7C-03

物体検出のためのバックボーンとしての再帰ニューラルネットワーク

立浪 祐貴 †　瀧 雅人 †
立教大学大学院人工知能科学研究科 †

1 はじめに

再帰ニューラルネットワーク (RNN) は画像認識にも利用できることが知られている．その代表的なアーキテクチャである Sequencer [1] は, ImageNet データセットで訓練された初めての RNN ベースのモデルで, 画像分類においては Vision Transformer(ViT)[2] などの最先端の手法にも匹敵する．しかしながら, Sequencer を物体検出フレームワークのバックボーンにしても, ResNet と比較しても見劣りする精度しか達成できない．本研究では, 物体検出にも対応できるアーキテクチャである Rotational Sequencer(RoS) を提案する．物体検出のバックボーンとしての RoS は, Sequencer を上回るだけでなく, 他のバックボーンと比較しても高い精度を得る．それだけでなく, RoS が画像分類においても Sequencer を凌駕する Top-1 精度が達成できることを示す．

2 手法

先行研究である Sequencer では, ResNet のようにダウンサンプリングを 4 回行うような 4 ステージ型の階層構造ではなく, 2 ステージ型の構造を採用している．このような事情から, ResNet 向けに開発された物体検出手法ではスケールを正しく捉えることが出来ず, 思うような精度にはならなかったと考えられる．そこで, 本稿では ResNet に合わせて 4 ステージ型の階層構造を採用した．ところが, Sequencer は特徴マップを列として扱うことから, この変更によって列が長くなり, RNN の特性によって速度が悪化してしまう．そこで, 速度が維持されるような工夫をいくつか導入した．Sequencer では ViT の自己注意にあたる空間集約の部分に, 2 次元双方向 LSTM, つまり 4 つの RNN を用いていた．RoS では, その代わりに RoS 層を使用する．RoS 層には主として 3 つの工夫を組み込んでいる．まず, 回転することで 4 つの RNN であったものを 1 つの RNN で重み共有, および一括で処理できるようにした．これによりパラメータの削減と速度の向上を試みている．次に, プーリングを活用し, 列を短くすることを試みた．これによって速度は向上するが, 弊害として, 近傍の情報集約が手薄になる．それを補うために 3x3 の depthwise 畳み込みの経路も導入した．さらに, 初期状態をより良く利用するために, 全体の情報を集約しそれを初期状態として注入する機構を導入した．提案するアーキテクチャは 4 ステージ型の階層構造を持ち, Transformer ブロックの自己注意層を RoS 層で置き換えたブロックの繰り返しを基調とする．ただし, 空間集約部分に RoS 層を使用している．RNN として GRU を用いた RoS-S/G と, LSTM を用いた RoS-S/L を提案する．いずれも今後の実験で検証に使用する．

Recurrent Neural Networks as Backbone for Object Detection
† Yuki Tatsunami, † Masato Taki
† Graduate School of Artificial Intelligence and Science

3 画像分類の実験

RoS-S/G および RoS-S/L を ImageNet-1K データセットを前処理して 224^2 の解像度にしてスクラッチ訓練した．[3] などで使用されているデータの水増しや正則化, ハイパーパラメータの設定を適用している．訓練した RoS-S/G および RoS-S/L に対して, ImageNet-1K の検証セットで Top-1 精度を報告する．表 1 がその結果であるが, オリジナルの Sequencer のネットワークも上回る精度を達成している．

Network	Param. (M)	FLOPs (G)	Top-1 Acc (%)
Deit-S [3]	22	4.6	79.8
Swin-T [4]	28	4.5	81.3
ConvNeXt-T [5]	29	4.5	82.1
Sequencer2D-S [1]	29	8.4	82.3
RoS-S/G	20	5.6	**82.6**
RoS-S/L	20	5.9	**82.6**

表 1　ImageNet-1k でスクラッチ訓練した画像分類モデルの Top-1 精度

4 物体検出の実験

物体検出の結果について紹介する．実験には人気のある物体検出のフレームワークの一つである RetinaNet [6] を使用した．そのバックボーンとして RoS を採用したモデルを, MS COCO データセットで訓練した．このモデルを別の画像分類モデルをバックボーンとした RetinaNet と比較する．いずれのモデルも, バックボーンには ImageNet で訓練した重みを初期値として使用する．表 2 はこれらのモデルたちを Average Precision(AP) で評価した結果である．RoS 搭載の RetinaNet はベースラインを十分上回る．

Backbone	AP	AP_S	AP_M	AP_L
ResNet50 [7]	36.3	19.3	40.0	48.8
Swin-T [4]	41.5	25.1	44.9	55.5
Sequencer2D-S [1]	33.6	15.3	37.5	50.2
RoS-S/G	**46.1**	**29.8**	49.9	**61.1**
RoS-S/L	**46.1**	28.5	**50.1**	60.7

表 2　RetinaNet を MS COCO で訓練した物体検出モデルの Average Precision

5 まとめと今後の課題

本稿では, 物体検出のバックボーン向きの RNN である RoS を提案した．RoS は画像分類と物体検出において優れた精度を達成できる．今後はよりパラメータ数の多いモデルでの実験や, 物体検出以外の下流タスクへの適用していきたい．

参考文献

[1] Yuki Tatsunami and Masato Taki. Sequencer: Deep lstm for image classification. In *NeurIPS*, 2022.

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

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

[4] Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[5] Zhuang Liu et al. A convnet for the 2020s. In *CVPR*, 2022.

[6] Tsung-Yi Lin et al. Focal loss for dense object detection. In *ICCV*, 2017.

[7] Kaiming He et al. Deep residual learning for image recognition. In *CVPR*, 2016.