INJECTING SEMANTIC CONCEPTS into END-TO-END IMAGE CAPTIONING

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan,
Lijuan Wang, Yezhou Yang, Zicheng Liu
Arizona State University, *Microsoft Corporation*

Captioning Model can be Better without Object Detector

Time Cost of a Vision-Language model with Detector:

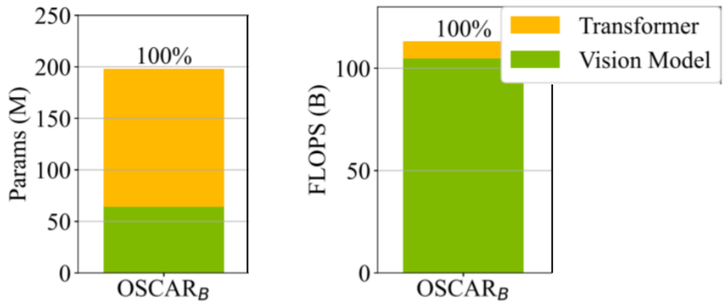


Figure 1. Number of parameters and the inference FLOPs (G) of OSCAR model. Though object detector only accounts for partial parameters, it takes up to more than 90% of the inference time due to regional operations (NMS, ROI).

Training Inflexibility:

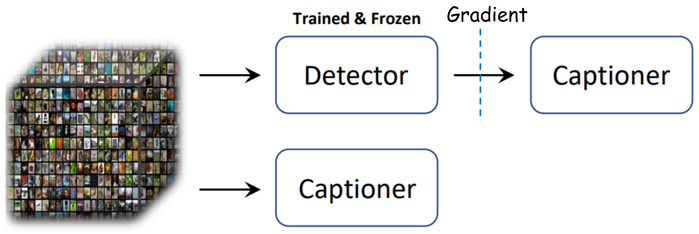
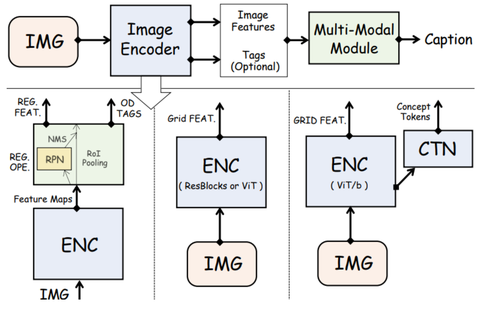


Figure 2. VL model with a frozen object detector leads to training inflexibility and domain gap challenge. In addition, training a decent object detector requires abundant bounding box annotations.

A Stronger E2E Captioning Model Without Detector

Model	E2E	Detector	Image Captioning	COCO-CIDEr
OSCAR_b	NO	YES	YES	137.6
VinVL_b	NO	YES	YES	**140.4**
ViLT	YES	NO	NO	-
SimVLM_b	YES	NO	YES	134.8
E2EVLM	YES	NO	YES	117.3
ViTCAP	YES	NO	YES	138.1



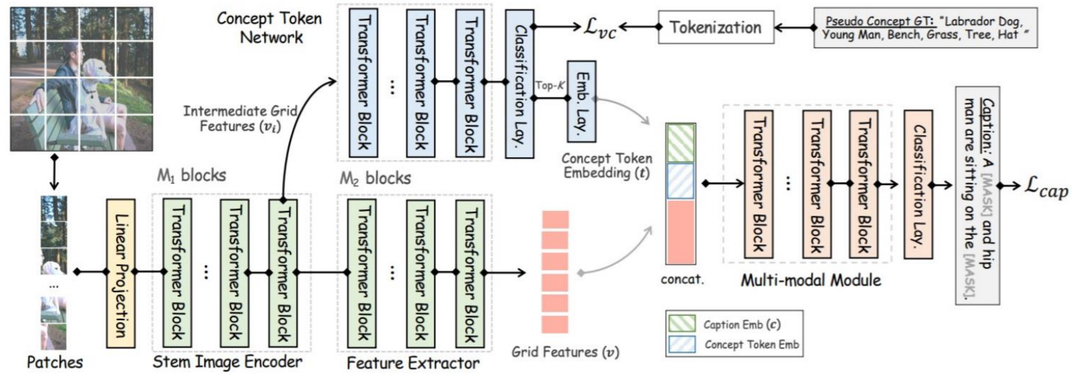


Figure 3. Architecture of our proposed ViTCAP image captioning model, which is detector-free based on the vision transformer architecture. The CTN branch is a shallow transformer architecture which produces concept tokens for decoding.

Injecting Semantic Concepts into End-to-End Image Captioning

- We present a detector-free image captioning model ViTCAP with fully transformer architecture, where it leverages grid representations without regional operations.
- We propose to inject semantic concepts into end-to-end captioning by learning from open-form captions. We find that our proposed concept classification training and concept tokens significantly benefit the captioning task.

The overall pipeline can be summarized as:

$$(v_i, v) = \mathbf{Encoder}(\mathbf{I}), \quad t = \mathbf{CTN}(v_i), \quad c = \mathbf{MM}(v, t).$$

Intermediate representations (vi) ; final grid representations (v) ; multi-modal module: MM ; Caption (c).

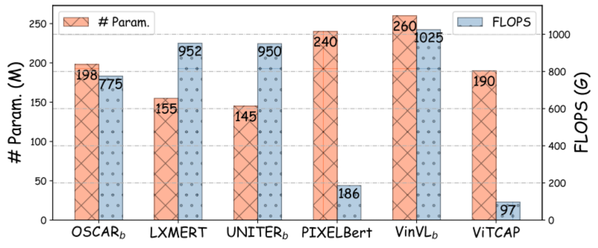
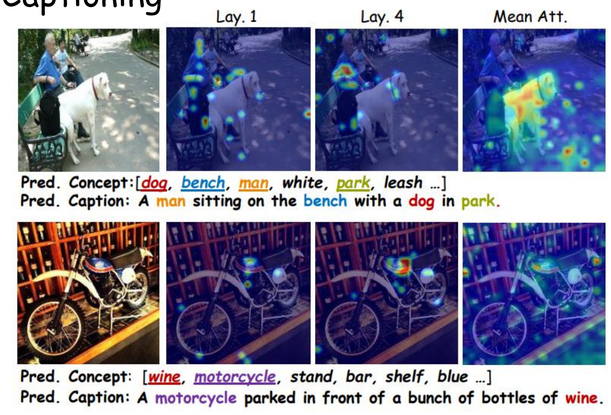


Figure 4. Inference speed in FLOPs (in G), number of parameters (in M) of multiple VL models and ViTCAP. ViTCAP consumes only ~10% FLOPs of the prevailing VL models (97G for ViTCAP vs. 1, 025G for VinVL).



Performance (w/o. pre-training)

Methods	V. ENC.	# *I-T*	Cross-Entropy Loss				
			B@4	M	R	C	S
Detector-Based Model							
AoANet [27]	F-RCNN$_{101}$	✗	38.9	29.2	58.8	129.8	22.4
M^2 Transfm. [11]	F-RCNN$_{101}$	✗	39.1	29.2	58.6	131.2	22.6
X-LAN [53]	F-RCNN$_{101}$	✗	39.5	29.5	59.2	132.0	23.4
RSTNet [81]	RESNeXt$_{152}$	✗	40.1	29.8	59.5	135.6	23.3
Detector-Free w.o. VLP							
ViTCAP (Ours)	ViT$_b$	✗	40.1	29.4	59.4	133.1	23.0

Acknowledgement

This work was supported by the National Science Foundation under Grant CMMI-1925403, IIS2132724 and IIS-1750082.

This work is done with Zhiyuan was a research intern in Azure Florence Team, Microsoft, 2021 Summer.

GitHub Repo.