学習サンプルが少ない状況下での Vision Transformer の性能評価に関する研究

計算機視覚工学研究室
2025050454
佐藤立樹

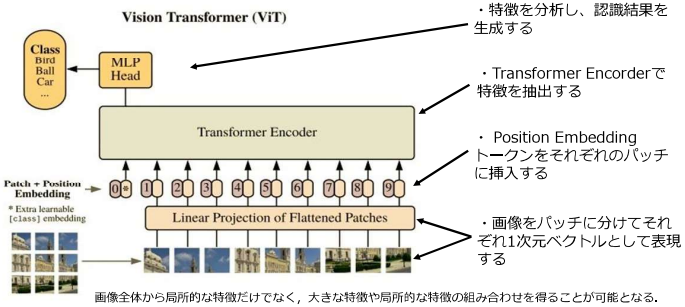
研究背景

近年話題となっているVision Transformerだが、Vision Transformerの性能が十分に発揮されるのは学習サンプルが数億枚以上必要である。しかし、そこまでのサンプルをそろえることが困難な状況は多々あるはずである。

目的

本研究の目的は、学習サンプルが少ない状況で、事前学習を用いた場合のVision Transformerの性能を検討することである。

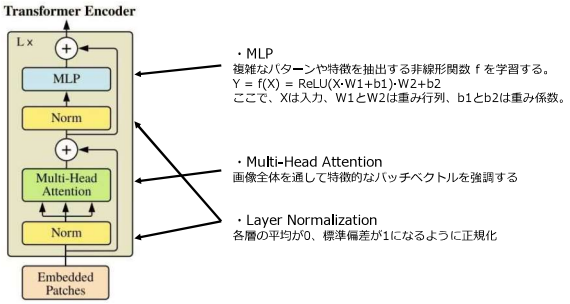
Vision Transformer



画像全体から局所的な特徴だけでなく，大きな特徴や局所的な特徴の組み合わせを得ることが可能となる．

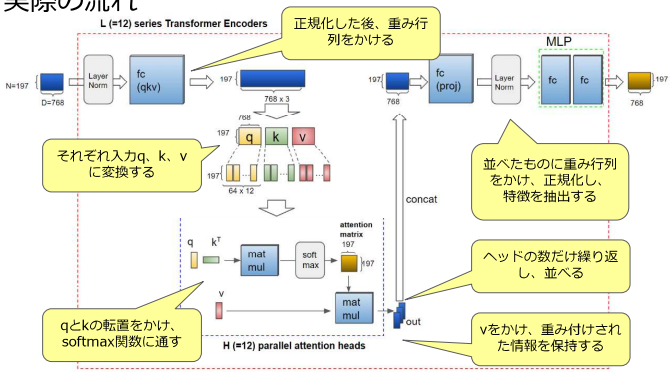
引用元：https://qiita.com/zisui-sukitarou/items/d990a9630ff2c7f4abf2

Transformer Encorder



引用元：https://qiita.com/zisui-sukitarou/items/d990a9630ff2c7f4abf2

実際の流れ



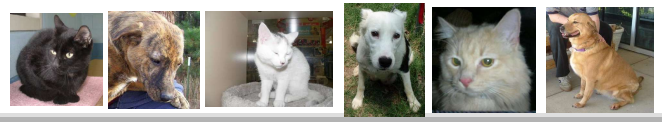
シミュレーション実験の説明

実験内容

以下の条件で、
Vision Transformer(ViT)、ResNet-152(ResNet)、EfficientNetV2-S(EfficientNet)
の犬猫判別時の性能を比較する。

実験条件１：事前学習無し、学習サンプル数 1000、100、50、20
実験条件２：事前学習有り、学習サンプル数 1000、100、50、20

犬猫の画像サンプル



事前学習

事前学習とは、大規模なデータセットでネットワークを初めて学習する手段

画像認識のタスクでは、事前学習によってネットワークはエッジ検出、テクスチャ認識などの一般的な視覚特性を学習する

その後、この事前学習モデルは特定の目的タスク（犬と猫の分類）に適応するために微調整される

これにより、事前学習がモデルに広範かつ一般的な知識を与え、適応能力を高める

学習サンプルの前処理

元の学習サンプル　224*224にリサイズした学習サンプル　ランダムに水平方向に反転・拡大した学習サンプル

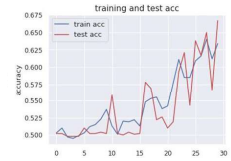
シミュレーション実験の結果

実験条件１　事前学習無し、学習サンプル数1000

ResNet　認識率：57.60%

EfficientNet　認識率：62.40%

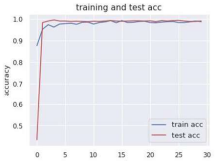
Vision Transformer　認識率：62.78%



精度のばらつき有り、70%を超えることがなかったためここで終了

実験条件2　事前学習有り、学習サンプル数1000

ResNet　認識率：98.98%



EfficientNet　認識率：99.10%

Vision Transformer　認識率：99.62%

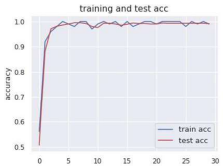
EfficientNetと0.5%程度の差だがViTのほうが高い評価を得られた

実験条件2　事前学習有り、学習サンプル数100

ResNet　認識率：97.16%

EfficientNet　認識率：73.10%

Vision Transformer　認識率：99.26%

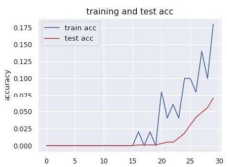


ResNetと2%の差をつけてViTが高い評価を得られた

実験条件2　事前学習有り、学習サンプル数50

ResNet　認識率：95.38%

EfficientNet　認識率：4.86%



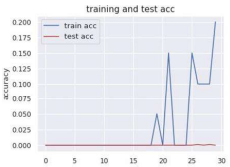
Vision Transformer　認識率：98.80%

EfficientNetがほとんど機能していない
ResNetと3.5%程度の差をつけてViTが高い評価を得られた

実験条件2　事前学習有り、学習サンプル数20

ResNet　認識率：93.64%

EfficientNet　認識率：0.04%

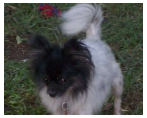
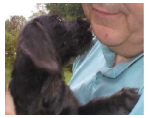


Vision Transformer　認識率：97.62%

ResNetと4%程度の差をつけてViTが高い評価を得られた

認識に失敗した画像の例




犬を猫だと認識された画像


猫を犬だと認識された画像

ViT（学習サンプル20）で学習した際に誤認識した画像は1000枚中58枚（犬画像56枚、猫画像2枚）

考察

・学習サンプル数1000の事前学習無しと事前学習有りの結果から、事前学習有りのほうが圧倒的に精度が良いという事がわかった。

・事前学習有りの方では、学習サンプル数として1000、100、50、20を用いた全ての場合において、ResNet及びEfficientNetよりVision Transformerの精度が優れていた。よって、学習サンプル数が少ない状況下で事前学習を用いればVision Transformerは最も有望であることが分かった。

今後の課題

・本研究では犬と猫の判別を行ったが、他の学習サンプルでも同じ結果が得られるかを検証していきたい。

Vision Transformerの利点

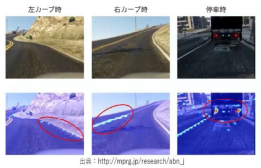
局所パターンの組み合わせで識別する事例

・**複数の箇所から判断しなければいけない状況に強い**

医用画像診断

複数の臓器の状態から
総合的に判断しないといけない
（医師のノウハウ）

自動運転のための画像認識



状況に応じて複数の場所に
注意を払わないといけない
（教習所で教えられること）