# IFCap: Image-like Retrieval and Frequency-based Entity Filtering for Zero-shot Captioning

Soeun Lee*, Si-Woo Kim*, Taewhan Kim, Dong-Jin Kim

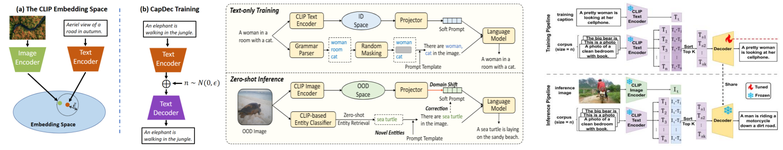
(Hanyang University, South Korea)

github.com/boreng0817/IFCap

## Introduction

### Text-only Training

- Leverages the ability of CLIP to effectively align images with related text [1, 2, 3].
- Recent research has focused on what additional cues can be used.




Noise injection into input text features [1]

(Train) Noun parsing / (Test) Noun retrieval from input text features [2]



Retrieved captions that similar to the input text [3]

### Limitation of previous work

① Modality gap

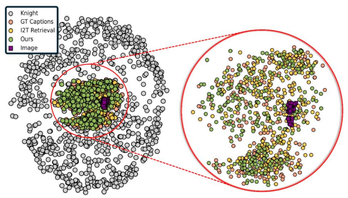
- Despite CLIP's strong performance in many works, a modality gap still exists.

② Rely on retrieved captions

- Relying solely on the retrieved sentences without using input senten ce.

③ Low detection rate of nouns

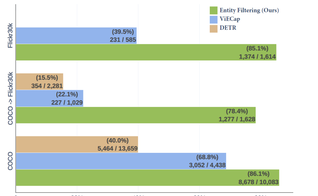
- Previous work using additional cues for training had insufficient info rmation.




text-to-text retrieval method ( from previous work

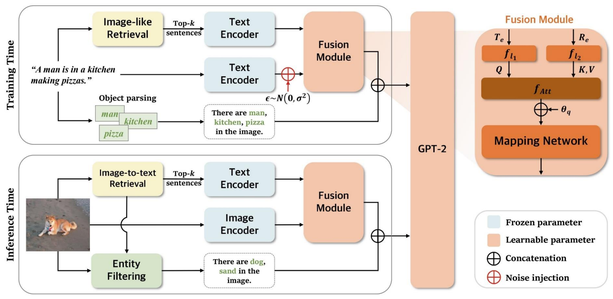
≠

image-to-text retrieval method ( )

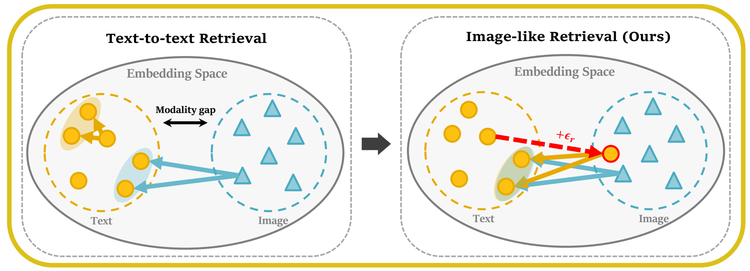



Precision score: $\frac{\#\ of\ correctly\ extracted\ words}{\#\ of\ all\ extracted\ words}$

## Proposed Framework




- **Image-like Retrieval (ILR)**




- (Problem) Information used during training ≠ Information used during testing
- (Solution) Noise injection ($N(0, \sigma^2) \sim \epsilon_r$)

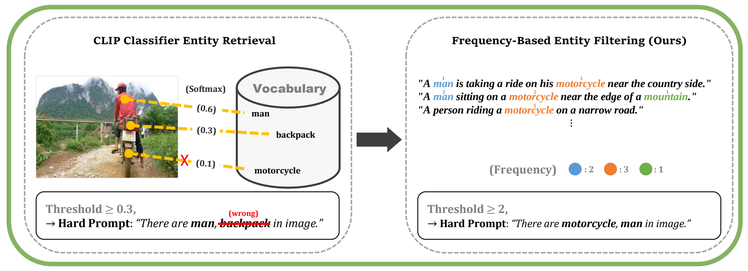
- **Fusion Module (FM)**
  - The input text is adjusted through noise injection to bring it closer to the image feature space.

- **Auto-regressive Loss**

$$L_\theta = -\frac{1}{N}\sum_{i=1}^{N} \log(y_i | F; h; y_{<i}; \theta)$$

$F$: Output from the mapping network
$h$: Hard prompt
$y$: Ground truth caption

- **Frequency-based Entity Filtering (EF)**




- (Problem) Vocabulary size ↑ ⇒ Probability of detecting each word ↓
- (Solution) Nouns are extracted from the top $l$ sentences retrieved via image-to-text retrieval, and the frequency of each noun is measured.

## Experimental Results

### Evaluation on In-domain Setup

| Method | Image Encoder | Text Decoder | COCO B@4 | COCO M | COCO C | COCO S | Flickr30k B@4 | Flickr30k M | Flickr30k C | Flickr30k S |
|---|---|---|---|---|---|---|---|---|---|---|
| CapDec (2022) | RN50x4 | $GPT\text{-}2_{Large}$ | 26.4 | 25.1 | 91.8 | 11.9 | 17.7 | 20.0 | 39.1 | 9.9 |
| DeCap (2023) | ViT-B/32 | $Transformer_{Base}$ | 24.7 | 25.0 | 91.2 | 18.7 | 21.2 | 21.8 | 56.7 | 15.2 |
| CLOSE (2022) | ViT-L/14 | $T5_{base}$ | - | - | 95.3 | - | - | - | - | - |
| ViECap (2023) | ViT-B/32 | $GPT\text{-}2_{Base}$ | 27.2 | 24.8 | 92.9 | 18.2 | 21.4 | 20.1 | 47.9 | 13.6 |
| $MeaCap_{InvLM}$ (2024) | ViT-B/32 | $GPT\text{-}2_{Base}$ | 27.2 | 25.3 | 95.4 | 19.0 | 22.3 | 22.3 | 59.4 | 15.6 |
| Knight (2023) | RN50x64 | $GPT\text{-}2_{Large}$ | 27.8 | 26.4 | 98.9 | 19.6 | 22.6 | **24.0** | 56.3 | 16.3 |
| ICSD♠ (2023) | ViT-B/32 | $BERT_{Base}$ | 29.9 | 25.4 | 96.6 | - | **25.2** | 20.6 | 54.3 | - |
| SynTIC♠† (2023) | ViT-B/32 | $Transformer^{L=4}_{H=4}$ | 29.9 | 25.8 | 101.1 | 19.3 | 22.3 | 22.4 | 56.6 | 16.6 |
| IFCap | ViT-B/32 | $GPT\text{-}2_{Base}$ | **30.8** | **26.7** | **108.0** | **20.3** | 23.5 | 23.0 | **64.4** | **17.0** |

### Evaluation on Cross-domain Setup

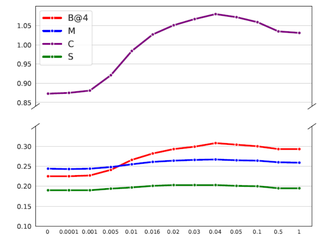
| Method | COCO ⟹ Flickr B@4 | M | C | S | Flickr ⟹ COCO B@4 | M | C | S |
|---|---|---|---|---|---|---|---|---|
| DeCap (2023) | 16.3 | 17.9 | 35.7 | 11.1 | 12.1 | 18.0 | 44.4 | 10.9 |
| ViECap (2023) | 17.4 | 18.0 | 38.4 | 11.2 | 12.6 | 19.3 | 54.2 | 12.5 |
| Knight (2023) | 21.1 | **22.0** | 48.9 | 14.2 | 19.0 | 22.8 | 64.4 | 15.1 |
| SynTIC (2023) | 17.9 | 18.6 | 38.4 | 11.9 | 14.6 | 19.4 | 47.0 | 11.9 |
| SynTIC-*TT* | 19.4 | 20.2 | 43.2 | 13.9 | **20.6** | 21.3 | 64.4 | 14.3 |
| IFCap* | 17.8 | 19.4 | 47.5 | 12.7 | 14.7 | 20.4 | 60.7 | 13.6 |
| IFCap-*TT* | **21.2** | 21.8 | **59.2** | **15.6** | 19.0 | **23.0** | **76.3** | **17.3** |

| Method | COCO ⟹ NoCaps Val In C | In S | Near C | Near S | Out C | Out S | Entire C | Entire S |
|---|---|---|---|---|---|---|---|---|
| DeCap (2023) | 65.2 | - | 47.8 | - | 25.8 | - | 45.9 | - |
| CapDec (2022) | 60.1 | 10.2 | 50.2 | 9.3 | 28.7 | 6.0 | 45.9 | 8.3 |
| ViECap (2023) | 61.1 | 10.4 | 64.3 | 9.9 | 65.0 | 8.6 | 66.2 | 9.5 |
| IFCap* | **70.1** | **11.2** | **72.5** | **10.9** | **72.1** | **9.6** | **74.0** | **10.5** |

### Evaluation on Video Captioning

| Method | MSR-VTT B@4 | M | C | S | MSVD B@4 | M | C | S |
|---|---|---|---|---|---|---|---|---|
| ZeroCap (2022b) | 2.3 | 12.9 | 5.8 | - | 2.9 | 16.3 | 9.6 | - |
| MAGIC (2022) | 5.5 | 13.3 | 7.4 | 4.2 | 6.6 | 16.1 | 14.0 | 2.9 |
| CLMs (2022) | 6.2 | 17.8 | 10.1 | 6.5 | 7.0 | 16.4 | 20.0 | 3.1 |
| CapDec (2022) | 8.9 | 23.7 | 11.5 | 5.9 | 7.9 | 23.3 | 34.5 | 3.2 |
| EPT (2022a) | 3.0 | 14.6 | 11.3 | - | 3.0 | 17.8 | 17.4 | - |
| Knight (2023) | 25.4 | **28.0** | 31.9 | **8.5** | 37.7 | **36.1** | 63.8 | 5.0 |
| IFCap | **27.1** | 25.9 | **38.9** | 6.7 | **40.6** | 34.2 | **83.9** | **6.3** |

### Hyper-parameter search




## References

[1] Nukrai, David, Ron Mokady, and Amir Globerson. "Text-only training f or image captioning using noise-injected clip." arXiv preprint arXiv:22 11.00575 (2022).

[2] Fei, Junjie, et al. "Transferable decoding with visual entities for zero-s hot image captioning." Proceedings of the IEEE/CVF International Co nference on Computer Vision. 2023.

[3] Wang, Junyang, et al. "From association to generation: Text-only capt ioning by unsupervised cross-modal mapping." arXiv preprint arXiv:2 304.13273 (2023).