Reactive Bias による ViT への人の知見の組み込みに関する研究

TR23002 鈴木雅司　　指導教授：藤吉弘亘

1. はじめに

AI の信頼性や性能向上を目的として，AI の学習過程に人が介入する Human-in-the-loop アプローチの研究が注目されている．三津原らは，Attention 機構の推論に影響を与える能力に着目し，CNN ベースの Attention Branch Network (ABN) に人の知見の組み込む手法を実現した．一方，Transformer ベースの ViT に人の知見を組み込む手法は，これまで実現されていない．そこで，本研究では ViT への人の知見の組み込みを実現するために，人の知見を学ぶためのバイアスである Reactive Bias を提案する．

2. Vision Transformer (ViT)

ViT は [2]，自然言語処理の手法である Transformer を画像認識タスクに応用した手法である．ViT は Multi-Head Self Attention と Multi Layer Perceptron，Layer Normalization により構成される Transformer block を複数重ね合わせた構造を持つ．Transformer block を定式化すると式 (1) のようになる．

$$\begin{aligned} \mathbf{z}'_l &= \mathrm{MHA}(\mathrm{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1} \\ \mathbf{z}_l &= \mathrm{MLP}(\mathrm{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l \end{aligned} \tag{1}$$

ここで，MHA は Multi-Head Self Attention，LN は Layer Norm，MLP は Multi Layer Perceptron を表す．$\mathbf{z}'_l$，$\mathbf{z}_l$ はそれぞれ l 層目の Multi-Head Self Attention および Transformer block の出力を表す．Multi-Head Self Attention を定式化すると式 (5) のようになる．

$$\mathbf{s}_h(\mathbf{x}) = \alpha(\mathbf{x})(\mathbf{x}\mathbf{W}^V) \tag{2}$$

$$\alpha(\mathbf{x}) = \mathrm{softmax}(\mathbf{e}(\mathbf{x})) \tag{3}$$

$$\mathbf{e}(\mathbf{x}) = \frac{(\mathbf{x}\mathbf{W}^Q)(\mathbf{x}\mathbf{W}^K)^T}{\sqrt{d_h}} \tag{4}$$

$$\mathrm{MSA}(\mathbf{x}) = \mathrm{Concat}\left(\mathrm{s}_1(\mathbf{x}), \mathrm{s}_2(\mathbf{x}), ..., \mathrm{s}_H(\mathbf{x})\right) \mathbf{W}^O \tag{5}$$

ここで，$\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ はそれぞれ入力を Query, Key, Value に変換するための線形射影，$\mathbf{W}^O$ は結合後の Self Attention へ重み付けをするための線形射影，d_h は各 Self Attention における入力特徴のチャンネル数，h は特定の head，$\mathbf{s}$ は Self Attention を表す．式 (4) に示すように Self Attention では，Query と Key の内積を用いて Value を重み付き和を計算する構造を持つ．これにより，Self Attetnion は入力特徴の間の類似度を基にして空間方向の特徴を大域的に抽出することが出来る．また，MLP は Transformer block 毎に全ての Token に対して同じ線形層を用いてチャンネル方向の特徴を抽出する．これらの処理により，ViT は Multi-Head Self Attention と MLP を繰り返すことで Token の空間方向とチャンネル方向の特徴を抽出する．

3. 提案手法

本研究では，ViT の各ブロックの Self Attention に人の知見をバイアスとして組み込む Reactive Bias を提案する．Reactive Bias のチューニングを通じて，人の知見である注目領域を ViT に組み込むことを実現する．三津原ら [1] の研究では，Attention Branch Network (ABN) の Attention 機構に対して人の知見をモデルへ組み込んでいる．そこで，Attention 機構と似た構造を持つ ViT の各層の Self Attention に人の知見を組み込むことを考える．しかし，ABN では 1 つのモデルに対して Attention 機構を 1 つ用いるのに対して，ViT の Self Attention は各ブロックの各 head に用いており，捉える特徴もそれぞれ異なる．そのため，ViT へ人の知見を組み込む場合は，人の知見を各 Self Attention に適した形に変換する必要がある．そこで，図 1 に示すように，Reactive Bias はサンプル毎の人の知見を組み込む役割を持つ Reactive mode と各 Self Attention のバイアスを学習する Learnable Bias から構成される．Reactive Bias は Learnable Bias のパラメータを基に各 Self Attention に適した形で人の知見を組み込む．

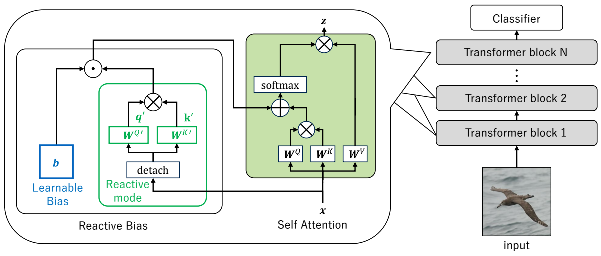

図 1: Reactive Bias アーキテクチャ

3.1. Learnable Bias

Learnable Bias は，データセットに対する Self Attention のバイアスを学ぶ役割を持ち，ファインチューニングを通じて各 Self Attention がどの領域を注視するか等のバイアスを獲得する．Learnable Bias のチューニングは，図 2(a) に示すように，Self Attention 内の Query と Key の内積に学習可能パラメータを加算によって実現する．ここで，Learnable Bias を $\mathbf{b} \in \mathbb{R}^{HW \times HW}$，入力を $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$，Self Attention の Query と Key の内積を $\mathbf{e} \in \mathbb{R}^{HW \times HW}$ とすると，Learnable Bias は式 (6) のようになる．

$$\mathbf{e}(\mathbf{x}) = \frac{(\mathbf{x}\mathbf{W}^Q)(\mathbf{x}\mathbf{W}^K)^T + \mathbf{b}}{\sqrt{d_h}} \tag{6}$$

式 (6) より，Learnable Bias は特定の head における Self-Attentnion の重みと Learnable Bias との要素和で表せる．これにより，Learnable Bias は Self Attention の重みと同一の勾配を用いてパラメータを更新できるため，Self Attention のデータセットに対する注視領域を獲得できる．

3.2. Reactive mode

Reactive mode はサンプル毎に人の知見を学習するために，入力に反応した値を出力するモジュールである．Self Attention は入力特徴を線形射影した Query と Key の内積により重みを計算するため，入力サンプルによって重みが変化する．そのため，人の知見を学習する場合にも，入力サンプルに対応したモジュールを構築する必要がある．そこで，Reactive mode は Self Attention における Query, Key の計算を模倣した構造とする．Reactive mode を $\mathbf{r} \in \mathbb{R}^{HW \times HW}$ とすると，Reactive mode は式 (7) のようになる．

$$\mathbf{e}(\mathbf{x}) = \frac{(\mathbf{x}\mathbf{W}^Q)(\mathbf{x}\mathbf{W}^K)^T + \mathbf{b} \odot \mathbf{r}(\mathbf{x})}{\sqrt{d_h}} \tag{7}$$

$$\mathbf{r}(\mathbf{x}) = (\mathbf{x}\mathbf{W}^{Q'})(\mathbf{x}\mathbf{W}^{K'})^T \tag{8}$$

ここで，$\mathbf{W^{Q'}}, \mathbf{W^{K'}}$ は Reactive bias における Query, Key の線形層を表す．式 (7) に示すように，Reactive Bias は Learnable Bias と Reactive mode の単純な要素積で示せる．

3.3. 人の知見の組み込み

図 2 に Reactive Bias の導入によるチューニング時の更新箇所を示す．Reactive Bias への人の知見の組み込みは Learnable Bias を用いて行うため，Reactive Bias は Learnable Bias と Reactive mode を段階的に学習する．Learnable Bias のチューニング時は，図 2(a) に示すように Learnable Bias のみを導入し，モデル全体をファインチューニングする．Reactive mode のチューニング時は，各 Self Attention に適した形で人の知見を組み込むために，図 2(b) に示すように各層で人の知見を基に作成したマスクと Learnable Bias を用いてチューニングを行い，Self Attention，Reactive mode，分類器のパラメータを更新す

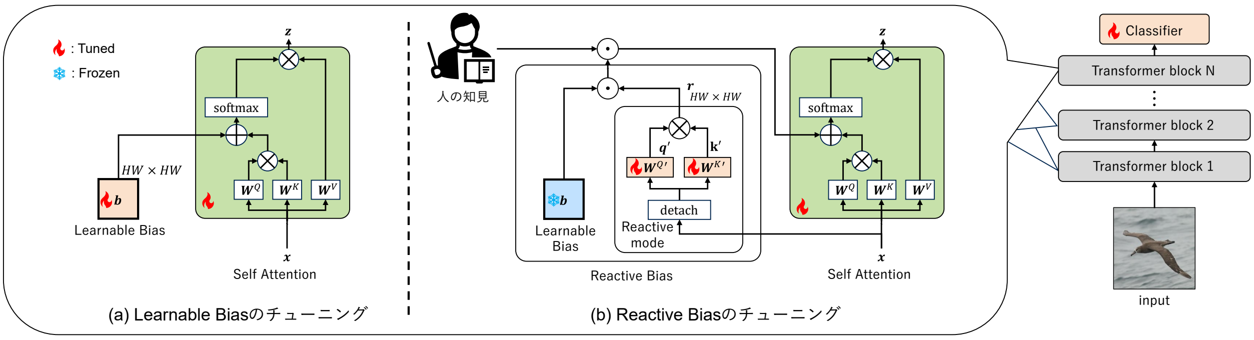

図 2: 提案手法のチューニング

表 1: 人の知見の組み込みによる精度変化の確認

Model	Reactive Bias	Human knowledge	CUB-200-2010 Acc [%]	CUB-200-2011 Acc [%]
ViT-T			64.39	80.77
	√	√	**67.00**	**81.27**
ViT-S			69.77	83.88
	√	√	**72.30**	**84.26**
ViT-B			70.33	84.76
	√	√	**72.63**	**85.86**

る．人の知見を用いて作成したマスクを $\mathbf{M} \in \mathbb{R}^{HW}$，マスクを導入した Reactive Bias を $\mathbf{B} \in \mathbb{R}^{HW \times HW}$ とすると，Reactive mode の学習時の勾配は式 (10) のようになる．

$$\mathbf{B}_{ij}(\mathbf{x}) = \mathbf{b}_{ij} \odot \mathbf{r}(\mathbf{x})_{ij} \odot \mathbf{M}_j \tag{9}$$

$$\frac{\partial \mathbf{B}(\mathbf{x})}{\partial \mathbf{W}^{\mathbf{Q}'}} = \sum_{i=1}^{HW} \mathbf{x}\mathbf{W}^{\mathbf{K}'} \odot \mathbf{b}_i \odot \mathbf{M} \odot \mathbf{x}_i$$

$$\frac{\partial \mathbf{B}(\mathbf{x})}{\partial \mathbf{W}^{\mathbf{K}'}} = \sum_{j=1}^{HW} \mathbf{x}\mathbf{W}^{\mathbf{Q}'} \odot \mathbf{b}_j \odot \mathbf{M}_j \odot \mathbf{x}_j \tag{10}$$

式 (10) より，$\mathbf{W}^{\mathbf{Q}'}$ はマスクされていない領域のトークンのみ Learnable Bias を基にパラメータを更新し，$\mathbf{W}^{\mathbf{K}'}$ は人の知見によってマスクした Query と Learnable Bias を基にパラメータを更新する．これにより，Reactive mode は人の注視領域に着目した学習が可能となり，Self Attention に適した形で人の知見を ViT へ組み込むことが可能となる．

4. 評価実験

本章では，提案手法の有効性を検証するために公開データセットを用いた人の知見の導入による精度向上の検証実験を行う．事前学習モデルには ImageNet-1k を用いて学習した DeiT [3] を利用し，評価実験ではベースモデルに ViT-T,ViT-S,ViT-B を用いる．人の知見の導入による評価には，鳥のデータセットである CUB-200-2010 および CUB-200-2011 と，人の知見を集めたデータセット bubble 情報や Gazed Human Attention (GHA) を用いたチューニングによる精度と説明性の比較を行う．

4.1. 評価結果

表 1 に人の知見と Reactive Bias を導入した時の精度を示す．表 1 より，通常の ViT と Reactive Bias を導入した ViT を比較すると，全てのモデルサイズにおいて Reactive Bias を用いることによる精度向上が確認できる．このことから，認識精度において人の知見の組み込みは有効であることが確認できた．

4.2. アテンションマップの比較

説明性の評価には Attention Rollout と RISE を使用し，ViT および Reactive Bias を導入した ViT のアテンションマップを比較する．Attention Rollout は Self Attention の重みを用い，RISE は入力画像へマスクを施し出力されるスコアからアテンションマップを可視化する．そのため，Attention Rollout は推論の過程で注目している領域を，RISE は画像空間上で推論に有効な領域を示す．

アテンションマップの可視化結果を図 3 に示す．図 3 は左から入力画像，人の知見，通常の ViT と Reactive Bias

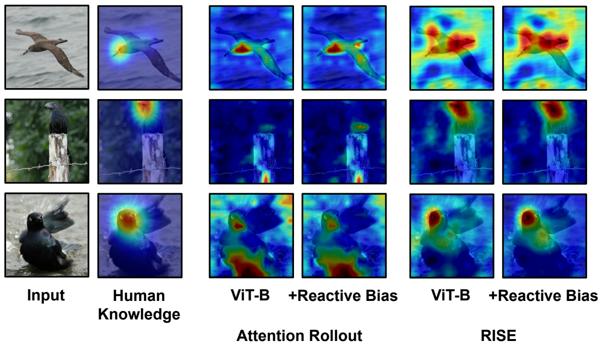

図 3: アテンションマップの可視化結果

を追加した ViT の Attention Rollout 及び RISE のアテンションマップを示す．図 3 より，Attention Rollout を用いた場合，通常の ViT と比較して Ractive Bias を用いた ViT のアテンションマップは物体上を注目する傾向が確認できる．これは，推論時に物体上の特徴をより強く推論に用いるように学習したことを表す．また，RISE を用いて可視化を行った場合，上から 2 枚目と 3 枚目のサンプルにおいて人の知見の組み込みによりアテンションマップが人の知見に近づいていることが確認できる．これは，入力画像中の人の知見の示す領域を重要視するように学習していることを示し，ViT への人の知見の組み込みが行えていることを示す．以上のことから，Reactive Bias による人の知見の組み込みは認識精度及び説明性の改善に有効であると確認できる．

5. おわりに

本研究では，ViT への人の知見を組み込みの実現を目指して，ViT の各層の Self Attention へ知見を導入する Reactive Bias を提案した．評価実験により，人の知見を用いた学習において提案手法は精度及び説明性の向上に有効であることを示した．しかし，Attention weight のアテンションマップにおける変化が少量であった．そのため今後は，より説明性の向上を得るために，学習条件の探索を行う．

参考文献

[1] M. Mitsuhara, *et al.*, “Embedding Human Knowledge into Deep Neural Network via Attention Map”, VISAPP, 2021.

[2] A. Dosovitskiy, *et al.*, “An image is worth 16x16 words:Transformers for image recognition at scale”, ICLR, 2021.

[3] H. Touvron, *et al.*. “Training data-efficient image transformers & distillation through attention”, ICML, 2021.

研究業績

[1] 鈴木雅司 等, “Attention Branch Transformer：Top-down Attention 機構を用いた頑健な ViT の提案”, 画像の認識・理解シンポジウム, 2023.

[2] 鈴木雅司 等, “Reactive Bias を用いた ViT への人の知見の組み込み”, 画像の認識・理解シンポジウム, 2024.