低ランク行列分解を用いたFFN 活性化刈り込みによるViTの推論高速化の検討 ~ViT版Deja Vuの開発~

伊藤 悠馬[1], 関川 雄介[2], 池畑 諭[2,3], 佐藤 育郎[1,2]

[1]東京科学大学, [2]デンソーITラボラトリ, [3]国立情報学研究所

Recognition, Control and Learning Algorithm Lab.

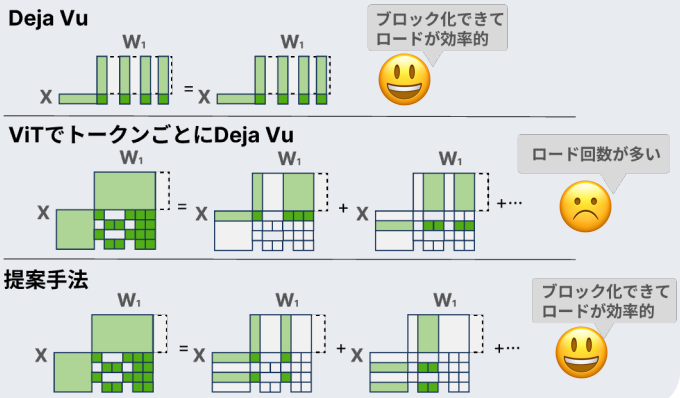
Institute of SCIENCE TOKYO / DENSO IT LAB / NII 国立情報学研究所

1.導入

背景
- 自動運転等の応用において、ViTの推論高速化には高いニーズがある
- 言語モデルに現れる活性化のスパース性(**文脈スパース性**)を利用した動的プルーニングにより6倍程度の推論高速化が報告されている (Deja Vu, ICML'23)

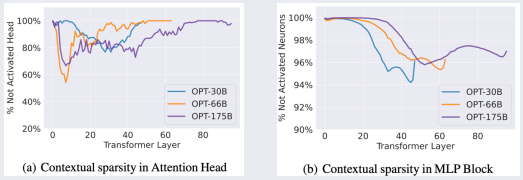
動機
- 文脈スパース性を利用した動的プルーニングにより、ViTの推論を高速化したい

研究概要
- FFNの中間層の活性化をGPU上で実行可能な低ランクの形に分割するアルゴリズムを開発
- 2値低ランク行列分解を用いることで、ロード回数を削減しながら、下流タスクの精度低下を抑えた

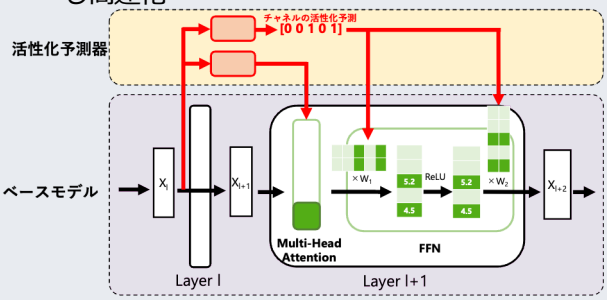


2.従来手法: Deja Vu [Liu et al., ICML, 2023]

概要
- 言語モデルにおいて文脈スパース性を活用した動的プルーニング手法
- 大規模言語モデルでは、アテンションヘッドにおいて約80%、FFNにおいて約95%が非活性である事例が知られている



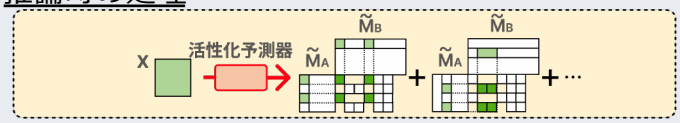
手法
- 各Transformerレイヤーごとに非活性な特徴次元を2層前の特徴から予測
- 予測器を用いてベースモデルの推論をプルーニングし高速化



3. 提案手法

予測器の学習データ作成

推論時の処理



$$FFN(X) \approx FFN_{masked}(X, \tilde{M}_A, \tilde{M}_B)$$
$$M \approx \tilde{M}_A \tilde{M}_B$$
$$M = Top\text{-}k(GELU(XW_1))$$
$$\tilde{M}_A = \widehat{M}_A > \tau, \quad \tilde{M}_B = \widehat{M}_B > \tau$$
$$\widehat{M}_A, \widehat{M}_B = \underset{M_A, M_B}{\operatorname{argmin}} WMSE(M, \sigma(M_A)\sigma(M_B))$$

FFN_{masked}: マスクをもとにスパースにFFNを計算する関数
$Top\text{-}k$: 絶対値の上位k個を$True$とする関数
$WMSE$: 重みづけ二乗誤差関数
σ: シグモイド関数
$X \in R^{t \times d}$: FFNの入力特徴料
$W_1 \in R^{d \times d_{ffn}}$: FFNの重み
$\widehat{M}_A, \widehat{M}_B \in (0,1)^{t \times k}$: 活性化マスク分解後の行列
$\tau \in R$: 閾値

- FFNの活性化関数GELUの出力の絶対値が下位の要素を刈り込むマスクを作成する
- 重みづけ二乗誤差により重要な要素を保持
- 行列積の実行時には$\widehat{M}_A, \widehat{M}_B$をもとに計算

重みづけ二乗誤差の設計
- 絶対値の上位40%の部分の喪失を大きく重みづけ
- GELUの出力の絶対値に比例して重みづけ

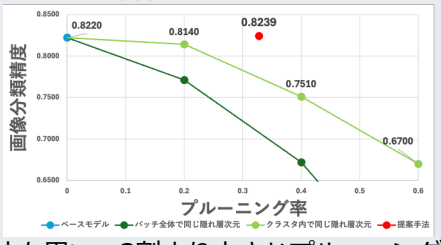
4.実験

FFNのプルーニング率 VS 画像分類精度

ベースライン
- ベースモデル: ViT-large(304M params)
- パッチ全体で同じ隠れ層次元を計算
- 同じパッチクラスタ内で同じ隠れ層次元を計算

実験設定
- 活性化マスクは予測器を使用せず、GELU出力を直接利用
- GELU出力の絶対値において各パッチの上位40%をTrueとし活性化マスクを作成
- ImageNet-1kの評価データセットからランダムで1100サンプルを利用し評価

結果



- 提案手法を用い、3割より大きなプルーニング率でベースモデル相当の性能を達成

5.今後の計画
- マスクをもとに高速にFFNを計算するFFN_{masked}の実装
- $\tilde{M}_A, \tilde{M}_B$を予測するアルゴリズムの設計

本研究は東京科学大学「DENSO IT LAB 認識・制御・学習アルゴリズム共同研究講座」における研究プロジェクト「機械学習の実応用に向けた品質基盤開発」の成果である