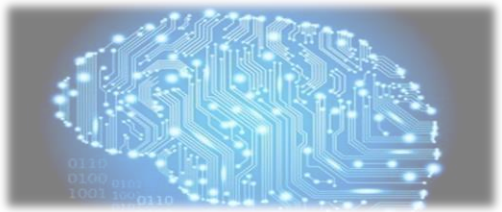
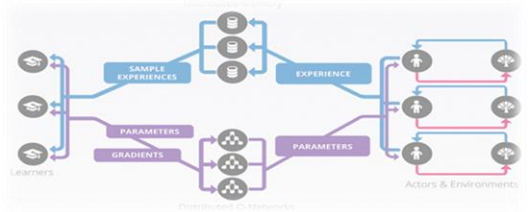



In Defense of Grid Features for Visual Question Answering

AutoML 연구실
석사과정 문아성

21.08.03

In Defense of Grid Features for Visual Question Answering

- In Defense of Grid Features for Visual Question Answering

Jiang, Huaizu, et al. "In defense of grid features for visual question answering."

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

The Winning Entry to the VQA Challenge 2020., Facebook AI Research

- Bottom-up Attention (Regions)
- Grid Features

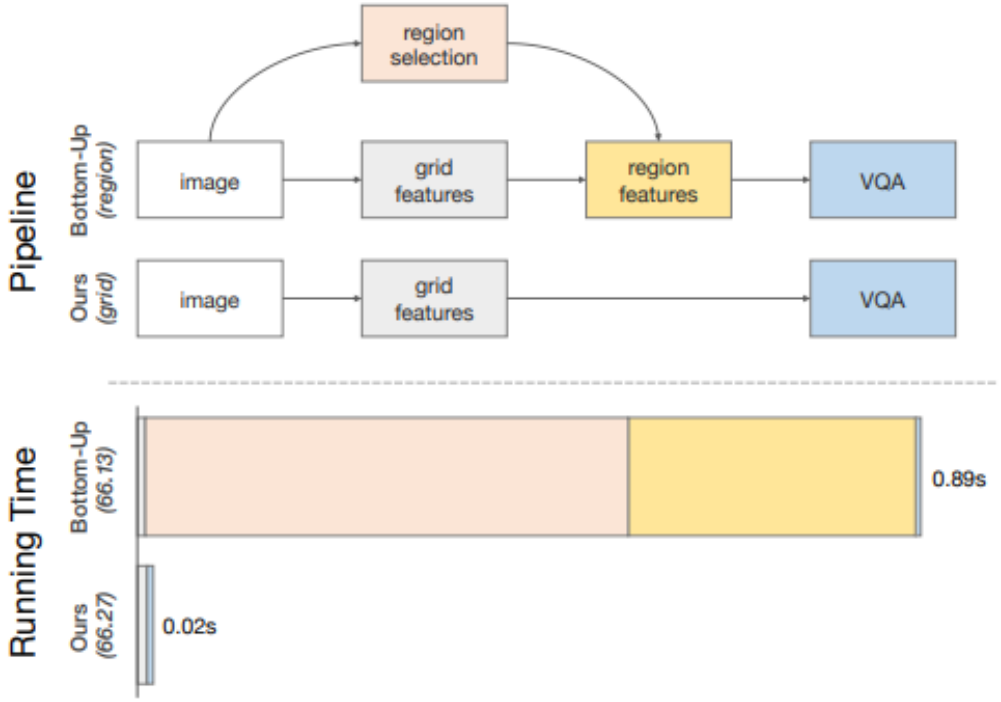
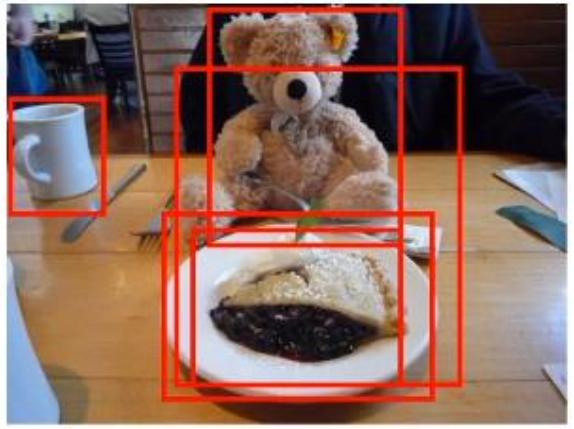


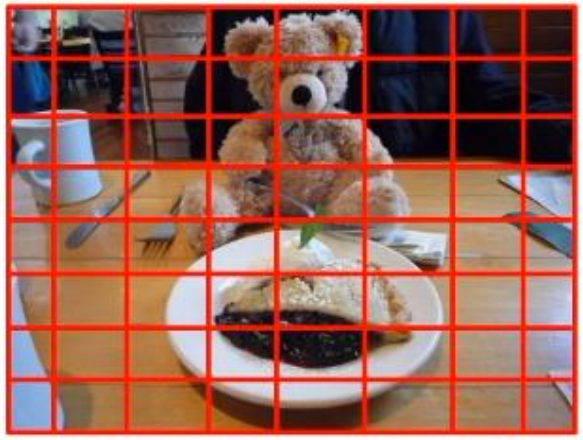
Figure 1: We revisit **grid**-based convolutional features for VQA, and find they can *match* the accuracy of the dominant **region**-based features from bottom-up attention [2], provided that one closely follow the pre-training process on Visual Genome [21]. As computing grid features skips the expensive region-related steps (shown in colors), it leads to significant speed-ups (all modules run on GPU; timed in the same environment).

Regions VS Grids

Question: Is the plate white?
GT-Answer: yes

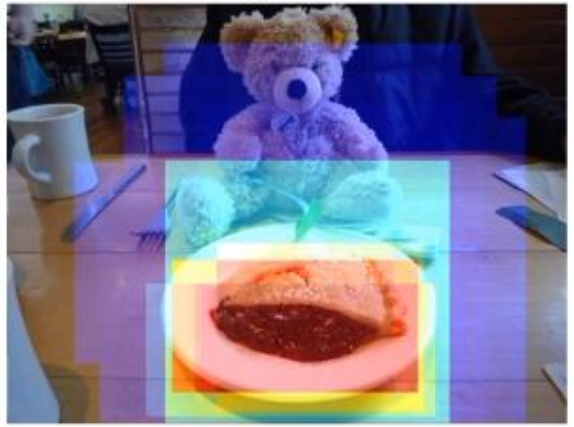


Region features [Anderson et al]
Answer: yes ✓



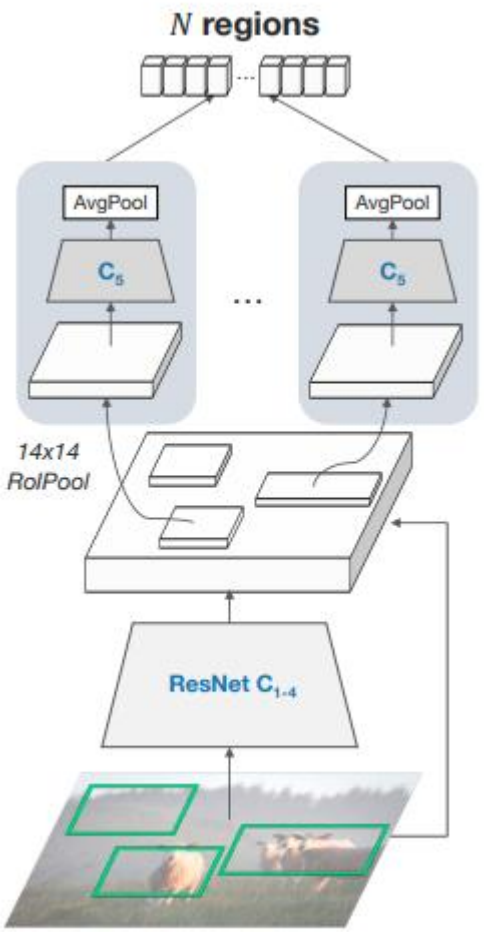
Grid feature (ours)
Answer: yes ✓

Grid features + MCAN:
state-of-the-art accuracy **(72.71)**
on VQA2.0 test-std

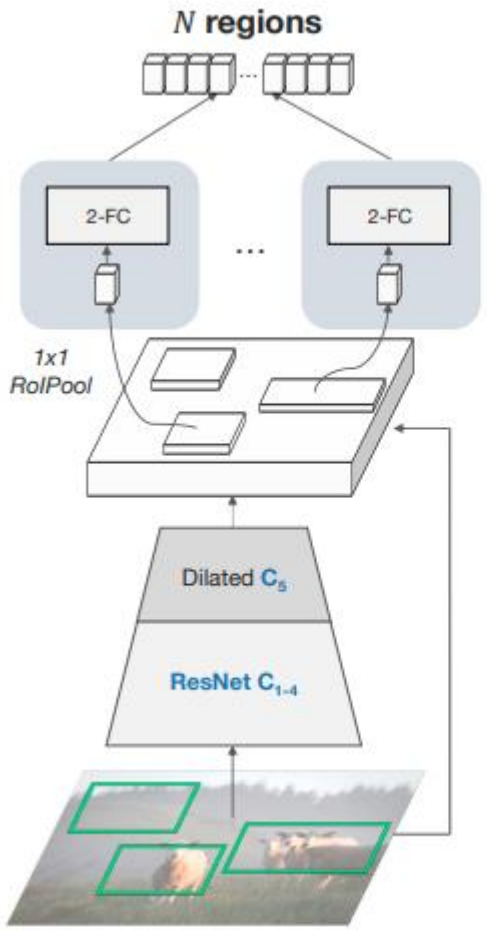


From Regions to Grids

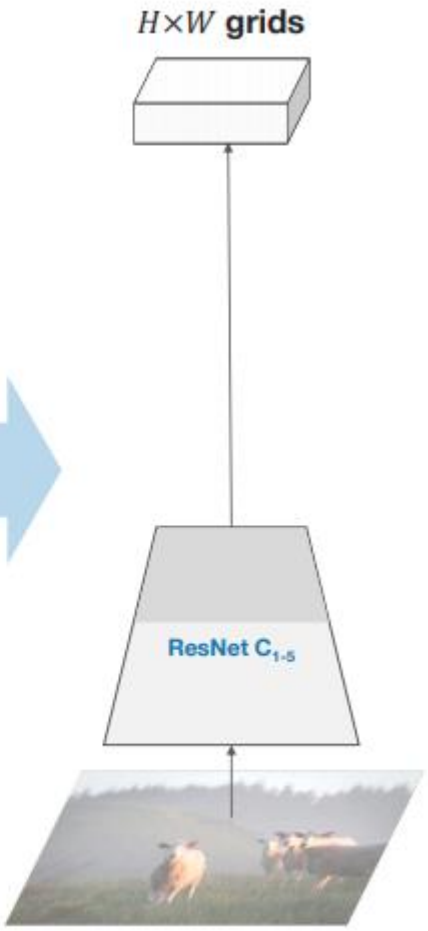
Pre-training for regions



Pre-training for grids



Grid feature extraction



Experimental Setting

Dataset: Visual Genome

108K images (train: 98k, test: 5k, validation 5k)

Backbone: Faster R-CNN (ResNet-50)

VQA Model: MCAN, Pythia

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In CVPR, 2019.

Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0.1: the winning entry to the vqa challenge 2018. arXiv preprint arXiv:1807.09956, 2018.

Main Results

#	feature	VG detection pre-train: RoIPool	region layers	AP	VQA: accuracy	Δ
1	R [2]	14×14	C_5 [15]	4.07	<u>64.29</u>	-
2		1×1	2-FC	2.90	63.94	*-0.35*
3	G	14×14	C_5	4.07	63.64	*-0.65*
4		1×1	2-FC	2.90	**64.37**	*0.08*
5		ImageNet pre-train			60.76	*-3.53*

Table 1: Main comparison. 'R' stands for region features as in bottom-up attention [2]. 'G' stands for grid features. All results reported on VQA 2.0 `vqa-eval`. We show that: **1)** by simply extracting grid features from the *same* layer C_5 of the same model, the VQA accuracy is already much closer to bottom-up attention than ImageNet pre-trained ones (row 1,3 & 5); **2)** 1×1 `RoIPool` based detector pre-training improves the grid features accuracy while the region features get worse (row 1,2 & 4). Last column is the gap compared to the original bottom-up features (underlined).

	# features (N)	`test-dev` accuracy	inference time breakdown (ms): shared conv.	region feat. comp.	region selection	VQA	total
R	100	66.13	9	326	548	6	889
	608	66.22	9	322	544	7	882
G	608	66.27	11	-	-	7	18

Table 2: Region *vs.* grid features on the VQA 2.0 `test-dev` with accuracy and inference time breakdown measured in milliseconds per image. Our grid features achieve comparable VQA accuracy to region features while being much faster without region feature computation and region selection.

Experiment Results

	accuracy	*time (ms)*
Pythia	68.31	-
R	68.21	929
G	67.76	39

	accuracy	*time (ms)*
MCAN	70.93	-
R	70.21	963
G	72.71	72

	accuracy	*time (ms)*
R	70.21	963
G	72.71	72
our	71.20	108