

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？？

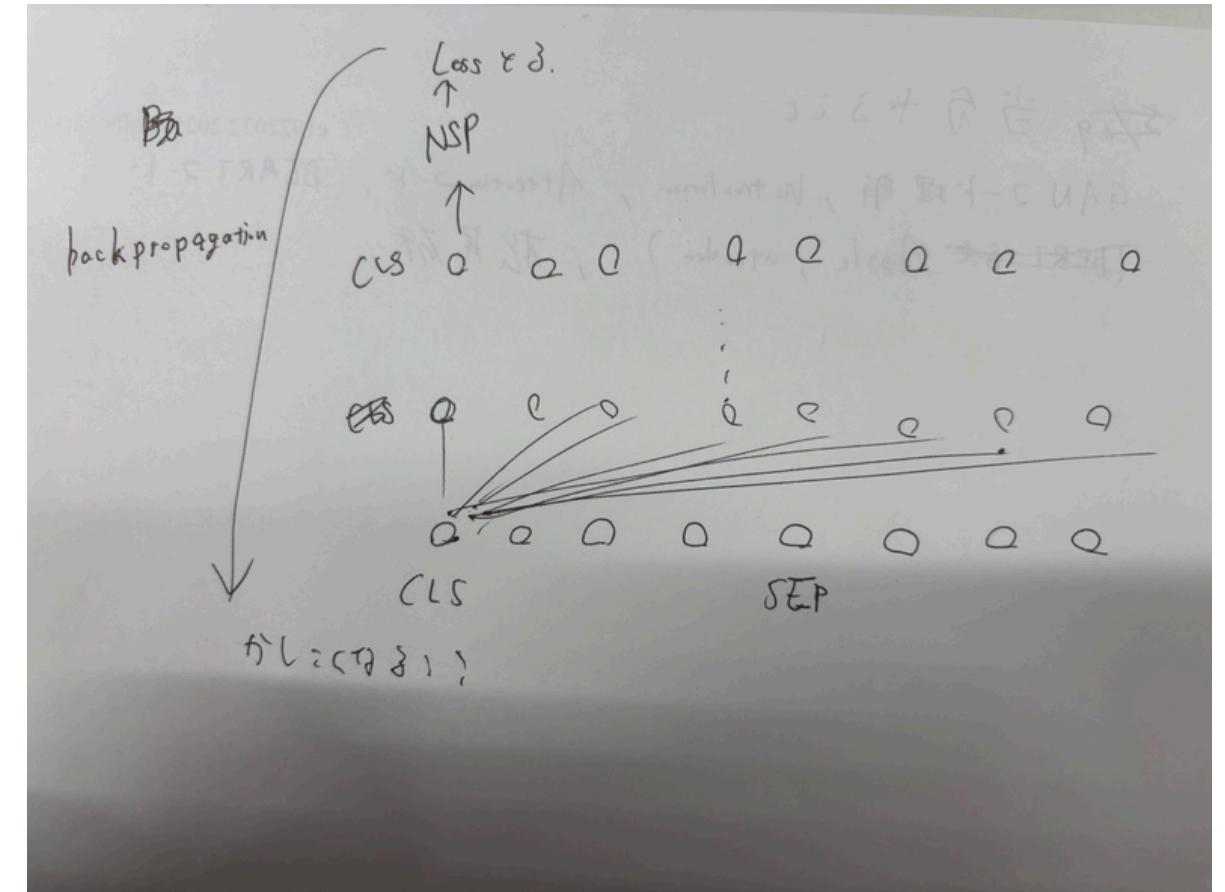
→後天的性質的な？

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？

→(先生)normはそもそもいらん情報を排除する（例えば輝度など）操作なので先にしてしまってからattentionを使う前



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 濑尾昌孝

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

- 置み込みが主流→NLP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-attention的の利点と似てるかも？
 - (先生) 文脈情報として一枚の画像での関連度を見つけるため
- [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

- ほとんどattention is all you needのencoder部分
- FeedForward NetworkではなくMLP
- Normの位置が違う

(先生) Eq1のEは埋め込み： $P^2 \times C$ をD次元に落とし込む

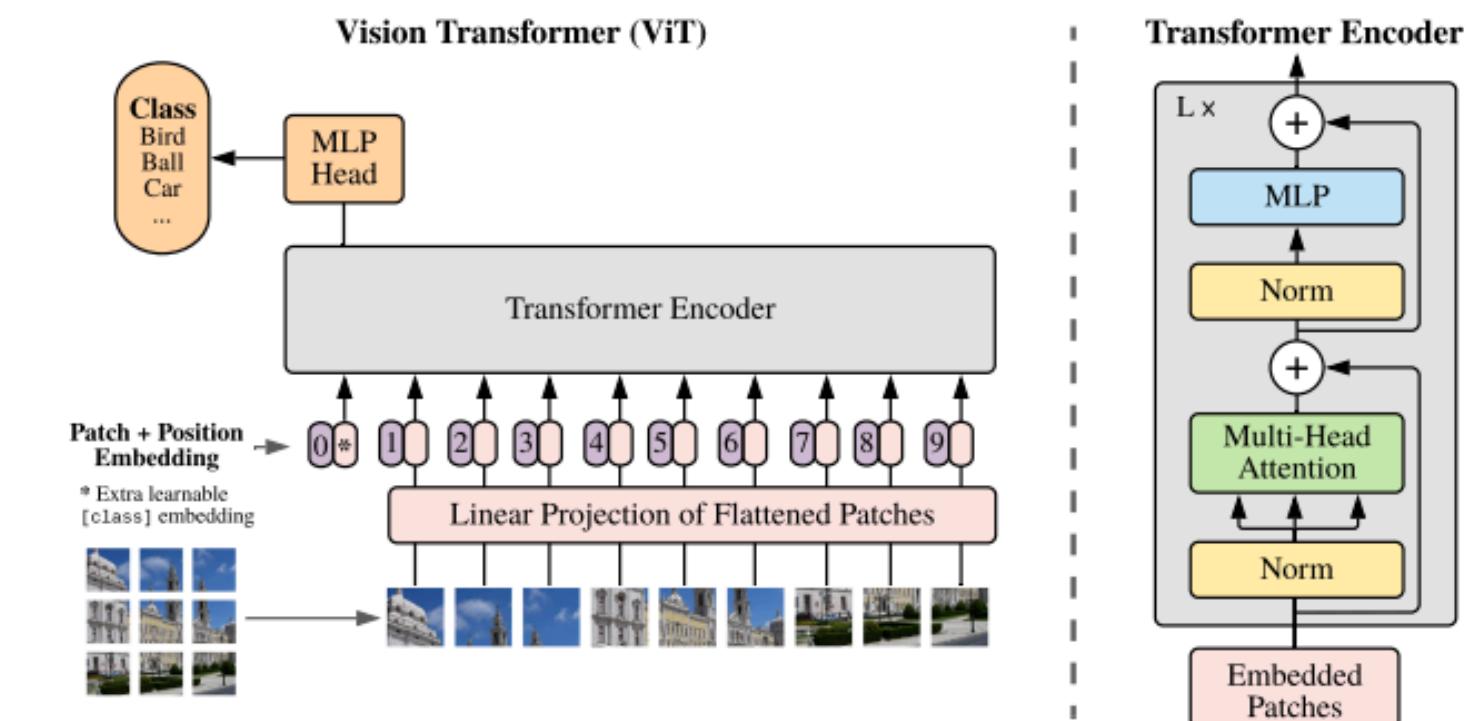
The MLP contains two layers with a GELU non-linearity.

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

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

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

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



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

EXPERIMENTS

実験設定

モデル比較

- ResNet、ViT、hybrid modelの表現学習能力を評価した

データセット

- 事前学習データセット
 - ImageNet (ILSVRC-2012)(クラス数：1,000,画像枚数：約1.3M)
 - ImageNet-21k:(クラス数：21,000,画像枚数：約14M)
 - JFT:(クラス数：18,000,画像枚数：約303M (高解像度))

EXPERIMENTS

実験設定

データ重複排除

- 下流タスクのテストセットと重複する画像を， Kolesnikov et al. (2020) の手法に従って除去

転移学習による評価タスク

- ImageNet (元の検証ラベルおよびReaLラベルによる評価：Beyer et al. 2020)
- CIFAR-10／CIFAR-100 (Krizhevsky, 2009)
- Oxford-IIIT Pets (Parkhi et al., 2012)
- Oxford Flowers-102 (Nilsback & Zisserman, 2008)
- 各データセットの前処理はKolesnikov et al. (2020) に準拠

EXPERIMENTS

実験設定

VTAB (Visual Task Adaptation Benchmark) 評価

- 合計19タスクから構成
- 各タスクに対して1,000例の訓練データを使用し、低データ環境での適応性能を測定 (Zhai et al., 2019b)
- タスクの分類：
 - Natural (自然画像, Pets, CIFARなど)
 - Specialized (医療画像, 衛星画像など)
 - Structured (位置推定などの幾何情報を要するタスク)

EXPERIMENTS

実験設定

モデルバリエント

- ViT: BERT と同じ “Base / Large” 設定 + さらに大きい “Huge”。
 - 記法 ViT-L/16 → Large + パッチ 16×16 。
 - トランスフォーマの 系列長 $\propto 1 / (\text{パッチサイズ}^2)$ → パッチを細かくすると計算量急増。
- ResNet (BiT): BatchNorm → GroupNorm、標準化畳み込みなど転移に強い改良版。
- ハイブリッド: ResNet 途中の特徴マップを パッチサイズ 1 “画素” として ViT に渡す。
 - 取り出し位置を変えて系列長を $1 \times$ or $4 \times$ に調整。

EXPERIMENTS

実験設定

学習とファインチューニング

フェーズ	手法	主なハイパーパラメータ
事前学習	Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)	バッチ 4,096 / Weight decay 0.1 / LR Warm-up → 線形減衰
FT	SGD + momentum	バッチ 512 / LR スケジュールは Appendix B.1.1
高解像度 FT (Table 2 用)	ViT-L/16: 512 px, ViT-H/14: 518 px	Polyak Averaging (係数 0.9999)

評価指標

- Fine-tuning Accuracy: それぞれの下流データセットで全層更新した後の精度。
- Few-shot Accuracy: 線形回帰（正則化付き）を閉形式で解き、凍結特徴の質を高速評価。
 - ViT の表現力を手軽に比較でき、ハイパラ探索時に便利。

EXPERIMENTS

実験設定

結果

- ViT-L/16 (JFT 学習) は BiT-L をすべての下流タスクで凌駕し、計算量は約半分。
- ViT-H/14 はさらに上回り、ImageNet・CIFAR-100・VTAB など「難易度高め」のデータセットで特に差が拡大。
- それでも Noisy Student より 計算コストが小さい。
- 公開データのみでも、ViT-L/16 + ImageNet-21k は競争力が高く、クラウド 8 TPUv3 (≈ 1 Pod) で 1 か月と現実的。

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	88.55 \pm 0.04	87.76 \pm 0.03	85.30 \pm 0.02	87.54 \pm 0.02	88.4 / 88.5*
ImageNet ReaL	90.72 \pm 0.05	90.54 \pm 0.03	88.62 \pm 0.05	90.54	90.55
CIFAR-10	99.50 \pm 0.06	99.42 \pm 0.03	99.15 \pm 0.03	99.37 \pm 0.06	—
CIFAR-100	94.55 \pm 0.04	93.90 \pm 0.05	93.25 \pm 0.05	93.51 \pm 0.08	—
Oxford-IIIT Pets	97.56 \pm 0.03	97.32 \pm 0.11	94.67 \pm 0.15	96.62 \pm 0.23	—
Oxford Flowers-102	99.68 \pm 0.02	99.74 \pm 0.00	99.61 \pm 0.02	99.63 \pm 0.03	—
VTAB (19 tasks)	77.63 \pm 0.23	76.28 \pm 0.46	72.72 \pm 0.21	76.29 \pm 1.70	—
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

EXPERIMENTS

実験設定

1. データ規模別事前学習実験

- 小規模データでは大モデル（ViT-Large）は Base に劣るが、
14k→300 M へ増えるほど大モデルの優位性が顕在化。
- BiT CNN と比較しても、データ量が増すと ViT が逆転して上
回る。

2. サブセット学習実験

- 9 M では ViT が過適合しやすく ResNet に劣るが、90 M 以上
で性能を逆転。
- 結論：小規模時は畳み込みのバイアスが有利、大規模時は ViT
の柔軟学習が有効。

EXPERIMENTS

実験設定

1. モデルの計算量 vs 転移性能比較

- ResNet 系 (R50~R200)、ViT 系 (B/32~H/14)、ハイブリッド
計 18 種を TPU コスト換算で比較。
- 結果：ViT は同等性能を得るのに ResNet の 2-4 倍少ない計算量
で済む。
- ハイブリッドは小規模予算で僅かに ViT を上回るが、大規模で差が
なくなる。
- ViT は試した範囲で飽和せず、さらなるスケーリングの余地あり。

EXPERIMENTS

実験設定

1. 入力投影フィルタの主成分

- パッチ埋め込みの線形投影で学習されるフィルタは、自然画像に適した低次元基底として妥当な形状を示す（エッジやガボール類似）。

2. 位置埋め込み

- 位置埋め込み間の類似度が画像上の距離に比例。行列構造や場合によっては正弦波構造も学習。

3. Attention Distance

- 低層から全画面を参照するヘッドもあり、グローバル統合能力を積極利用。
- 一部のヘッドは局所注意に特化し、初期 CNN 層と似た役割を果たす可能性。
- ネットワーク深度とともに注意距離が拡大。

EXPERIMENTS

実験設定

4.6 Self-Supervision

- Masked Patch Prediction
- BERT の MLM に倣い、パッチの一部を隠して予測する自己教師あり事前学習を実施。
- ViT-B/16 で ImageNet 上 79.9 % (訓練 from scratch 比 +2 %) を達成。
- しかし監視あり事前学習 (約84 %前後) にはまだ及ばず、コントラスト学習等の検討が残課題。

研究ってなにするの？？

what's contents

