What-Where Transformer: 物体の意味と位置情報を並行的に処理する画像バックボーンの基礎検討

Recognition, Control and Learning Algorithm Lab.

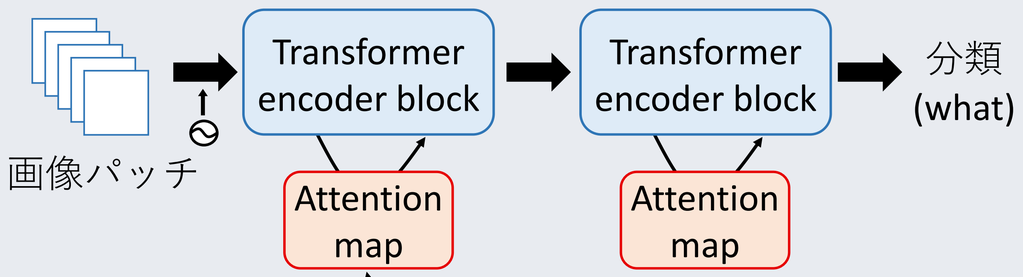
吉橋 亮太[1], 加太 将弘[1] , 池畑 諭[2,3], 川上 玲[1],佐藤 育郎[1,2]

[1]東京科学大学, [2]デンソーITラボラトリ, [3]国立情報学研究所

研究の背景と目的

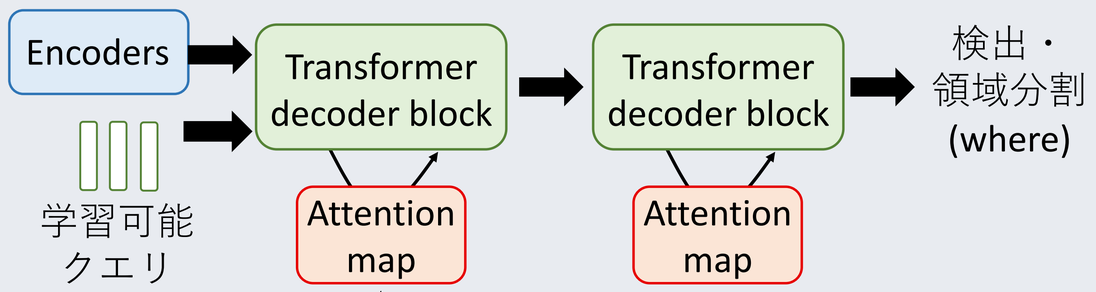
◆画像認識におけるTransformer

- Vision Transformer (ViT) [1]



注意マップ：空間的な情報集約により特徴抽出を補助

- Detection Transformer (DETR) [2]



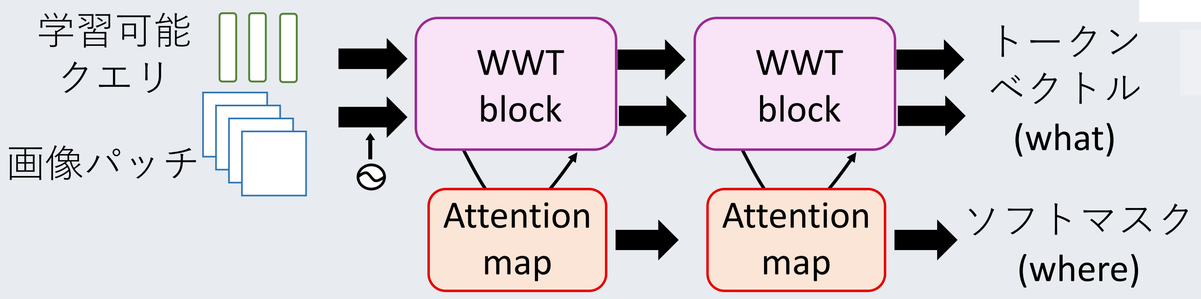
従来：分類と検出・領域分割に異なるエンコーダ・デコーダ構造が必要 = whatとwhereを別個処理

Q. 構造を統一しシンプル化・物体位置に関する情報が特徴抽出にも介在するようにできないか？

→ViT型バックボーン内部に物体位置情報を扱う機構を取り込む**What-Where Transformer (WWT)** を考案

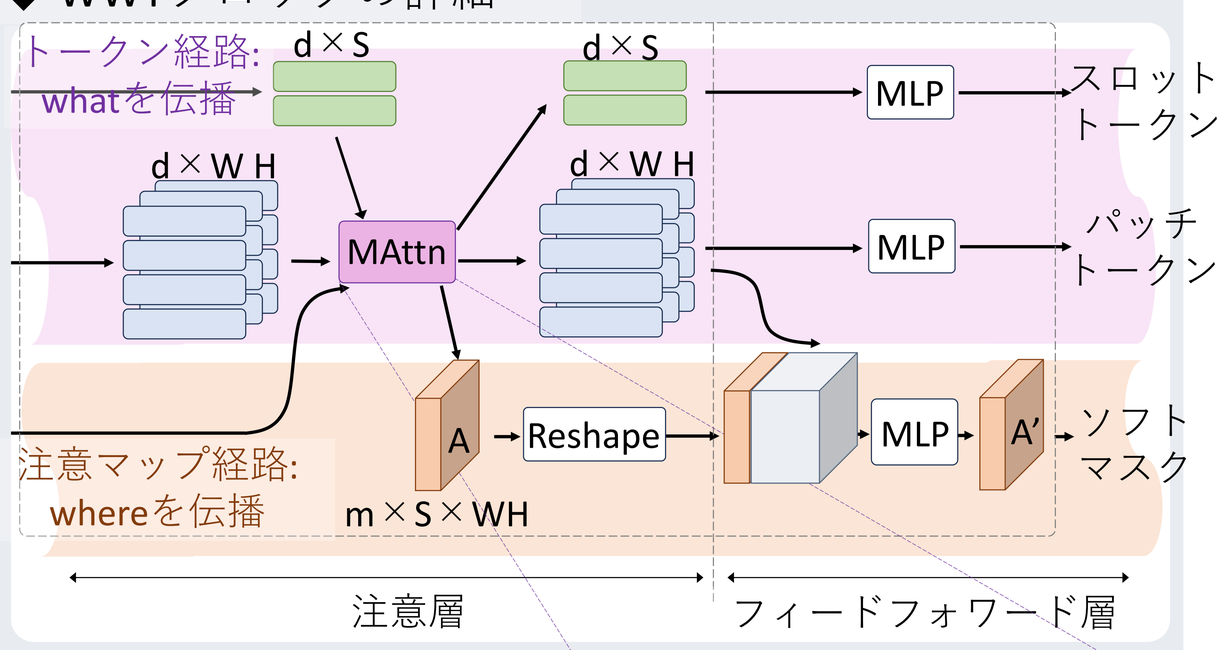
[1]: Dosovitskiy+, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, ICLR2021
[2]: Carion+, *End-to-End Object Detection with Transformers*, ECCV2020

提案手法: What-Where Transformer

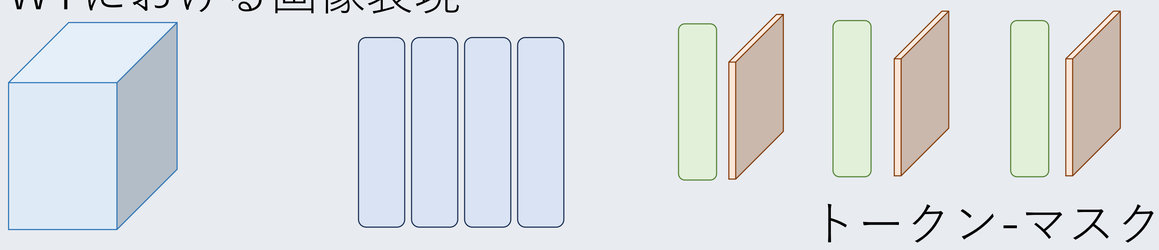


- トークン経路と注意マップ経路の複数ストリーム構造
  注意マップもネット内部を伝播・出力として再利用
- 学習可能クエリで画像の要素（例：物体）ごとのトークンや注意マップを学習：この単位を”スロット”と呼ぶ

◆ WWTブロックの詳細

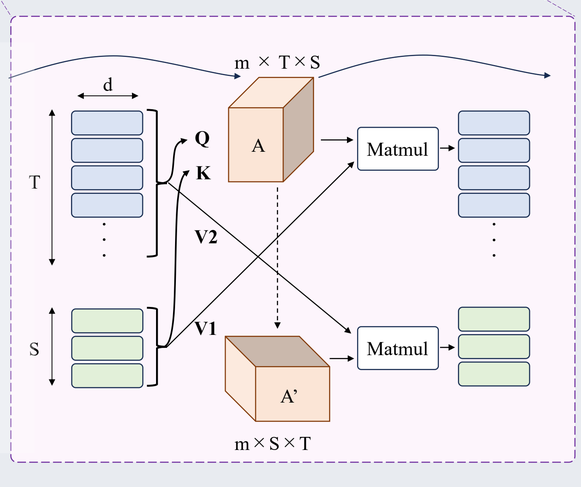


◆ WWTにおける画像表現



CNN: 特徴マップテンソル　ViT: トークン系列　**WWT**: トークン-マスク対の集合: Whatとwhereを分離的にエンコード

◆相互注意（MAttn）モジュール：
パッチトークンとスロットトークンを内積型注意により相互作用させ両方を更新



基礎的な評価

◆ Multi-MNISTデータセットで

- マルチラベル分類
- 入力の再構成 (自己符号化)
- 教師あり領域分割

を学習し各タスク性能と出力ソフトマスクを観察

マルチラベル分類

モデル	テスト正解率 (%)	学習正解率 (%)
CNN	**68.94**	74.88
ViT	20.18	99.48
ViT + 相互注意	23.20	86.09
WWT (ours)	32.01	**99.61**

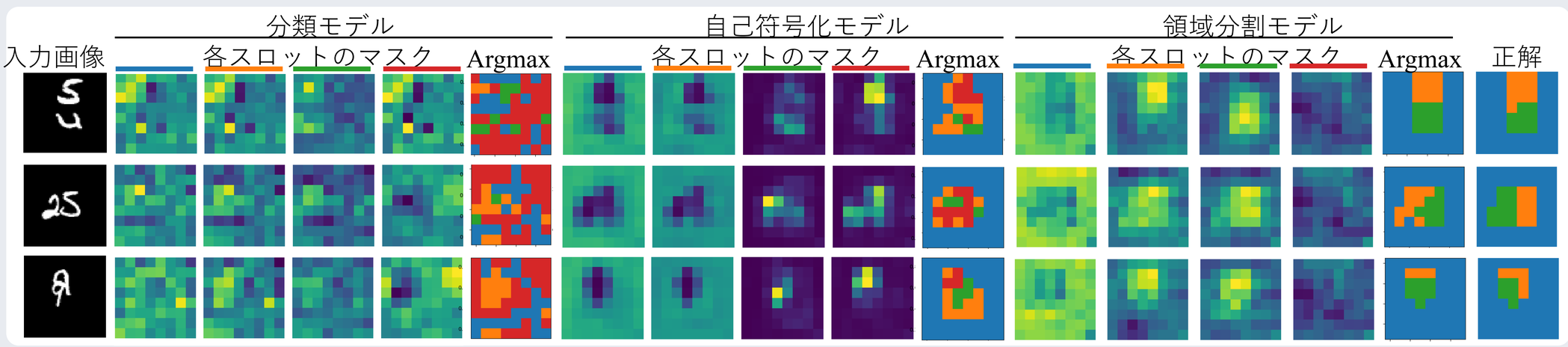
自己符号化

モデル	テスト mIoU	テスト再構成誤差
Slot Attention	17.20	0.0141
ViT	25.56	**0.0091**
WWT (ours)	**36.47**	0.0166

教師あり領域分割

モデル	テスト mIoU
ViT	**91.60**
WWT (ours)	88.38

◆出力マスクの可視化:
自己符号化 = 教師なし学習で物体ごとのマスクをある程度獲得



◆今後の展望：ImageNetなどで実用的なサイズのデータセットへのスケール性を確認

本研究は東京科学大学「DENSO IT LAB 認識・制御・学習アルゴリズム共同研究講座」における研究プロジェクト「機械学習の実応用に向けた品質基盤開発」の成果である