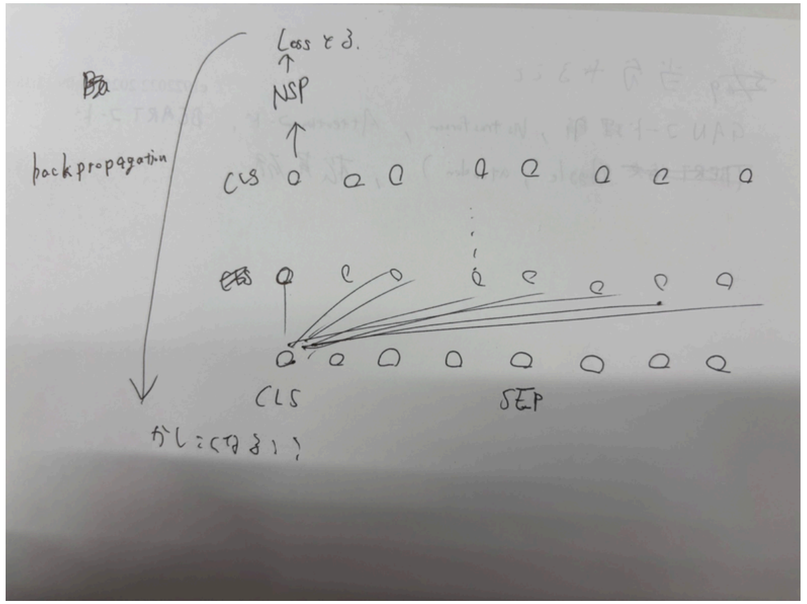
疑問



1. CLSトークンの学習の仕方は？
→△self-attentionで学習?
→BERTではNSPなどでは正解クラスと比較して誤差逆伝番して求める?
2.式１～４これどこでどうやって使うの？
→transformer encoderで使います
3.inductive biasって何？？
→極力ないほうがいいの？？
→後天的性質的な？
4.fine-tuning時の位置埋め込みはそのままpre-trainingの位置埋め込みを使うってこと？
→Yes?
5,normが違う理由は？
→画像をembeddingしたときにベクトルがあれそう

2025.05.13 個人ゼミ

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

大阪工業大学大学 ロボティクス&デザイン工学部　システムデザイン工学科

工藤滉青　瀬尾昌孝

アウトライン

1. ABSTRACT
2. INTRODUCTION
3. RELATED WORK
4. METHOD

ABSTRACT

要旨

1. これまでの課題

コンピュータービジョンへの応用が限られている。
ビジョン領域ではattentionはCNNと組み合わせて使うかCNNの一部の構成要素を置き換える形で使用されますが、全体的な構造は維持されることがほとんど。

2. 解決案

画像パッチの列に対して直接Transformerを適用する

INTRODUCTION

はじめに（１/２）

コンピュータービジョン

- 畳み込みが主流→NPLの成功→self-attentionへの関心が高まる
- Ramachandran
  - CNNを完全に置き換え
  - 特殊な注意機構を使用しているため、現代のハードウェアアクセラレータ上ではまだ効果的にスケーリングできなかった。

研究への応用案

- 標準的な Transformer を可能な限り最小限の変更で画像に直接適用する。
- そのために、画像をパッチに分割し、それらのパッチの線形埋め込みの列を Transformer への入力として与える

INTRODUCTION

はじめに（２/２）

結果

- ImageNet のような中規模データセット
  - 同等のサイズの ResNet よりも数ポイント低い精度
  - 原因：Transformer がCNNのinductive biases(equivarianceとlocality)を欠いたため。
- 大規模なデータセット（1400万～3億枚の画像）
  - CNNのinductive biasesよりも優れた。

RELATED WORK

関連研究

参考にした研究

- パッチを用いた完全なselfattentionモデル（ Cordonnier 2020）
- CNNとselfattentionの様々な形の組み合わせたモデル（Bello 2019,Hu , 2018；Carion , 2020、Wu , 2020など）
- 画像の解像度と色空間を削減した上で、画像ピクセルにTransformerを適用したモデル：imageGPT

METHOD

ViTモデル（１/２）

モデル構造

- 入力値をreshape
  - $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ から $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$ に変換
  - 計算量を軽くするため（my thoughts）
  - 学習がそれぞれでできるため（my thoughts）
    - multi-head-attetntion的の利点と似てるかも？
- [class] tokenの配置（ほぼBERTと同じ）
- 位置情報はパッチ埋め込みにポジション埋め込みを加算
- transformer encoderの詳しい構造は次のページ

METHOD

ViTモデル（２/２）

transformer encoderについて

- ほとんとattetnion is all you needのencoder部分
- FeedForword Networkではなく MLP
- Normの位置が違う

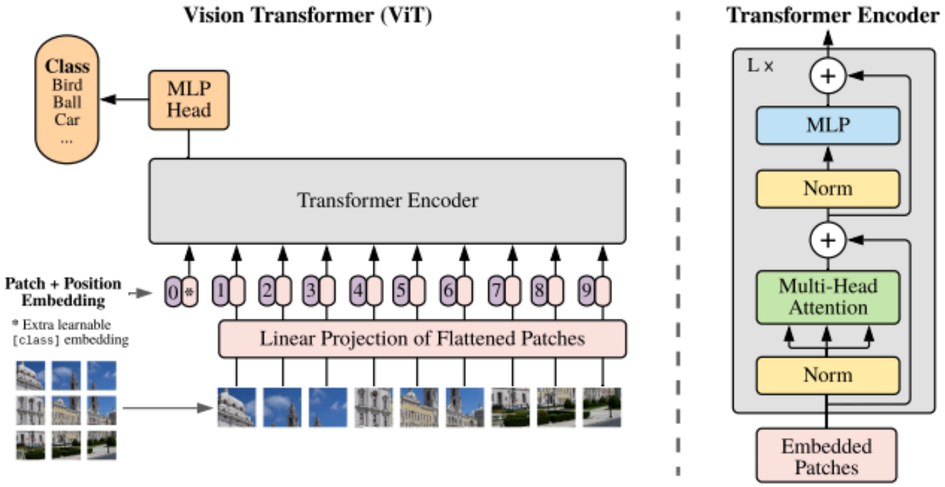
The MLP contains two layers with a GELU non-linearity.

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$$

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \ldots L$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \qquad \ell = 1 \ldots L$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0)$$



METHOD

Inductive biasとHybrid Architecture

Inductive bias(CNNと比較して少なめ)

- MLPのみlocalとtranslationally equivariant
- self-attentionはglobal
- two-dimentional neighborhood
    - 画像をパッチに切り分ける時
    - 異なる解像度の画像に対する位置埋め込みを調整するためのfine-tuning時

Hybrid Architecture

- 生画像パッチの代わりに、CNN（LeCun et al., 1989）で得られた特徴マップを系列入力として用いる手法

METHOD

FINE-TUNING AND HIGHER RESOLUTION

Fine-TuningとPre-Training

- higher resolutionをpre-trainingで使うよりもfine-tuningで用いるほうが性能向上につながる。（らしい）
    - pre-trainingで高解像度の画像を使ったら本当に必要な情報を取得するときに余計なものをとってしまうから？？
- 高解像度の画像を入力する場合でもパッチサイズは変えないため、系列長が大きくなる。
- Vision Transformer はメモリ制約の範囲内で任意の系列長を処理できるが、pre-training時の位置埋め込みは解像度が変わると意味をなさなくなる可能性がある。
    - pre-trainingとfine-tuningとで同じ位置埋め込みを使うから？