H-040

BitNet を用いた Vision Transformer の構築

Construction of Vision Transformer using BitNet

井波 辰朗[1)]　神野 健哉[1)]
Tatsuro Inami　Kenya Jin'no

概要

近年,LLM は自然言語処理において高い処理能力を得ているが, その反面膨大な計算量を有しており学習を行う事自体が非常に困難という問題を抱えている. それに対し,Wang らは Transformer の全結合層を{-1, 0, 1}の重みを持つ BitLinear に置き換えた BitNet を考案した. そこでは元の Transformer に対して精度を落とすことなくモデルサイズや計算量を大幅に削減することが出来ると発表されている. そこで, 本研究では Transformer を用いた画像分類モデルである Vision Transformer においても BitLinear が有効なのを実験により検討し, 深層学習におけるモデルサイズと量子化について考察する.

1　まえがき

近年における深層学習は膨大なデータセットをモデルに学習させることによって, 画像分類や翻訳のみならず, 画像生成や文章生成においても高い精度を出すことが可能となっている. しかし, それらの学習には膨大な計算が必要であり, それらの計算は GPU を中心とした多くの計算機が必要になる. これにより, モデルの学習環境における経済的なコストや消費電力による環境的なコストが問題になっている.

それに対し,Wang らは自然言語処理モデルの一つである Transformer[3] の重み部分を{-1, 1}の 2 値で表現した BitNet を発表した. そこでは, 重みを{-1, 1}の 2 値に制限することで計算量を減少させ, 計算エネルギーの減少が可能であると主張されている [1]. また, その後, 重み部分を{-1, 0, 1}の 3 値で表現した BitNet b1.58 を発表し, よりモデルサイズの大きい LLaMA モデルにおいても計算エネルギーの減少が可能であると主張している [2].

BitNet は深層学習モデル内における全結合層 (nn.Linear 層) を量子化した重み層である BitLinear に置き換えたモデルである. 本研究では Transformer における BitNet 及び BitLinear の有効性を元に,Transformer を画像分類に応用した Vision Transformer[4] においてもそれらが有効であるかを CIFAR-10 データセットを用いた画像分類により検討を行う.

1)　東京都市大学大学院 総合理工学研究科 情報専攻 Informatics, Graduate School of Integrative Science and Engineering, Tokyo City University

2　BitNet[1] [2]

BitNet は Transformer 内にある全結合層を BitLinear に置き換えた自然言語処理モデルである.Wang らは BitNet を用いることで学習や推論における計算エネルギーを減少させることが可能であると主張している. 図 1 に BitLinear と BitNet のモデル概要を示す.

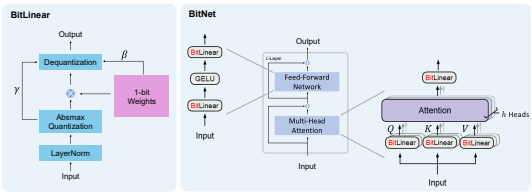

図 1: BitLinear と BitNet のモデル概要 [1]

3　BitLinear

BitLinear は BitNet を内で用いられている{-1, 1}の 2 値もしくは{-1, 0, 1}の 3 値の重みを持った全結合層である. 本研究では{-1, 0, 1}の 3 値を持った層を用意しそれらを実験に用いる.

BitLinear を用いたモデルの学習は小数点を含んだ通常の重み (W_{nm}) と 3 値に量子化された重み ($\tilde{W}$) を用いて行われる. 順伝搬時には通常の重みから計算された量子化後の重みが使用され逆伝搬時には通常の重みが損失関数によって求められた誤差によって更新される. これにより学習時の計算エネルギーは BitLinear を用いないモデルとあまり変わらないものの, 推論時には量子化された重みのみが使用されるため計算の高速化やエネルギーの減少が可能である主張されている. 量子化計算は以下の通りである.

$$\tilde{W} = \mathrm{RoundClip}\left(\frac{W}{\gamma+\epsilon}, -1, 1\right) \tag{1}$$

$$\mathrm{RoundClip}(x, a, b) = \max(a, \min(b, \mathrm{round}(x))) \tag{2}$$

$$\gamma = \frac{1}{nm}\sum_{ij}|W_{ij}| \tag{3}$$

4　Vision Transformer[4]

Vision Transformer は Dosovitskiy らによって発表された画像分類モデルである. このモデルは ResNet [5] や VGG [6] といった一般的な画像分類モデルと異なり, 畳み込み層を用いず全結合層のみを用いて画像分類を行う. 具体的には, 自然言語処理モデルである Transformer のエンコーダを元にモデルが作られており,Attention(注

意機構) という処理機構を持っている.Attention は全結合層のみで構成されており, 重みの種類を増やしそれぞれの出力を再び処理することで精度の向上に貢献していると考えられている.

また,Vision Transformer は画像をパッチに分割し, クラスと位置を埋め込んだ特徴ベクトルを生成した後, エンコーダを用いた画像分類を行うため画像の位置情報を保持したまま処理が可能であると主張されている.

本研究では Transformer における BitNet の有効性を基に、同様の機構を持つ Vision Transformer においてもそれらが有効であるかの検討を行う.

図 2 に Vision Transformer のモデル概要を示す.

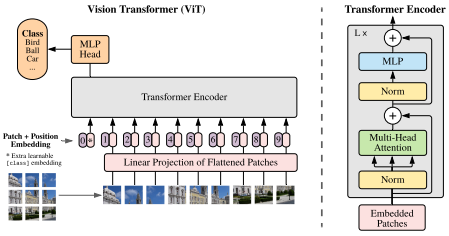

図 2: Vision Transformer のモデル概要 [4]

5 実験

本研究では Vision Transformer における BitNet 及び BitLinear の有効性を検討するために,CIFAR-10 データセットを用いた画像分類を行う. 表 1 のパラメータにより用意された Vision Transformer モデルと, 最後の分類層を除いたすべての全結合層を BitLinear に置き換えたモデルを用意し訓練データとテストデータそれぞれで学習を行う. そこで得られた損失誤差と画像分類の正答率を比較する. なおパッチ分割する際の画像サイズを Patch size,TransformerEncoder の数を Layers と表記する.

表 1: 実験に用いるパラメータ

Epochs	100
Loss	Cross-entropy Loss
Optimizer	Adam
Learning Rate	0.001
Patch size	16
Layers	6

6 結果

実験によって得られた 1epoch あたりの損失誤差 (Loss) と正答率 (Accuracy) を図 3 に示す.

訓練データでの損失誤差と正答率において,BitLinear を用いたモデルは通常のモデルよりもわずかに性能が劣るものの, その差は統計的に有意ではない. これは,BitLinear による量子化がモデルの学習能力に大きな影響を与えないことを示唆している. 対してテストデータでは損失関数は両モデルとも途中から増加してしまい, 正答率においても 5 割程度に留まってしまった. しかし,BitLinear を用いたモデルは通常のモデルより低い誤差が出ていることが分かる.

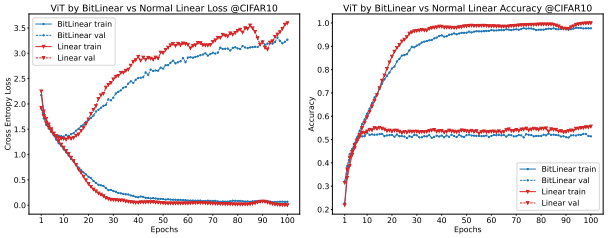

図 3: 1epoch あたりの損失誤差と正答率

7 まとめ

実験結果より, 訓練データによる両モデルの損失や正答率がほぼ変わらなかったことから限られた重みを持つ BitLinear を用いても学習が可能であり, 深層学習モデルにおける計算エネルギーの減少が可能であることを示唆している. また, テストデータにおける誤差が通常のモデルよりも低い値になったことから BitLinear は重みにおける表現力を保持したまま過学習を防ぐ役割があることがが考えられる. テストデータにおける精度低下においては dropout 層の追加やパラメータの調整により改善が見込めるため今後の課題として検討を行う. また, 推論時の具体的な計算エネルギーやメモリサイズは実行環境の調整が必要であるためこれらも合わせて今後の研究を行う.

謝辞

本研究の一部は JSPS 科研費 23K11266, 23H03387, 24K15115, 東北大学電気通信研究所共同プロジェクト研究, 東京都市大学重点推進研究未来知能ユニットの助成によるものです.

参考文献

[1] Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, and Furu Wei. Bitnet: Scaling 1-bit transformers for large language models. arXiv preprint arXiv:2310.11453, 2023.

[2] Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, and Furu Wei. The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits. arXiv:2402.17764 ,2024

[3] Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. In International Conference on Learning Representations (ICLR, 2021).

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[5] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[6] Simonyan, K. "Very deep convolutional networks for large-scale image recognition." Proc ICLR (2015).