FY2022/2022年度

Robotics for Infrastructure Management Team
Takayuki Okatani
インフラ管理ロボット技術チーム　岡谷 貴之

RIKEN

Center for Advanced Intelligence Project

当チームの目標と研究内容

目標と背景

- 目標：橋梁などインフラ構造物の維持管理を，AI・ロボット技術で効率化・高度化
- 技術課題①：画像等を自動撮影可能なロボットの開発 ← Skydio社のドローンなど
- 技術課題②：画像等を元に**対象物の状況を理解し，説明できるAI**の開発

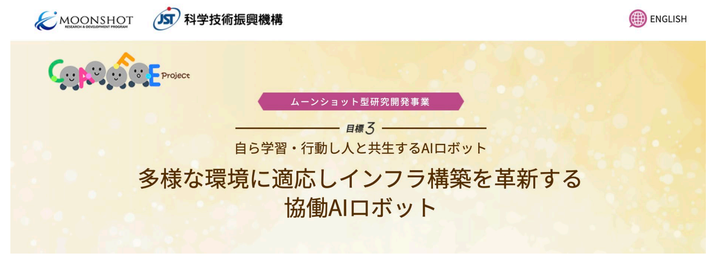
⇩

- 大規模言語モデルの発展（ChatGPTなど）→ 画像の言語化技術の重要度が増している

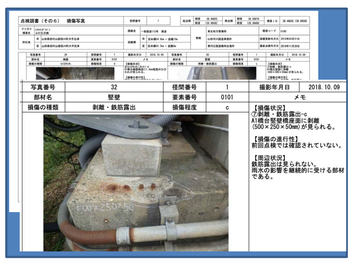
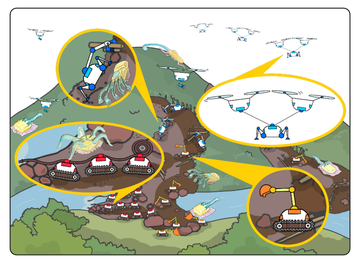
研究内容

- 画像と言語のマルチモーダルAIの研究を優先的に実施中
- 他の成果①：比喩表現を用いた画像記述 [Tran-Okatani, ACCV2022]
- 他の成果②：画像特徴の自己教師学習の研究 [Liu-Suganuma-Okatani, NeurIPS2022]

内閣府ムーンショット
CAFÉプロジェクトに
当チームが参加中



橋梁点検から災害現場・月面基地建設までをターゲットに，環境理解AIを研究中

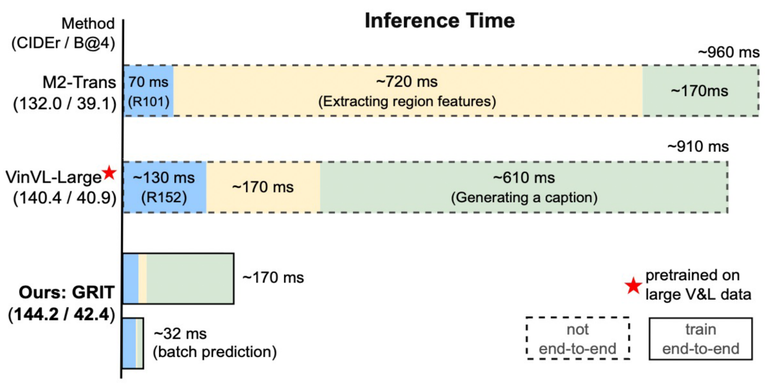




高速・高性能な画像記述AI

Nguyen, Suganuma, Okatani, GRIT: Faster and Better Image-captioning Transformer Using Dual Visual Features, ECCV2022

計算効率の高い画像記述AIの必要性

- 近年，学習が大規模化
  - モデルサイズ：順伝播計算1回～秒のオーダ
  - 訓練データサイズ：数百万～数十億の画像・テキストペアを使用
- 巨大な計算リソースがなければ，研究が不可能に
- 最適なアーキテクチャ設計で速度と性能の両方を向上



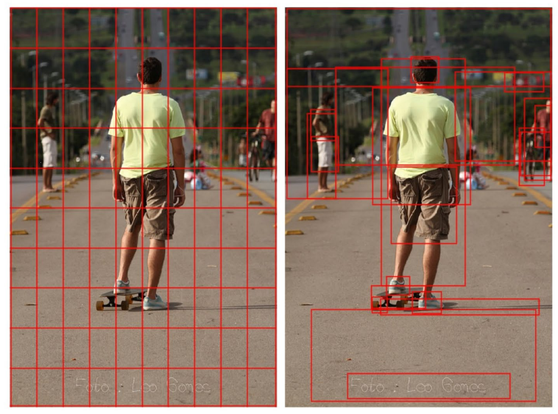
画像から取り出すべき特徴とは？

Grid-feature

- 個別物体の特徴は取れない ✕
- 物体と物体の関係性 ○
- 計算量少ない ○

Region-feature

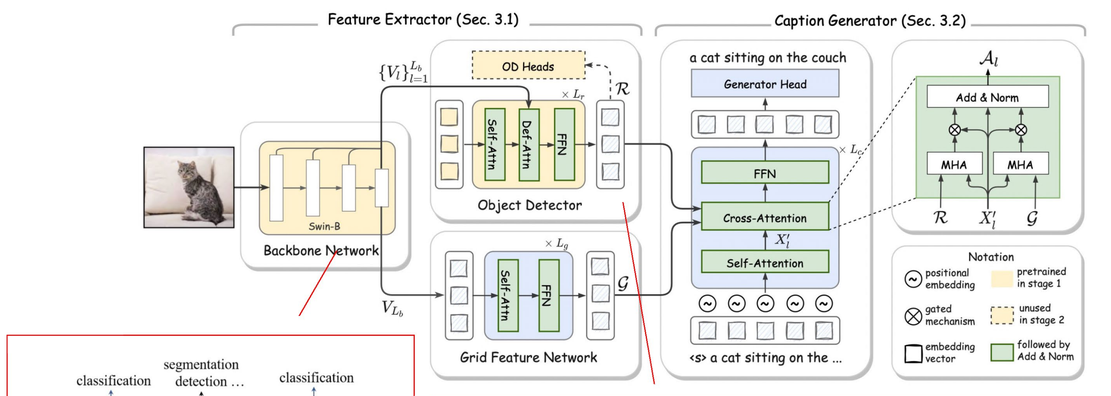
- 個別の物体の特徴を取れる ○
- 物体間の関係性 ✕
- 計算量多い ✕
- 誤検出時の悪影響 ✕

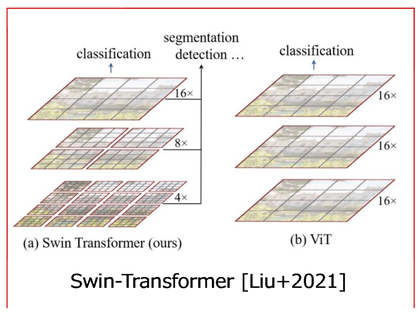


[Anderson+2018]

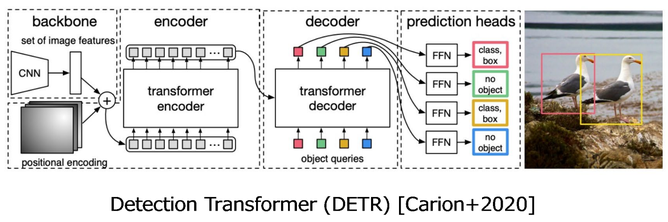
提案した画像記述AIのアーキテクチャ(GRIT)

- Grid/Region特徴の融合利用 + ”All Transformer” 構造





Swin-Transformer [Liu+2021]



Detection Transformer (DETR) [Carion+2020]

実験結果：記述精度

- 現時点で最良 (軽量かつ高性能)のアーキテクチャ

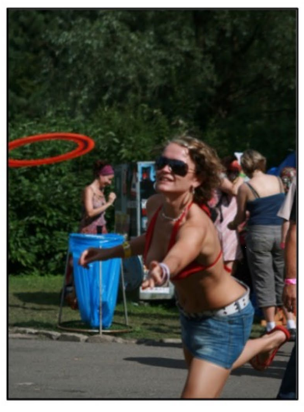
Method	V. E. Type	# VL Data	Performance Metrics B@1	B@4	M	R	C	S
w/ VL pretraining								
UVLP [55]	$\mathcal{R}$	3.0M	-	39.5	29.3	-	129.3	23.2
$\text{Oscar}_{\text{base}}$ [25]	$\mathcal{R}$	6.5M	-	40.5	29.7	-	137.6	22.8
$\text{VinVL}^{\dagger}_{\text{large}}$ [53]	$\mathcal{R}$	8.9M	-	**41.0**	31.1	-	140.9	**25.2**
$\text{SimVLM}_{\text{huge}}$ [45]	$\mathcal{G}$	1.8B	-	40.6	**33.7**	-	**143.3**	25.4
w/o VL pretraining								
SAT [43]	$\mathcal{G}$	-	-	31.9	25.5	54.3	106.3	-
SCST [38]	$\mathcal{G}$	-	-	34.2	26.7	55.7	114.0	-
LSTM-A [52]	$\mathcal{G}$	-	78.6	35.5	27.3	56.8	118.3	22.0
RSTNet [54]	$\mathcal{G}$	-	81.8	40.1	29.8	59.5	135.6	23.0
Up-Down [4]	$\mathcal{R}$	-	79.8	36.3	27.7	56.9	120.1	21.4
RFNet [20]	$\mathcal{R}$	-	79.1	36.5	27.7	57.3	121.9	21.2
GCN-LSTM [51]	$\mathcal{R}$	-	80.5	38.2	28.5	58.3	127.6	22.0
LBPF [35]	$\mathcal{R}$	-	80.5	38.3	28.5	58.4	127.6	22.0
SGAE [49]	$\mathcal{R}$	-	80.8	38.4	28.4	58.6	127.8	22.1
AoA [15]	$\mathcal{R}$	-	80.2	38.9	29.2	58.8	129.8	22.4
GET [16]	$\mathcal{R}$	-	81.5	39.5	29.3	58.9	131.6	22.8
ORT [13]	$\mathcal{R}$	-	80.5	38.6	28.7	58.4	128.3	22.6
ETA [24]	$\mathcal{R}$	-	81.5	39.3	28.8	58.9	126.6	22.6
$\mathcal{M}^2$ Transformer [7]	$\mathcal{R}$	-	80.8	39.1	29.2	58.6	131.2	22.6
X-LAN [33]	$\mathcal{R}$	-	80.8	39.5	29.5	59.2	132.0	23.4
TCIC [10]	$\mathcal{R}$	-	81.8	40.8	29.5	59.2	135.4	22.5
Dual Global [46]	$\mathcal{R}+\mathcal{G}$	-	81.3	40.3	29.2	59.4	132.4	23.3
DLCT [31]	$\mathcal{R}+\mathcal{G}$	-	81.4	39.8	29.5	59.1	133.8	23.0
GRIT	$\mathcal{R}+\mathcal{G}$	-	83.5	41.9	30.5	60.5	142.2	24.2
GRIT†	$\mathcal{R}+\mathcal{G}$	-	**84.2**	**42.4**	**30.6**	**60.7**	**144.2**	**24.3**

画像記述の例

正解(人間)
GT-1: An elderly man looks at a cell phone.
GT-2: An old man holding up a cell phone to his face.
既存手法
M²: a man is taking a picture of himself on a motorcycle
提案手法
GRIT: a man sitting in a chair holding a cell phone

GT-1: a table top with some plates of food on it
GT-2: Two plates of breakfast foods on a restaurant table.
M²: a plate of food with eggs and meat on a table
GRIT: two plates of food on a table with a fork

GT-1: there are many people in the beach playing volley ball
GT-2: some males on some sand are playing volleyball
M²: a group of people playing soccer on the beach
GRIT: a group of men playing volleyball on the beach



正解(人間)
GT-1: A woman throwing a frisbee outside at a park
GT-2: a woman is throwing a disk outside in the sun
既存手法 (M2-transformer)
M²: a woman holding a blue umbrella in the street
提案手法
GRIT: a woman is throwing a frisbee in the street

GT-1: Two frisbees laying on the ground next to a sports water bottle.
GT-2: Two flying disks on the ground next to a water bottle
M²: a knife and a knife on a table with a
GRIT: two frisbees laying on the ground next to a bottle

GT-1: Two knives are lying on a dark red surface.
GT-2: Two knives placed on a dining table
M²: a close up of a red tie with a
GRIT: two knives are on a red table with