GRIT: Faster and Better Image captioning Transformer Using Dual Visual Features

Van-Quang Nguyen[1], Masanori Suganuma[1,2], Takayuki Okatani[1,2] [1]GSIS, Tohoku University, [2]RIKEN Center for AIP

ECCV TEL AVIV 2022

Abstract

We propose Transformer-only neural architecture named **GRIT** that effectively utilizes dual visual features to generate better captions:

- GRIT replaces the CNN-based detector employed in previous methods with a DETR-based one, making it computationally faster.
- Its monolithic design consisting only of Transformers enables end-to-end training of the model.

Question: How do we extract and fuse good visual representations?

New (Grid Features + Region Features) → Faster + Better

~~less accurate~~ ~~very slow~~

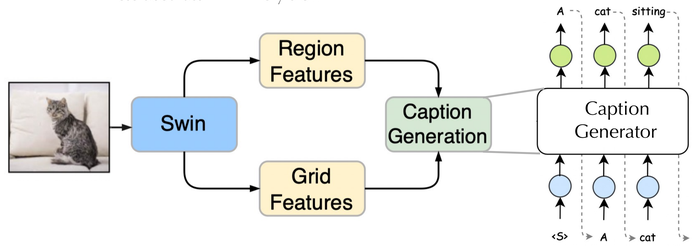
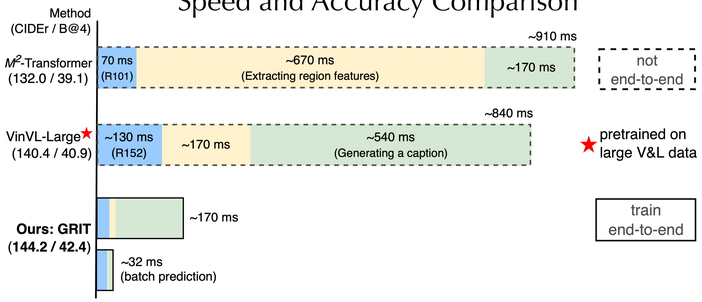


Figure: GRIT: **G**rid- and **R**egion-based **I**mage captioning **T**ransformer. It utilizes both region and grid features, making more accurate. It uses DETR-based detector, making it faster.

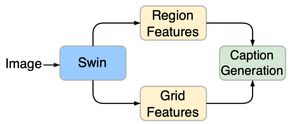
Speed and Accuracy Comparison



Previous Region-based methods

(M^2-Transformer [1], VinVL [2], etc.)

Our method - GRIT



Our Method

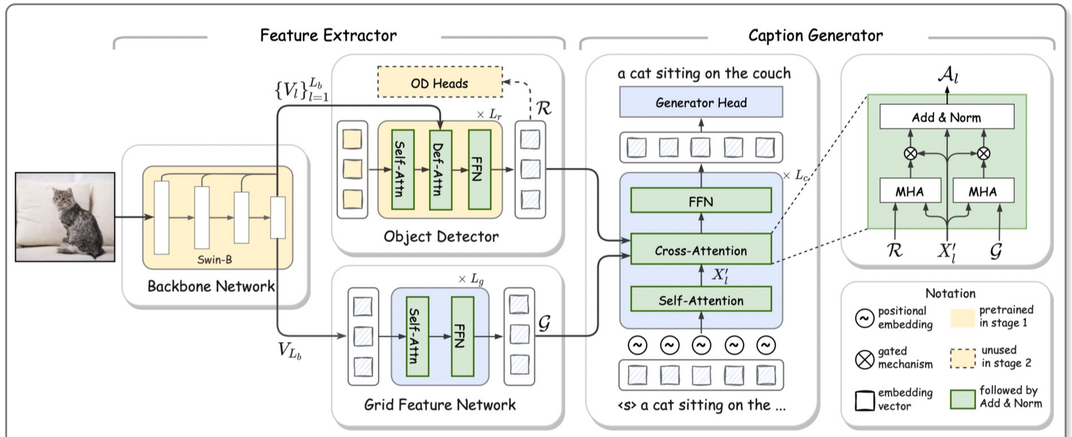


Figure: (1) Backbone Network to extract initial multi-level feature maps (2) Object Detector to extract region features (3) Grid Feature Network to extract grid features (4) Caption Generator to generate a caption.

Object Detector

To extract region features

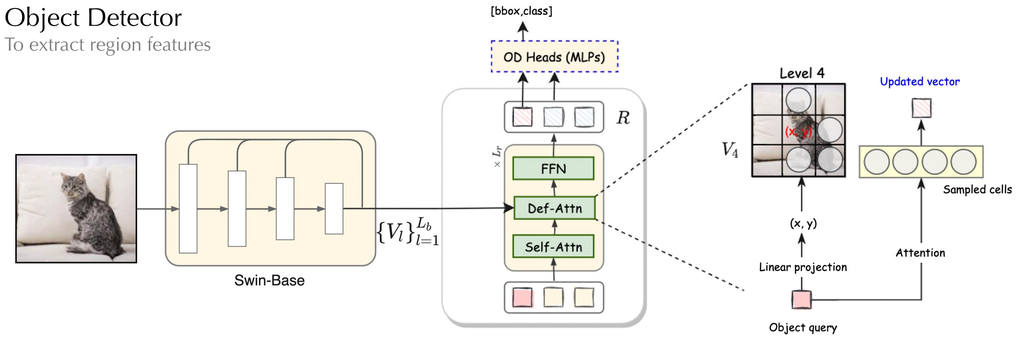


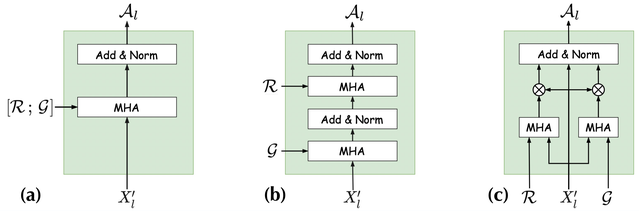
Figure: Use Swin-Base to extract multi-level feature maps. Build the object detector by stacking several deformable transformer layers to perform attention on a few of sampled input keys. The region features (R) are extracted from the object detector.

Cross-Attention Mechanism

The mechanism to use the dual visual features

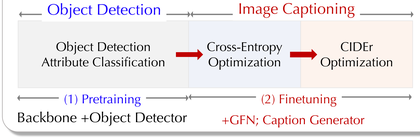
(a) Concatenated Cross-Attention
(b) Sequential Cross-Attention
(c) Parallel Cross-Attention

-> Parallel Cross-Attention yields the best!



Experimental Results

End-to-End Training



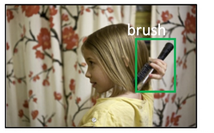
Training Datasets

- Object Detection: VG or [VG,Object365,OI,COCO]❖
- Image Captioning: COCO Image Captioning

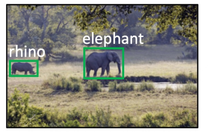
Results on the Karpathy test split of COCO

Method	V. E. Type	# VL Data	Performance Metrics B@1	B@4	M	R	C	S
w/ VL pretraining								
UVLP	$\mathcal{R}$	3.0M	-	39.5	29.3	-	129.3	23.2
$\text{Oscar}_{\text{base}}$	$\mathcal{R}$	6.5M	-	40.5	29.7	-	137.6	22.8
$\text{VinVL}^{\dagger}_{\text{large}}$	$\mathcal{R}$	8.9M	-	**41.0**	31.1	-	140.9	25.2
$\text{SimVLM}_{\text{huge}}$	$\mathcal{G}$	1.8B	-	40.6	**33.7**	-	**143.3**	**25.4**
w/o VL pretraining								
SAT	$\mathcal{G}$	-	-	31.9	25.5	54.3	106.3	-
RSTNet	$\mathcal{G}$	-	81.8	40.1	29.8	59.5	135.6	23.0
Up-Down	$\mathcal{R}$	-	79.8	36.3	27.7	56.9	120.1	21.4
AoA	$\mathcal{R}$	-	80.2	38.9	29.2	58.8	129.8	22.4
$\mathcal{M}^2$ Transformer	$\mathcal{R}$	-	80.8	39.1	29.2	58.6	131.2	22.6
TCIC	$\mathcal{R}$	-	81.8	40.8	29.5	59.2	135.4	22.5
Dual Global	$\mathcal{R}+\mathcal{G}$	-	81.3	40.3	29.2	59.4	132.4	23.3
DLCT	$\mathcal{R}+\mathcal{G}$	-	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 (❖: 4 datasets)	$\mathcal{R}+\mathcal{G}$	-	**84.2**	**42.4**	**30.6**	**60.7**	**144.2**	**24.3**

Qualitative Examples



M^2: a young girl holding a baseball bat in a
GRIT: a little girl brushing her hair with a brush



M^2: a group of elephants grazing in a field
GRIT: an elephant and a rhino standing in a field

M^2: a kitten laying on the floor next to a skateboard
GRIT: a cat laying on a skateboard on the floor

References & Our Code

[1] Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In CVPR, 2020.
[2] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In CVPR, 2021.

Code and pretrained models given at: https://www.github.com/davidnvq/grit