ViPCap: Retrieval Text-based Visual Prompts for Lightweight Image Captioning

Taewhan Kim, Soeun Lee, Si-Woo Kim, Dong-Jin Kim
(Hanyang University, South Korea)

Introduction

Lightweight image captioning

< Using small mapping network [1] >   < Using datastore [2] >

Problems in prior research

① Computational cost
- Despite of mapping network, high computational costs are still required.

② Neglect of rich image descriptions
- Prior works only use retrieval data as a text prompt, not a visual prompt.

③ Rely on frozen CLIP image encoder
- Prior works rely solely on the frozen CLIP encoder.

Retrieval : a brown dog with big eyes on a chair
a brown dog peeks over the edge of a table
a beagle dog looking innocent standing by a fence

GT : a dog is peaking its head behind a chair

SmallCap : a beagle is peeking out of a window

ViPCap : a dog peeking out from under a wooden chair

Problem: Because the visual information in the text is **not** properly **reflected** in the image features.

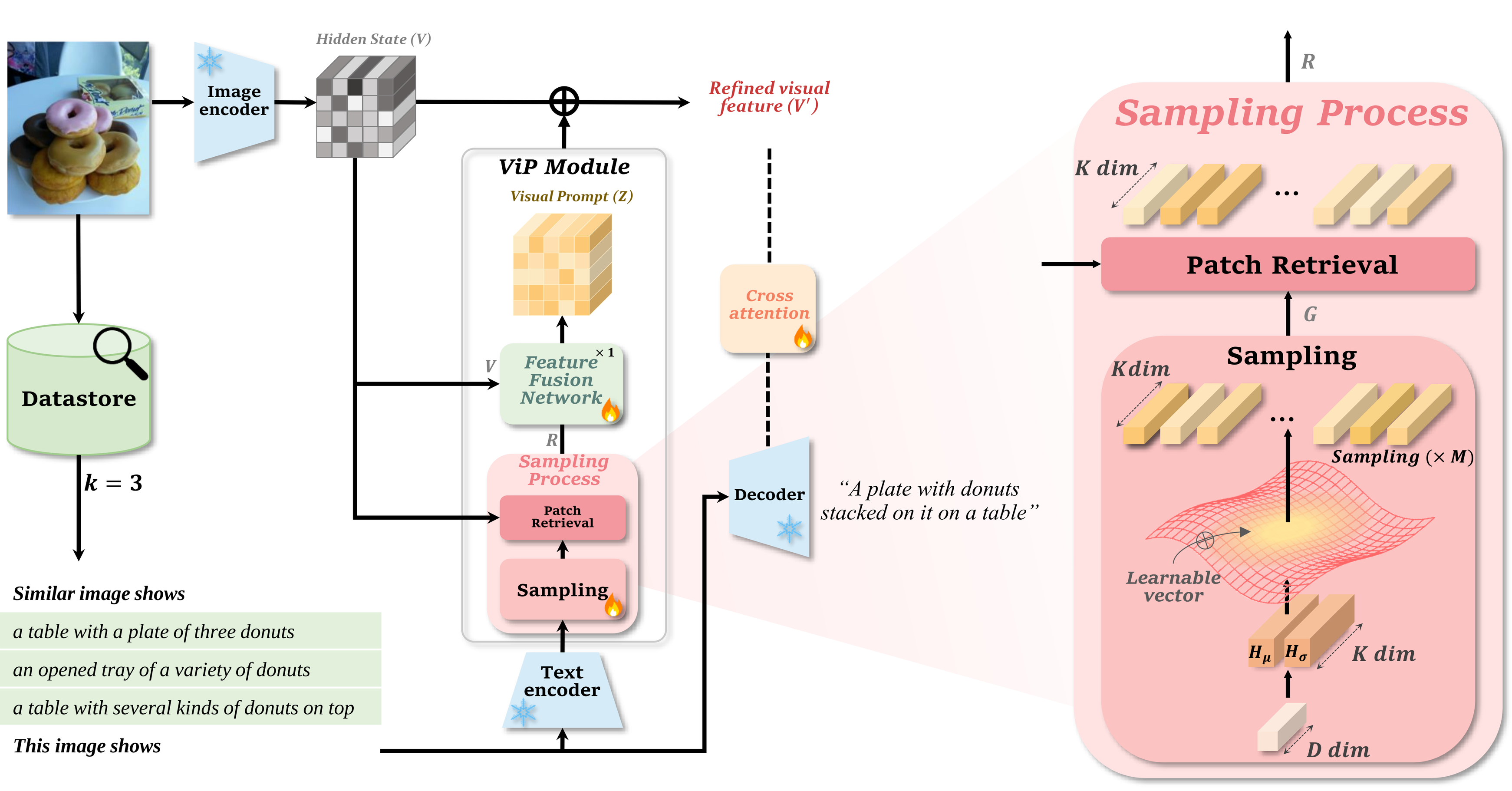
➡ **Solution**: Utilize retrieved texts (=hard prompt) as visual prompts

Proposed Framework

Contributions

- Generate **visual prompts** from **text prompts**.
- **Align** sampled semantic text features with visual representations.
- Flexible integration with **various models** and **prompts (Plug-and-play).**

Overall framework



Method

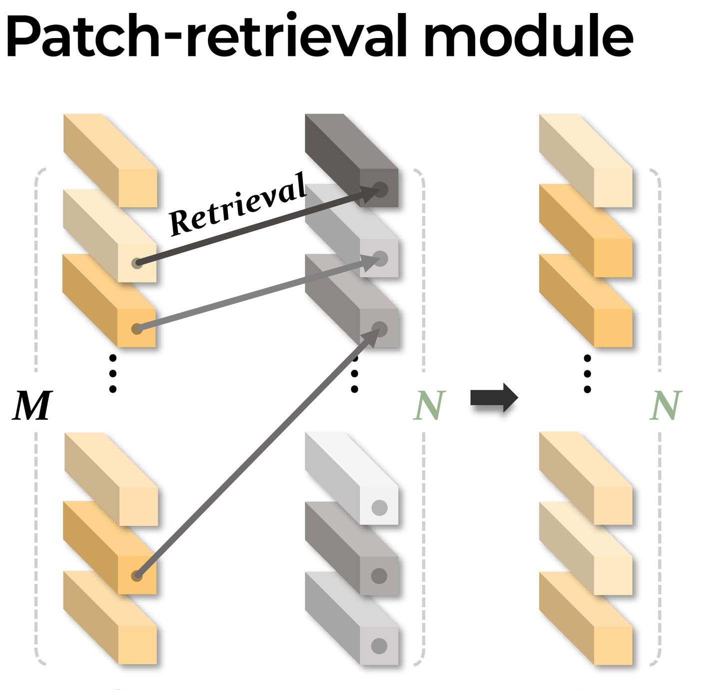
- Converts text prompts into semantic features using learnable distribution.
- Retrieves semantic features aligned with patch-level image features.
- Fuse these features with **FFN** to generate visual prompts.

Sampling process

- Sampling from Gaussian distribution
  - Capture for fine-grained visual information
- Add learnable vectors

$\times M \longrightarrow$ $\mu = H_\mu(T) + \alpha \cdot \omega,\quad \sigma = H_\sigma(T)$

Patch-retrieval module



G: Semantic feature V: Image feature
R: Retrieved semantic feature

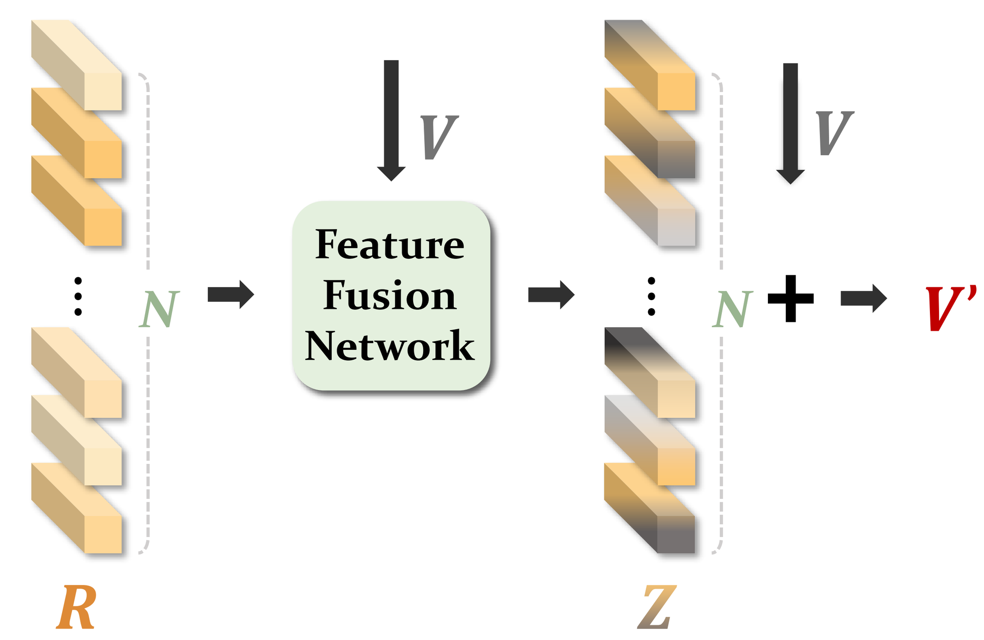
- Select **Semantic features** (G) to patch-level using cosine similarity

$$\mathcal{L}(j) = argmax_{i \in [1:M]} \mathrm{sim}(g_i, v_j)$$

- Provide the most relevant information for each patch (one to many)

$$R = \{g_{\mathcal{L}(j)}\}_{j=1}^{N} \in \mathbb{R}^{N \times K}$$

Feature Fusion Network (FFN)



Z: Visual prompt
V': Refined visual feature

- FFN (x 1 layer)
- Visual prompt
  $Z =$ FFN(V, R)
- Refined visual feature
  $V' = V + Z$

Experimental Results

Evaluation on COCO, Flickr30k, Nocaps.

Method	Training Param θ	COCO Test B@4	M	C	S	Flickr30k Test C	S	NoCaps Val In	Near	Out	Entire
Large scale training models											
$\text{OSCAR}_{\text{Large}}$ (2020)	338M	37.4	30.7	127.8	23.5	-	-	78.8	78.9	77.4	78.6
$\text{LEMON}_{\text{Huge}}$ (2022)	675M	41.5	30.8	139.1	24.1	-	-	118.0	116.3	120.2	117.3
$\text{SimVLM}_{\text{Huge}}$ (2022a)	632M	40.6	33.7	143.3	25.4	-	-	113.7	110.9	115.2	112.2
$\text{BLIP2}_{\text{ViT-g OPT}_{2.7B}}$ (2023a)	1.1B	43.7	-	145.8	-	-	-	123.0	117.8	123.4	119.7
CogVLM (2024)	1.5B	-	-	148.7	-	94.9	-	-	-	132.6	128.3
$\text{PaLI}_{\text{mT5-XXL}}$ (2023)	1.6B	-	-	149.1	-	-	-	-	-	-	127.0
Lightweight models											
CaMEL (2022)	76M	**39.1**	**29.4**	**125.7**	**22.2**	-	-	-	-	-	-
$\text{I-Tuning}_{\text{Medium}}$ (2023)	44M	35.5	28.8	120.0	22.0	**72.3**	**19.0**	89.6	77.4	58.8	75.4
ClipCap (2021)	43M	33.5	27.5	113.1	21.1	-	-	84.9	66.8	49.1	65.8
$\text{I-Tuning}_{\text{Base}}$ (2023)	14M	34.8	28.3	116.7	21.8	61.5	16.9	83.9	70.3	48.1	67.8
SmallCap (2023)	7M	37.0	27.9	119.7	21.3	60.6	-	87.6	78.6	68.9	77.9
$\text{SmallCap}_{\text{d=16, Large}}$ (2023)	47M	37.2	28.3	121.8	21.5	-	-	-	-	-	-
ViPCap (Ours)	14M	37.7	28.6	122.9	21.9	66.8	17.2	**93.8**	**81.6**	**71.5**	**81.3**

→ ViPCap demonstrates competitive performance

Evaluation on text-only captioning models.

Method	In-Domain COCO B@4	M	C	S	Flickr30k B@4	M	C	S	Cross-Domain COCO ⇒ Flickr30k B@4	M	C	S	Flickr30k ⇒ COCO B@4	M	C	S	COCO ⇒ NoCaps In	Near	Out	Entire
CapDec (2022)	26.4	25.1	91.8	-	17.7	20.0	39.1	-	17.3	18.6	35.7	-	9.2	16.3	27.3	-	60.1	50.2	28.7	45.9
CapDec+ViP	27.0	25.6	94.2	18.8	18.6	20.1	44.4	14.4	15.7	18.0	35.8	11.8	9.5	16.3	30.7	9.2	60.2	50.9	33.7	47.8
Δ	0.6	0.5	2.4	-	0.9	0.1	5.3	-	-1.6	-0.6	0.1	-	0.3	-	3.4	-	0.1	0.7	5.0	1.9
ViECap (2023)	27.2	24.8	92.9	18.2	21.4	20.1	47.9	13.6	17.4	18.0	38.4	11.2	12.6	19.3	54.2	12.5	61.1	64.3	65.0	66.2
ViECap+ViP	27.3	25.1	93.6	18.4	21.2	20.2	48.8	13.9	17.4	18.1	40.2	11.1	13.6	19.3	55.2	12.7	62.2	64.9	67.1	67.2
Δ	0.1	0.3	0.7	0.2	-0.2	0.1	0.9	0.3	-	0.1	1.8	-0.1	1.0	-	1.0	0.2	1.1	0.6	2.1	1.0

→ Demonstrate its potential as image feature

Ablation Studies

Prompt	ViP ×	ViP ✓
"This image shows"	111.1	116.0 (**4.9** ↑)
Retrieval prompt	117.3	119.9 (**2.6** ↑)

→ Achieves notable performance without using retrieval prompts

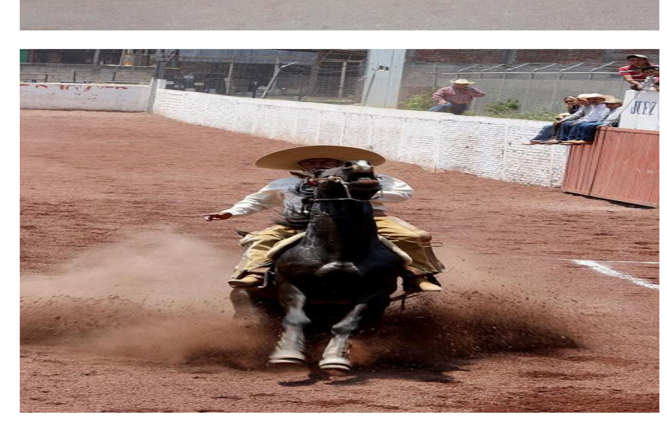
Method	Enc.	Dec.	ViP	Ret	CIDEr
ViPCap (Ours)	ViT-B/32	OPT-125M	✓	×	122.0
				✓	122.5 (**0.5** ↑)
		XGLM		×	116.8
				✓	121.2 (**4.4** ↑)
EVCap	EVA-CLIP-g	Vicuna-13B	×	×	140.1
			✓	×	141.3 (**1.2** ↑)

→ Outperformance based on various models

Qualitative Results

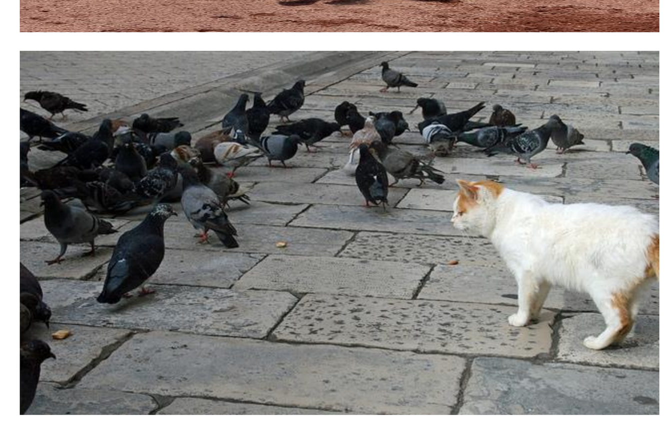
R a cow stands on the ground in front of a pink building
a cow who is looking in the door of a house on a street
a cow walking by a white building and a large decorated pole

G A white cow walking down the street next to some motorcycles
S a cow walking down a street next to a building
V a cow walking down a street next to a pink building



R the charro or mexican cowboy is skilled in many techniques for control of animal behavior including using roping
a man is riding a bucking brown horse
a man in an arena rides a bucking horse

G A man in a hat riding a horse
S a man riding a horse in a dirt field
V a man in cowboy hat riding a horse



R a cat sits in the middle of pavement as a bird swoops down
a bull and flock of pigeons on a street
an animal surrounded by a flock of birds

G A cat looking at a large group of pigeons.
S a white and black cat sitting on top of pigeons
V an orange and white cat sitting on the ground near a bunch of pigeons

References

[1] Junnan Li, Dongxu Li et al., "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models", ICML 2023
[2] Ramos, Rita and Martins, Bruno and Elliott et al., "SmallCap: Lightweight Image Captioning Prompted With Retrieval Augmentation", CVPR 2023
[3] Li, Jiaxuan and Vo et al., "EVCap: Retrieval-Augmented Image Captioning with External Visual-Name Memory for Open-World Comprehension", CVPR 2024