配を計算できないことで，そのため，近似的な勾配を用いて学習を行います．b1.58 論文の場合，量子化関数はあたかも存在しなかったかのように，勾配をそのまま素通ししています．このような手法を Straight-Through Estimator と呼びます．

b1.58 論文の概要

b1.58 論文では，LLM の中のリニア層（パラメーター行列と入力行列との行列積を行う層です）を量子化します．パラメーターが取り得る値を 3 種類（すなわち，{-1, 0, 1} のみです）に制限し，入力のそれを 8bit に制限します．これにより，乗算器が不要になり，加算器だけで行列積の計算が行えるようになる，というのが著者らの主張です．

LLaMA 3B モデルと，その量子化版である BitNet b1.58 3B モデルを比較し，パラメーター数が同じであるにもかかわらず，ARC-Easy などいくつかのタスクにおいて，量子化版モデルのほうが高い精度を示しました．当然ですが，パラメーター

数が同じであれば，モデルを保持するために必要なメモリ容量は量子化版モデルのほうが圧倒的に小さくなります．

700M，1.3B の 2 種類のモデルサイズではこのような逆転現象は起こっていなかったことから，「モデルを大きくすると精度の逆転現象が起こるのだとすると，量子化というのはこれまで想定されていたよりもかなり優れたアイディアなのではないか？」と，世間から注目を集めています．

その他，論文中には「BitNet b1.58 は，7nm チップ上での行列の乗算における算術演算のエネルギー消費を 71.4 倍節約します．（BitNet b1.58 saves 71.4 times arithmetic operations energy consumption for matrix multiplication on 7nm chips.）」「BitNet b1.58 は，モデルの性能と推論コストに関する新しいスケーリング則を可能にしています．（BitNet b1.58 is enabling a new scaling law with respect to model performance and inference cost.）」といった魅力的な文言が並んでおり，これらも注目を集めた理由と言えるでしょう．

精度の逆転現象は起き得るのか

前述の精度の逆転現象に類似する事例は，画像処理の分野で報告があります．文献 3）では，CIFAR-10 データセットを用いて量子化 ResNet を学習した実験では，パラメーターを量子化したほうがむしろ精度が向上する場合があることを報告しています．

しかし，このような逆転現象は，データセットが小規模で，なおかつ，ネットワークが十分に大きい場合に限られます．筆者はニューラルネットワークの量子化に関する論文を定期的にチェックしていますが，CIFAR-10 や SVHN といったかなり小規模なデータセット以外で，このような精度の逆転現象が見られたという報告を見たことがありません．

文献 4）では，Encoder-Decoder 型 Transformer のリニア層のパラメーターを二値化した場合に，既存の実数値モデルと類似のスケーリング則が成立することを実験によって示しています．複数のデータセットを用いた実験を行い，Google の社内データセットを用いた実験ではパラメーターを二値化すると精度が明確に落ちるが，WMT17 En-De dataset を使った実験では精度の落ち方がゆるやかであることから，前者のデータセットの方が難しく，学習時に大きなモデルキャパシティを必要とするのではないか，と推測しています．この考察は文献 3）での結果とも整合的です．

文献 5）は BART（こちらも Encoder-Decoder 型 Transformer の一種です）のリニア層のパラメーターを三値にし，入力を 8bit にした場合，いくつかの NLP タスクにおいて，0.5 ～ 1pt 程度の精度低下があったと報告しています．また，パラメーターを二値にした場合，三値の場合からさらに 1 ～ 1.5pt の低下があったと報告しています．

ここまでをまとめると，量子化の有無による精度の逆転現象は画像処理関連の既存研究でも見られますが，必要条件として，問題設定と比較してネットワークサイズが十分に大きな場合に限られます．また，これまで，LLM においては，量子化に関する研究はすでにいくつか行われていますが，精度の逆転現象が報告された事例はほかにはありません．

もっとも，b1.58 論文の主張は「700M モデルや 1.3B モデルでは精度の逆転は起きず，3B モデルでは逆転が見られた」という内容であり，既存研究との矛盾はありません．第三者による再現実装もすでにいくつか現れており，特に，文献 6）はコードだけではなくモデルウェイトも公開しています．しかし，b1.58 論文と同等の精度は再現できておらず，7 タスクの平均精度では，微妙に fp16 モデルに負けています．とはいえ，その差はわずかに 0.1pt です．モデルパラメーターが三値化されていることを考えると，大したものだと言えるでしょう．

果たして，より大きな 7B モデルや 30B モデルでは，逆転現象は起きるのでしょうか．結果はまだ誰も知りません．これは面白くなってきました．

71.4 倍の電力効率向上について

論文中には電力効率が 71.4 倍とありますが，この数字をどうやって出したのか，具体的な計算式はありません．引用されている文献 7）には，fp16（16bit 浮動小数点数）の加算器の電力が 160fJ，乗算器の電力が 340fJ，int8（8bit 整数）加算器の電力が 7fJ とあり，(160＋340)/7 ＝ 71.4 となるので，まず間違いなく，これらの数字が根拠になっているものと思われます．しかし，この計算式には，考慮されていない点がいくつかあります．

- -1 を掛けるためには二の補数の計算が必要になります．二の補数の操作にはビット反転と ＋1 の加算の回路が必要になりますから，結局，加算器と同程度のトランジスタが必要です．
- 乗算器が完全に不要になるわけではなく，-1 をかけるか，0 を出力するか，何もしないか，操作の選択を行う回路が必要になります．このためには選択操作を行うマルチプレクサと呼ばれる回路が，8 × 2 ＝ 16 個必要になります．
- 加算器が 8bit でよいのは初回の加算だけです．実際には加算は複数回行いますから，すべての加算を int8 加算器で実行することはできません．

そもそも，提案手法のメリットは推論時に現れるわけですから，推論でよく使われるビット幅で比較するべきでしょう．近年では，推論時の数値型は int8 や fp8 で十分であるということが明らかになってきています[8]．fp16 は比較対象としてやや甘いと言わざるを得ません．たとえば，int8 を比較対象にすると，文献 7）によると int8 の乗算は 70fJ，加算は上述の通り 7fJ ですから，電力効率向上の計算式は (7＋70)/7 ＝ 11 倍となります．

上述の乗算器まわりの回路は最低でも int8 加算器と同じ程度の規模になりますから，11 をさらに 2 で割って，5.5 倍程度が期待できる電力効率向上の上限になります．依然として大きなインパクトではありますが，71.4 倍とは比べるべくもありません．

既存プロセッサでの効果

ハードウェアを新しく作れば電力効率が向上することは分かりましたが，CPU や GPU などの既存プロセッサでの速度向上は望めるのでしょうか．実は，推論に関しては，高速化が期待できます．Transformer の推論時のボトルネックは DRAM 帯域です．このため，モデルサイズが小さくなれば，それはそのまま推論の高速化に貢献します．

一方，学習時には，fp16 もしくは fp32 でパラメーターを保持していて，そちらの値を更新していくので，データと計算が増える分，単純に遅くなります．

高速化は楽しいし役に立つ

さて，ここまで，b1.58 論文の中身について解説してきましたが，いかがでしたでしょうか．個人的には，この論文には賛否両論があると考えています．

否定的な見地からは，論文としての品質が十分に高いとは言えない点が挙げられます．ここまで，問題点をいくつか具体的に挙げてきましたが，ほかにも，先行研究として引用すべき文献，たとえば文献 4）や文献 5）が引用されていない点も気になります．

一方で，肯定的な見地からは，精度の逆転現象が本当ならば大きな発見であり，自然言語処理分野への大きな貢献となり得る，と言えます．もしこの実験結果が本当ならば実用上の効果は大きいですし，それだけではなく，この発見をきっかけにニューラルネットワークの性質についての新しい知見が得られる可能性もあります．

LLM に限らず，昨今のニューラルネットワークの計算量は，10 年前と比較しても圧倒的に大きくなってきています．ニューラルネットワークの高速化，効率化の価値は，以前よりもさらに増していると言えるでしょう．

筆者の所属する LeapMind では，これまで，量子化ニューラルネットワーク推論用の半導体 IP の

開発に取り組んできました．また，昨年（2023年）からLLMなどの生成AIの学習のための半導体チップの開発に取り組んでおり，こちらでも，低ビット表現を利用した効率化を図っています．筆者は，ニューラルネットワークの高速化や効率向上は，技術的に面白く，需要も急拡大しつつある，魅力的な分野であると考えています．しかしながら，残念なことに日本ではこういった技術の研究や開発に取り組んでいる会社や組織は多くありません．

本稿がニューラルネットワークの高速化への関心を喚起する一助となり，日本からも多くの挑戦者が現れることを期待しています．

参考文献

1) Ma, S., Wang, H., Ma, L., Wang, L., Wang, W., Huang, S., Dong, L., Wang, R., Xue, J. and Wei, F. : The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits (2024).
2) Wang, H., Ma, S., Dong, L., Huang, S., Wang, H., Ma, L., Yang, F., Wang, R., Wu, Y. and Wei, F. : Bitnet: Scaling 1-bit Transformers for Large Language Models (2023).
3) Zhu, C., Han, S., Mao, H. and Dally, W. J. : Trained Ternary Quantization, arXiv preprint arXiv:1612.01064 (2016).
4) Zhang, Y., Garg, A., Cao, Y., Lew, L., Ghorbani, B., Zhang, Z. and Firat, O. : Binarized Neural Machine Translation. Advances in Neural Information Processing Systems, 36 (2024).
5) Liu, Z., Oguz, B., Pappu, A., Shi, Y. and Krishnamoorthi, R. : Binary and Ternary Natural Language Generation, In Rogers, A., Boyd-Graber, J. and Okazaki, N. editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.65-77, Toronto, Canada (July 2023). Association for Computational Linguistics.
6) 1bit LLM/bitnet_b1_58-3b, Hugging Face, https://huggingface.co/1bitLLM/bitnet_b1_58-3B
7) Zhang, Y., Zhang, Z. and Lew, L. : PokeBNN: A Binary Pursuit of Lightweight Accuracy, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.12475-12485 (2022).
8) Baalen, M. v., Kuzmin, A., Nair, S. S., Ren, Y., Mahurin, E., Patel, C., Subramanian, S., Lee, S., Nagel, M., Soriaga, J. and Blankevoort, T. : FP8 Versus INT8 for Efficient Deep Learning Inference (2023).

（2024年3月15日受付）
（2024年4月4日note公開）

徳永拓之　tokunaga@leapmind.io

大阪大学基礎工学部システム科学科卒業，東京大学大学院情報理工学系研究科修了．数社でエンジニアとして働いた後，2018年よりLeapMind（株）にて取締役CTO．機械学習とカレーが好き．