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

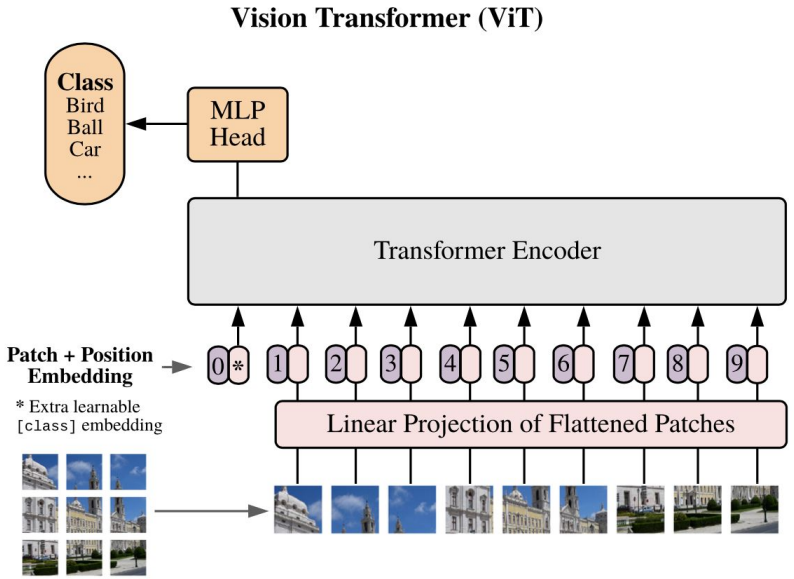
機械学習特論期末課題　TP22013 下村晃太

概要

- Vision Transfomerとは
- 従来手法との相違点・特徴
- モデルアーキテクチャ
- 実験結果

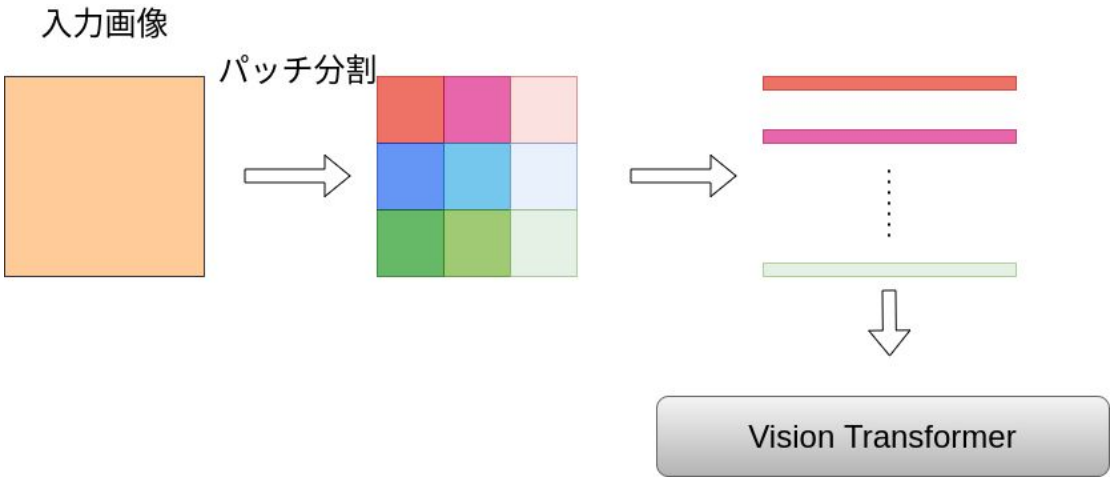
Vision Transformerとは

- NLP分野のTransfomerを画像分野に応用したモデル
- 入力画像をパッチ分割をして単語のように扱う
- Transformer Encoderを使用



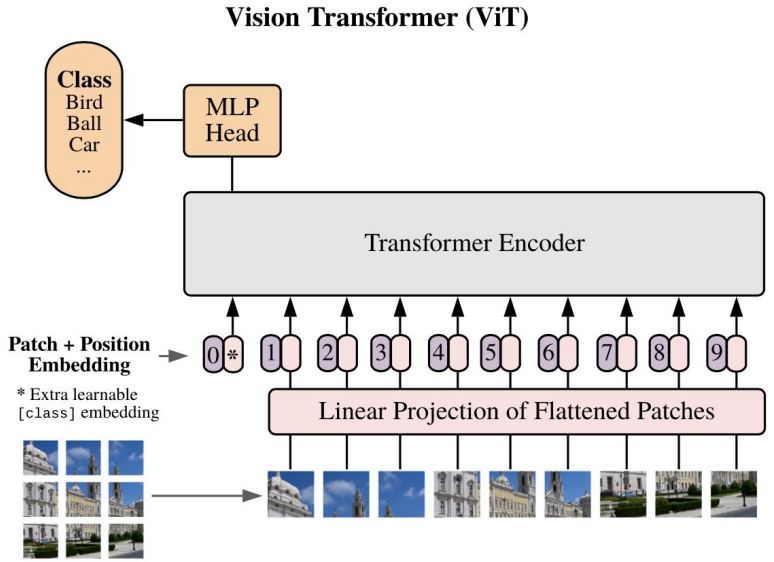
従来手法との相違点・特徴

- 分割したパッチ間の特徴を学習
- CNNとは違い離れた画像情報を学習可能
- 入力画像をパッチ分割しVision Transformerに入力



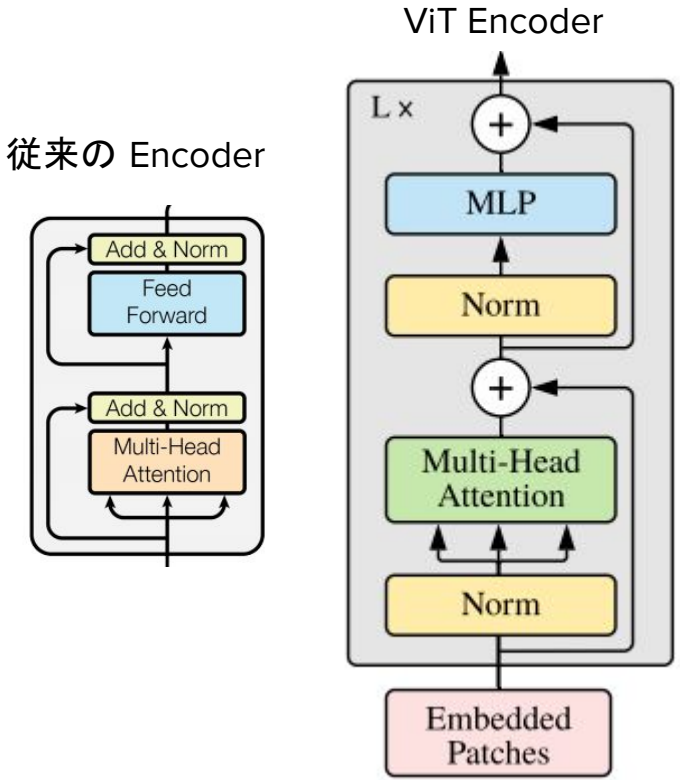
モデルアーキテクチャ[1/n]

- 分類タスク用にMLPHeadを追加
- Transformer Encoderがメイン
- Position Embeddingにはパッチの位置情報を用いる



モデルアーキテクチャ[2/n]

- 従来のTransformer Encoderとの相違点
  - layer NormがMHAの前に位置
  - MLPの活性化関数にGELU
    - 従来はReLU



実験結果(実験コードはここ)

- 論文でのベンチマークとほぼ同等の精度を再現できた
  - ベンチマーク : Accurasy99.00
  - 再現結果 : Accurasy 98.77

```
**************infer_start**************
100% |████████████████████████████████| 40/40 [00:19<00:00, 2.07it/s]
Accuracy of the network on the 10000 test images: 98.770000 %
```

name	Epochs	ImageNet	ImageNet ReaL	CIFAR-10	CIFAR-100	Pets	Flowers	exaFLOPs
ViT-B/32	7	80.73	86.27	98.61	90.49	93.40	99.27	55
ViT-B/16	7	84.15	88.85	99.00	91.87	95.80	99.56	224
ViT-L/32	7	84.37	88.28	99.19	92.52	95.83	99.45	196
ViT-L/16	7	86.30	89.43	99.38	93.46	96.81	99.66	783
ViT-L/16	14	87.12	89.99	99.38	94.04	97.11	99.56	1567
ViT-H/14	14	88.08	90.36	99.50	94.71	97.11	99.71	4262
ResNet50x1	7	77.54	84.56	97.67	86.07	91.11	94.26	50
ResNet50x2	7	82.12	87.94	98.29	89.20	93.43	97.02	199
ResNet101x1	7	80.67	87.07	98.48	89.17	94.08	95.95	96
ResNet152x1	7	81.88	87.96	98.82	90.22	94.17	96.94	141
ResNet152x2	7	84.97	89.69	99.06	92.05	95.37	98.62	563
ResNet152x2	14	85.56	89.89	99.24	91.92	95.75	98.75	1126
ResNet200x3	14	87.22	90.15	99.34	93.53	96.32	99.04	3306
R50x1+ViT-B/32	7	84.90	89.15	99.01	92.24	95.75	99.46	106
R50x1+ViT-B/16	7	85.58	89.65	99.14	92.63	96.65	99.40	274
R50x1+ViT-L/32	7	85.68	89.04	99.24	92.93	96.97	99.43	246
R50x1+ViT-L/16	7	86.60	89.72	99.18	93.64	97.03	99.40	859
R50x1+ViT-L/16	14	87.12	89.76	99.31	93.89	97.36	99.11	1668