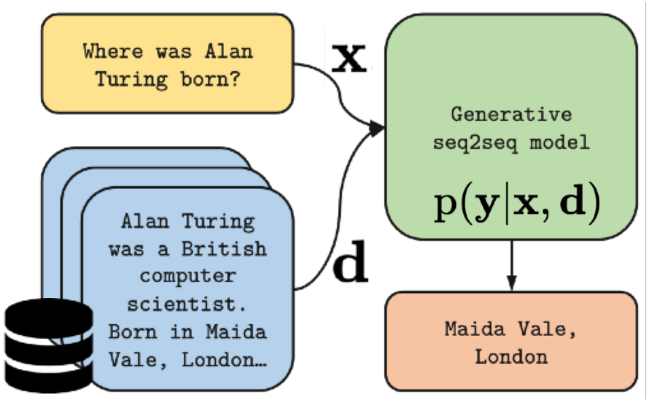
Experiments in Retrieval-Augmented Image Captioning

~~Rita Ramos~~ Desmond Elliott
Department of Computer Science
University of Copenhagen

Retrieval Augmented Generation

- Combine the power of in-weights learning with in-context adaptation through retrieval augmentation
- Given a datastore of facts, knowledge, documents, etc.
  - Combine the most relevant items from the datastore (d) with the input (x) for your task



Izacard and Grave. EACL 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering.

Multimodal Retrieval Augmentation

- Combine the most relevant items from the datastore with the input

a white polar bear laying on top of a rock
a brown bear sleeping on a big rock
a dog is sleeping on a rock
a white bear sleeping on a rocky ledge
a white bear is laying out on the rocks

?

IMAGE

NEAREST CAPTIONS

GENERATED CAPTION

Multimodal Retrieval Augmentation

- Combine the most relevant items from the datastore with the input

a white polar bear laying on top of a rock
a brown bear sleeping on a big rock
a dog is sleeping on a rock
a white bear sleeping on a rocky ledge
a white bear is laying out on the rocks

?

IMAGE

NEAREST CAPTIONS

GENERATED CAPTION

Multimodal Retrieval Augmentation

- Combine the most relevant items from the datastore with the input

a white polar bear laying on top of a rock
a brown bear sleeping on a big rock
a dog is sleeping on a rock
a white bear sleeping on a rocky ledge
a white bear is laying out on the rocks

→ a white polar bear sleeps on a rock

IMAGE | NEAREST CAPTIONS | GENERATED CAPTION

… at ACL 2025

… at ACL 2025

Towards Text-Image Interleaved Retrieval

Xin Zhang[1,2*], Ziqi Dai[1*], Yongqi Li[2], Yanzhao Zhang, Dingkun Long
Pengjun Xie, Meishan Zhang[1†], Jun Yu[1], Wenjie Li[2], Min Zhang[1]
[1]Harbin Institute of Technology, Shenzhen [2]The Hong Kong Polytechnic University
{zhangxin2023,ziqi.dai}@stu.hit.edu.cn zhangmeishan@hit.edu.cn
Release at https://github.com/vec-ai/wikiHow-TIIR

… at ACL 2025

Towards Text-Image Interleaved Retrieval

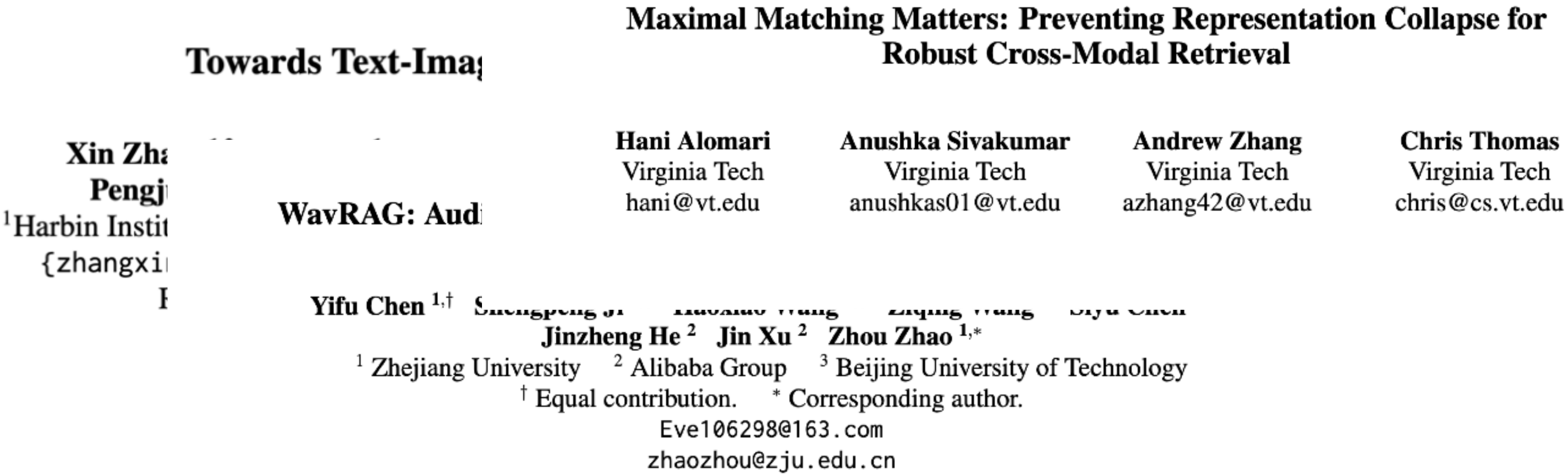
Xin Zha
Pengj
[1]Harbin Instit
{zhangxi
F

WavRAG: Audio-Integrated Retrieval Augmented Generation for Spoken Dialogue Models

Yifu Chen [1,†] Shengpeng Ji [1,†] Haoxiao Wang [1,†] Ziqing Wang [3] Siyu Chen [1]
Jinzheng He [2] Jin Xu [2] Zhou Zhao [1,*]

[1] Zhejiang University [2] Alibaba Group [3] Beijing University of Technology
[†] Equal contribution. [*] Corresponding author.
Eve106298@163.com
zhaozhou@zju.edu.cn

… at ACL 2025



Maximal Matching Matters: Preventing Representation Collapse for Robust Cross-Modal Retrieval

Hani Alomari Virginia Tech hani@vt.edu
Anushka Sivakumar Virginia Tech anushkas01@vt.edu
Andrew Zhang Virginia Tech azhang42@vt.edu
Chris Thomas Virginia Tech chris@cs.vt.edu

Towards Text-Ima

Xin Zha
Pengj
[1]Harbin Instit
{zhangxi

WavRAG: Audi

Yifu Chen [1,†]
Jinzheng He [2] **Jin Xu** [2] **Zhou Zhao** [1,*]
[1] Zhejiang University [2] Alibaba Group [3] Beijing University of Technology
[†] Equal contribution. [*] Corresponding author.
Eve106298@163.com
zhaozhou@zju.edu.cn

… at ACL 2025

Towards Text-Ima

Xin Zh
Pengj
[1]Harbin Instit
{zhangxi

WavRAG: Aud

Yifu Chen [1,†]
Jinzheng He [2] **Jin Xu** [2] **Zhou Zhao** [1,*]
[1] Zhejiang University [2] Alibaba Group [3] Beijing University of Technology
[†] Equal contribut
Eve
zhao

Maximal Matching Matters: Preventing Representation Collapse for Robust Cross-Modal Retrieval

Hani Alomari Virginia Tech hani@vt.edu
Anushka Sivakumar Virginia Tech anushkas01@vt.edu
Andrew Zhang Virginia Tech azhang42@vt.edu
Chris Thomas Virginia Tech chris@cs.vt.edu

VISA: Retrieval Augmented Generation with Visual Source Attribut

Xueguang Ma [*,1] **Shengyao Zhuang** [*2,3] **Bevan Koopman** [2,3]
Guido Zuccon [3] **Wenhu Chen** [1] **Jimmy Lin** [1]
[1]University of Waterloo [2]CSIRO [3] University of Queensland
x93ma@uwaterloo.ca, s.zhuang@uq.edu.au

… at ACL 2025

Towards Text-Imag

Xin Zha
Pengj
[1]Harbin Instit
{zhangxi

WavRAG: Aud

Maximal Matching Matters: Preventing Representation Collapse for Robust Cross-Modal Retrieval

Hani Alomari	Anushka Sivakumar	Andrew Zhang	Chris Thomas
Virginia Tech	Virginia Tech	Virginia Tech	Virginia Tech
hani@vt.edu	anushkas01@vt.edu	azhang42@vt.edu	chris@cs.vt.edu

OMGM: Orchestrate Multiple Granularities and Modalities for Efficient Multimodal Retrieval

Wei Yang[*], Jingjing Fu[†], Rui Wang, Jinyu Wang, Lei Song, Jiang Bian
Microsoft Research Asia
wyang6621@gmail.com, {jifu,ruiwa,jinywan,lesong,jiabia}@microsoft.com

eneration with Visual Source Attribut

gyao Zhuang[*2,3] **Bevan Koopman[2,3]**
Wenhu Chen[1] **Jimmy Lin[1]**
[2]CSIRO [3] University of Queensland
o.ca, s.zhuang@uq.edu.au

… at ACL 2025

Towards Text-Imag

Xin Zha
Pengj
[1]Harbin Instit
{zhangxi

WavRAG: Audi

Maximal Matching Matters: Preventing Representation Collapse for Robust Cross-Modal Retrieval

Hani Alomari	Anushka Sivakumar	Andrew Zhang	Chris Thomas
Virginia Tech	Virginia Tech	Virginia Tech	Virginia Tech
hani@vt.edu	anushkas01@vt.edu	azhang42@vt.edu	chris@cs.vt.edu

fficient

Ask in Any Modality
A Comprehensive Survey on Multimodal Retrieval-Augmented Generation

Mohammad Mahdi Abootorabi[‡♣†], Amirhosein Zobeiri[◇], Mahdi Dehghani[¶], Mohammadali Mohammadkhani[§], Bardia Mohammadi[♠], Omid Ghahroodi[†], Mahdieh Soleymani Baghshah[§,*], Ehsaneddin Asgari[†,*]

[†]Qatar Computing Research Institute, [‡]Saarland University, [♣] Zuse School ELIZA, [◇]University of Tehran, [♠]Max Planck Institute for Software Systems, [¶]K.N. Toosi University of Technology, [§]Sharif University of Technology

Correspondence: soleymani@sharif.edu and easgari@hbku.edu.qa

https://multimodalrag.github.io

.com

eneration with Visual Source Attribut

gyao Zhuang[,*2,3] Bevan Koopman[2,3]
Wenhu Chen[1] Jimmy Lin[1]
[2]CSIRO [3] University of Queensland
o.ca, s.zhuang@uq.edu.au

… at ACL 2025

Towards Text-Ima

Xin Zha
Pengj
[1]Harbin Instit
{zhangxi

WavRAG: Aud

Maximal Matching Matters: Preventing Representation Collapse for Robust Cross-Modal Retrieval

Hani Alomari	Anushka Sivakumar	Andrew Zhang	Chris Thomas
Virginia Tech	Virginia Tech	Virginia Tech	Virginia Tech
hani@vt.edu	anushkas01@vt.edu	azhang42@vt.edu	chris@cs.vt.edu

MegaPairs: Massive Data Synthesis for Universal Multimodal Retrieval

Junjie Zhou[1,2*], **Yongping Xiong**[1*], **Zheng Liu**[2,5†], **Ze Liu**[3,2], **Shitao Xiao**[2], **Yueze Wang**[2], **Bo Zhao**[4], **Chen Jason Zhang**[5], **Defu Lian**[3†]
[1] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
[2] Beijing Academy of Artificial Intelligence
[3] University of Science and Technology of China
[4] Shanghai Jiao Tong University, [5] The Hong Kong Polytechnic University
{junjiebupt, zhengliu1026}@gmail.com, liandefu@ustc.edu.cn

A Comprehensive Survey o

Mohammad Mahdi Abootorabi[‡♣†], **An**
Bardia Mohammadi[♠], **Omid Gh**
[†]Qatar Computing Research Institut
[♠]Max Planck Institute for Software Systems, K.N. Toosi University of Technology, Sharif University of Technology
Correspondence: soleymani@sharif.edu and easgari@hbku.edu.qa

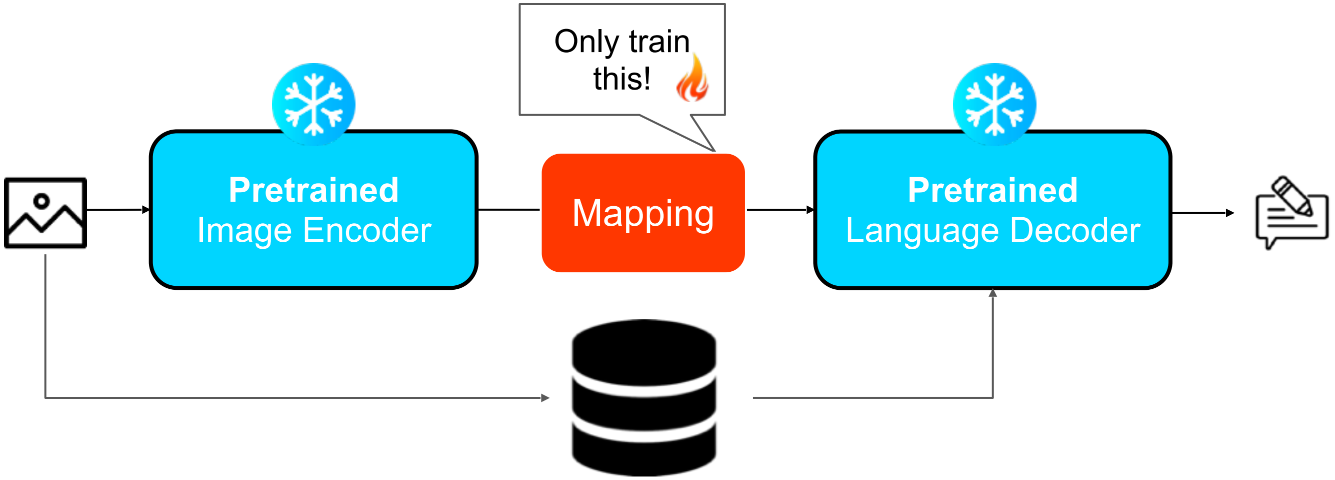
https://multimodalrag.github.io

ource Attribut

oopman[2,3]
n[1]
ueensland
o.ca, s.zhuang@uq.edu.au

Our Motivation

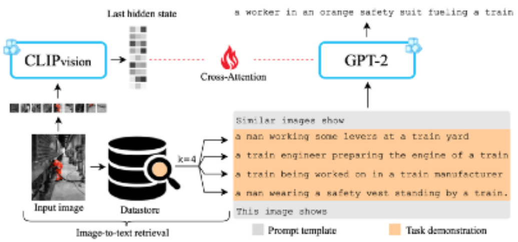
- Train lightweight image captioning models using frozen backbones
  - CLIPCap (Mokady et al. 2021), I-Tuning (Luo et al. 2023)
- … and using retrieval augmentation to *assist* the decoder



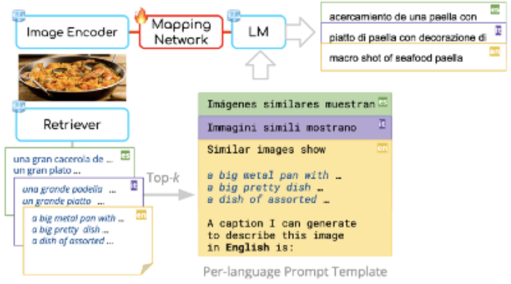
Mokady et al. 2021. ClipCap: CLIP Prefix for Image Captioning.
Luo et al. ICASSP 2023. I-Tuning: Tuning Frozen Language Models with Image for Lightweight Image Captioning.

Overview

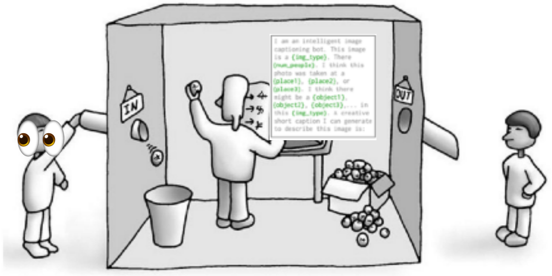
1. **Lightweight** RAG Captioning



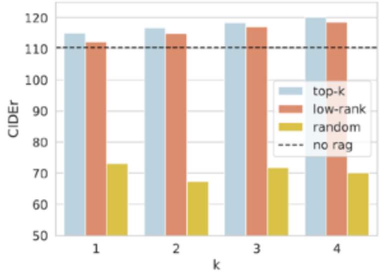
2. Lightweight **Multilingual** Training



3. **Image-blind** captioning



4. **Understanding** Multimodal RAG



SmallCap: Lightweight Image Captioning Prompted with Retrieval Augmentation

CVPR 2023


R. Ramos

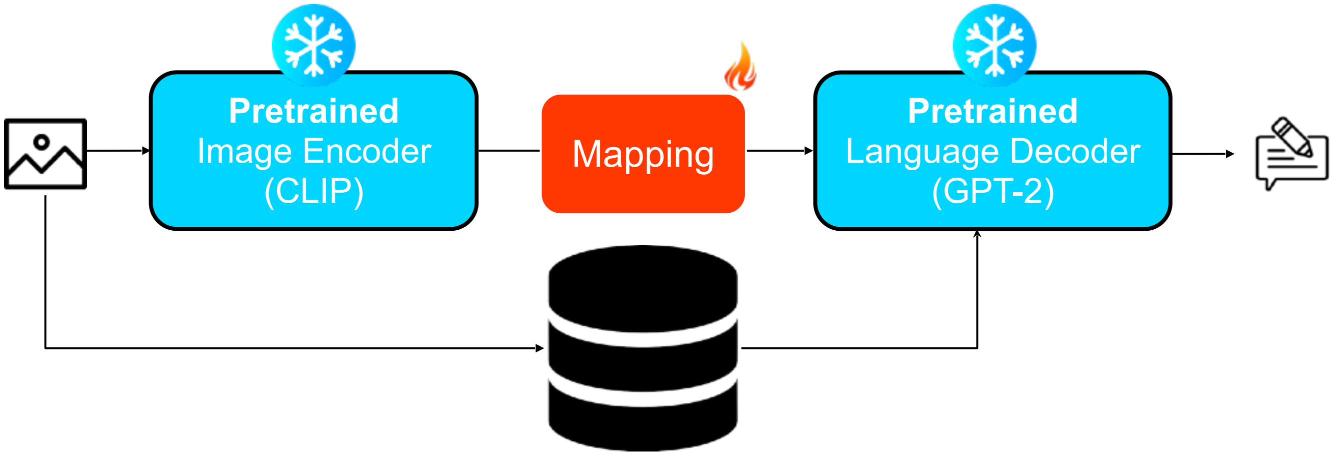

B. Martins


D. Elliott

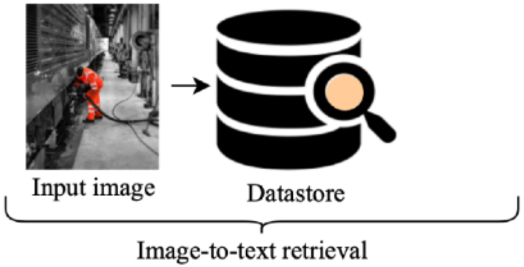

Y. Kementchedjhieva

Lightweight Training trough Retrieval

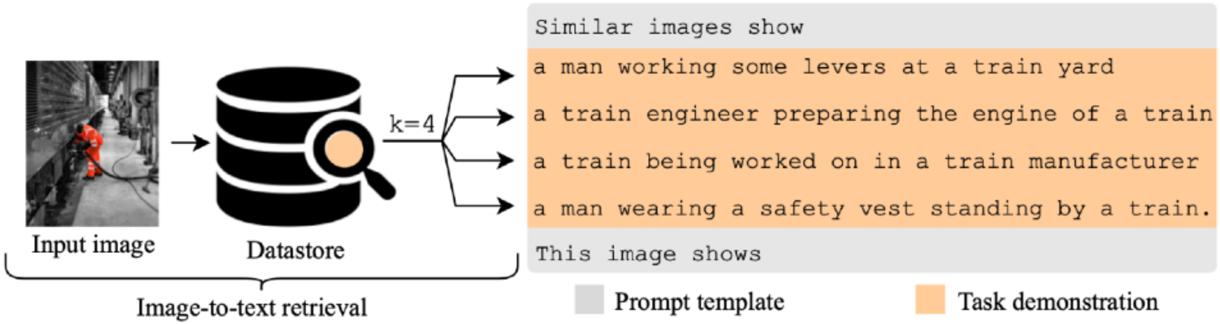
- Given the success of retrieval augmented generation, can we extend this to multimodality with a lightweight training paradigm?



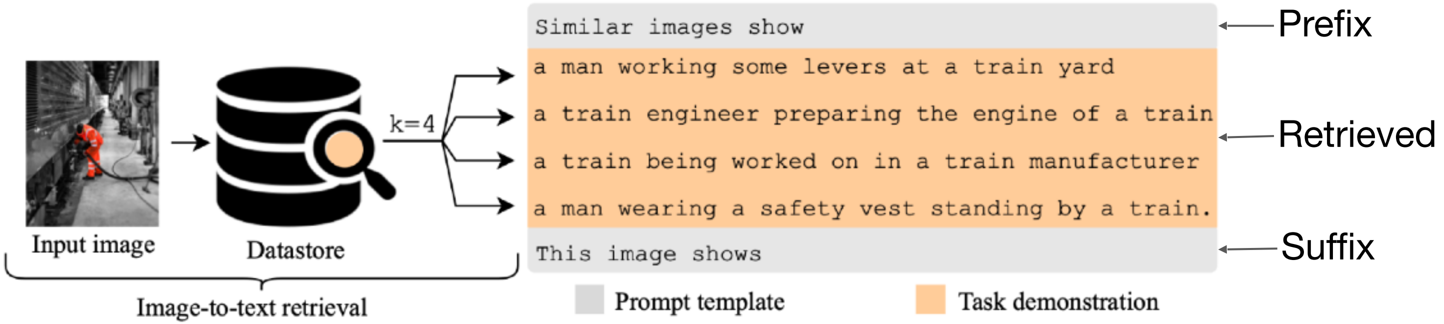
SmallCap Model



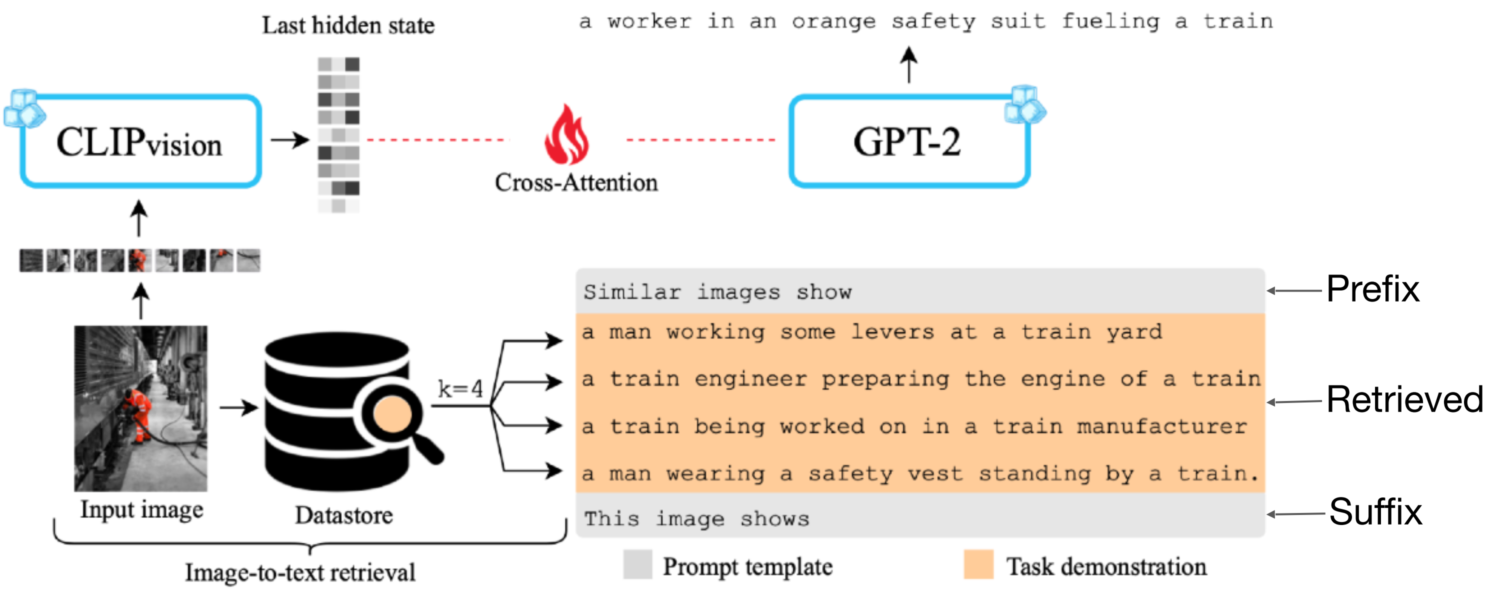
SmallCap Model



SmallCap Model

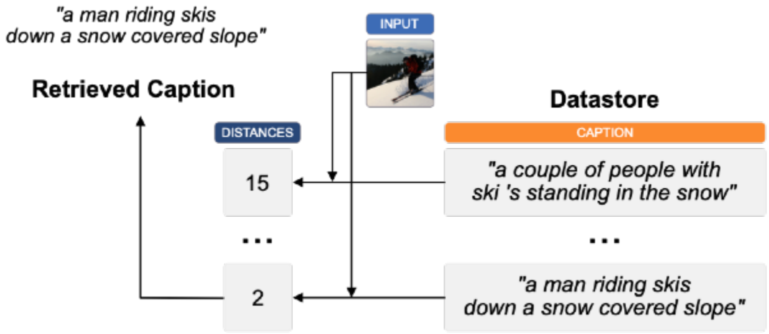


SmallCap Model



Retrieval System

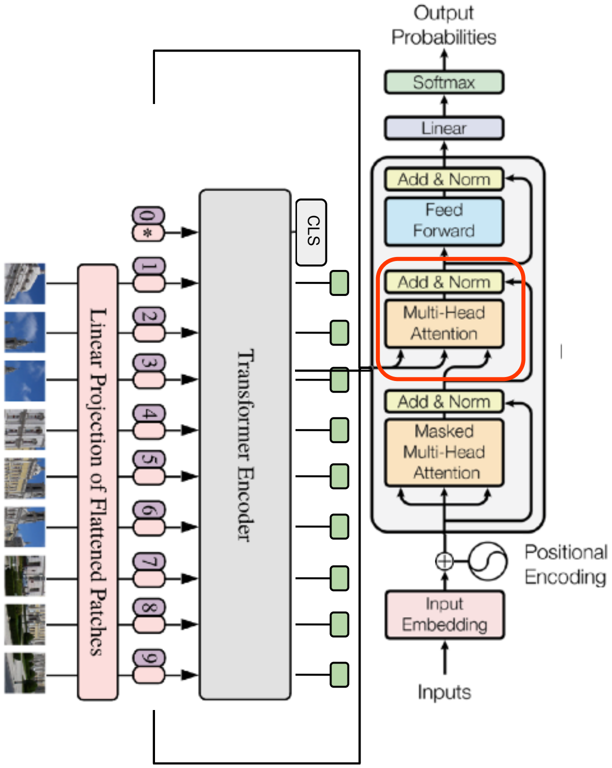
- Build a datastore with high-dimensional dense vectors
  - FAISS: Facebook AI Similarity Search for nearest-neighbor search
  - Captions of images represented with CLIP embeddings
- Retrieve k nearest-neighbours captions from datastore
  - Image embedding compared against datastore caption vectors



Johnson et al. IEEE Big Data 2019. Billion-scale similarity search with GPUs.

Trained Cross-Attention Layers

- Autoregressive Transformer LMs only contain a multi-head **self-attention mechanism**
- We insert a randomly initialized **cross-attention mechanism** to attend to the visual encoder output embeddings



Experimental Setup

- Pretrained CLIP-ViT-B/32 and GPT/OPT backbone models
- Randomly initialize the cross-attention layer
- Train only on COCO in only 8 hours on 1 x 40GB NVIDIA A100 GPU

Low-rank cross-attention

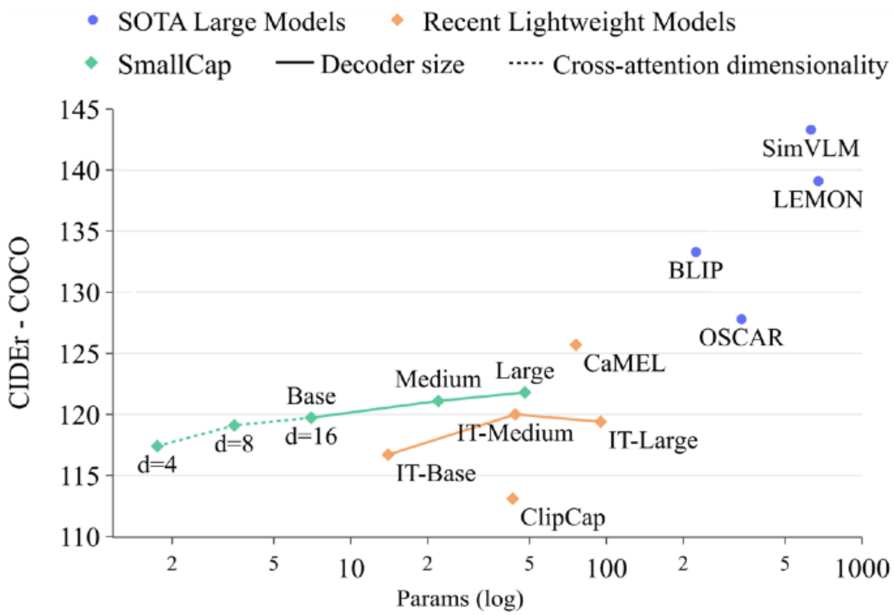
$$\mathrm{Att}(\mathbf{QW}_i^Q, \mathbf{KW}_i^K, \mathbf{VW}_i^V)$$

$$W_i^K, W_i^Q, W_i^V$$

$$\in R^{d\_encoder \times \mathbf{d}}$$

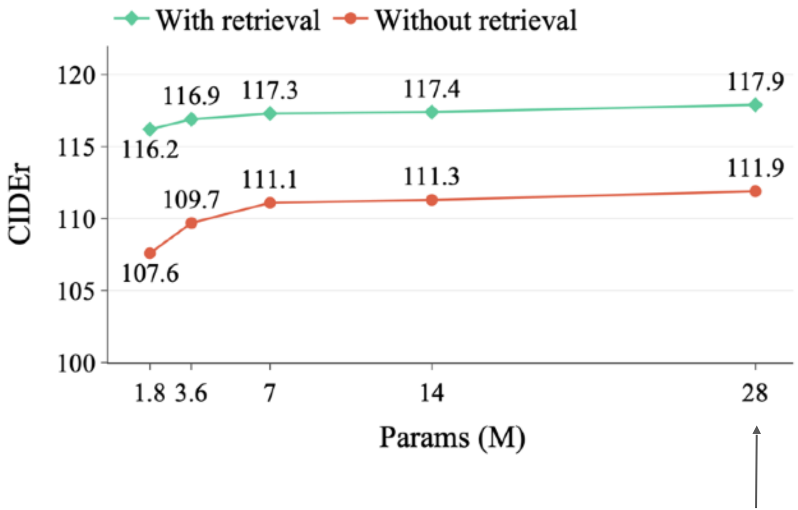
Attention rank	Params
d=64 (Full)	22M
d=16	7M
d=8	3.6M
d=4	1.8M

Results



- Outperform other lightweight approaches
- Effective with low-rank matrices: 4,8,16 << 64
- Larger pretrained decoders further improve performance

Importance of Retrieval Augmentation

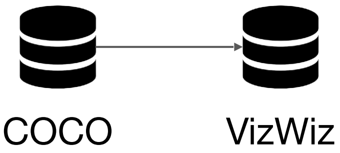


Full rank

- With retrieval:
  - Performance is stable across the range of cross-attention sizes
- Without retrieval:
  - Drop in performance
  - SmallCap model performance degrades at a higher rate

Training-Free Domain Transfer

- SmallCap was trained on COCO but we can easily swap the datastore



	Flickr30k	VizWiz	MSR-VTT
ClipCap	41.2	28.3	12.5
CaMEL	55.2	37.6	20.7
SmallCap	**60.6**	**55.0**	**28.4**

Qualitative Example from VizWiz

COCO

- some carrots potatoes garlic an onion and some chicken broth
- a selection of ingredients for soup includes carrots, meat, and prepackaged broth
- this is the makings of a meal with chicken and vegetables
- the meal has chicken, bread, and cole slaw

Generated caption:

a close up of a plate of food on a table

Qualitative Example from VizWiz

COCO

- some carrots potatoes garlic an onion and some chicken broth
- a selection of ingredients for soup includes carrots, meat, and prepackaged broth
- this is the makings of a meal with chicken and vegetables
- the meal has chicken, bread, and cole slaw

Generated caption:

a close up of a plate of food on a table

VizWiz

- a can of swanson fat free chicken broth
- a can of swanson brand chicken broth with less sodium
- a 14,5 ounce can of swanson branded chicken broth
- a can of swanson chicken broth on a table

Generated caption:

a can of swanson brand chicken broth on a table

Qualitative Example from VizWiz

COCO

- some carrots potatoes garlic an onion and some chicken broth
- a selection of ingredients for soup includes carrots, meat, and prepackaged broth
- this is the makings of a meal with chicken and vegetables
- the meal has chicken, bread, and cole slaw

Generated caption:

a close up of a plate of food on a table

VizWiz

- a can of swanson fat free chicken broth
- a can of swanson brand chicken broth with less sodium
- a 14,5 ounce can of swanson branded chicken broth
- a can of swanson chicken broth on a table

Generated caption:

a can of swanson brand chicken broth on a table

“swanson” does not appear anywhere in the COCO training dataset

Q: What about multilingual captioning?

PAELLA: Parameter-Efficient Lightweight Language-agnostic Captioning Model

Findings of NAACL 2024


R. Ramos


E. Bugliarello


B. Martins


D. Elliott

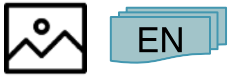
Multilingual Image Caption Training

- Common approach in the literature is to machine translate and train

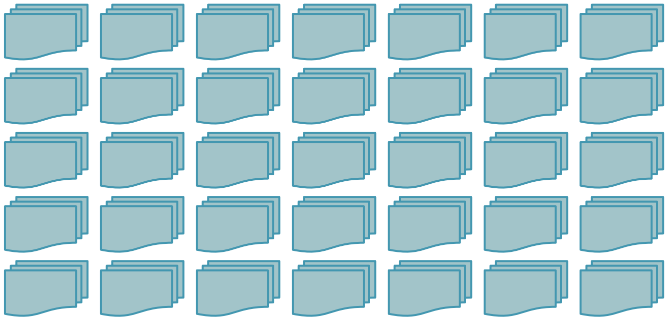
COCO —translate→ **COCO-35**

COCO
- 113k images
  - 5 english captions

COCO-35
- 113k images
  - 5*35L captions



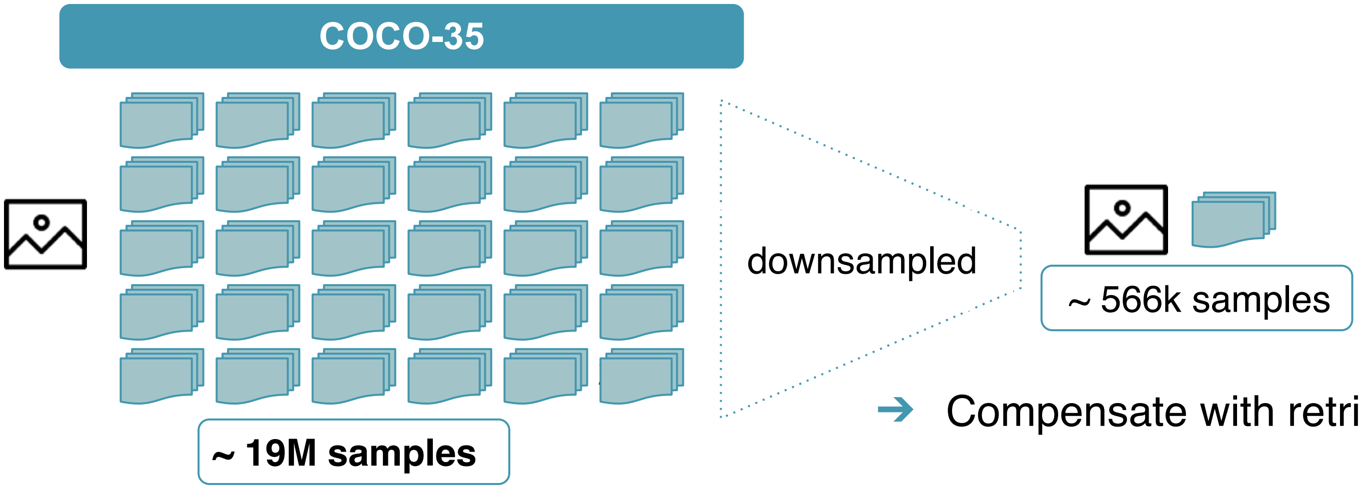
~ 566k samples



~ 19M samples

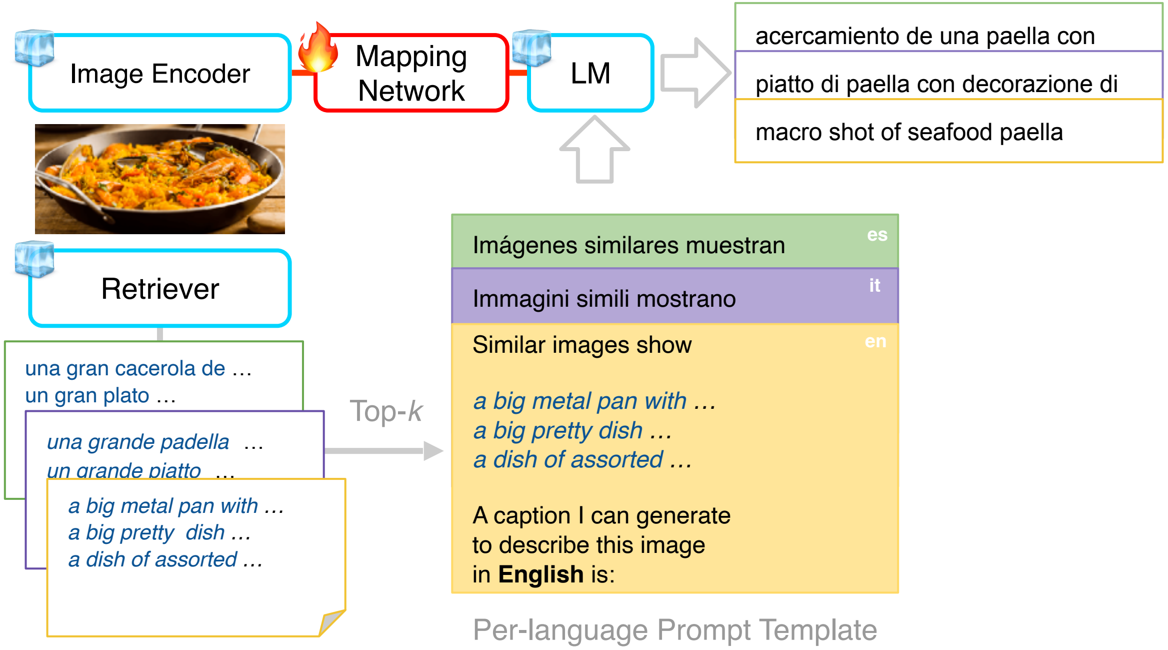
Data-Efficient Multlingual Training

- Only train on a subset of COCO-35:
  - Sample uniformly across 35 languages
  - **Match the size** of the English COCO dataset



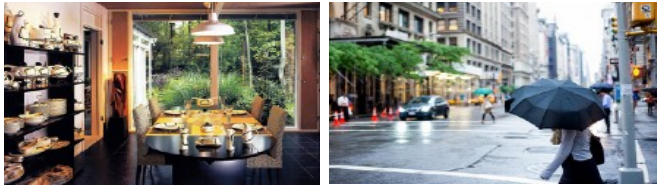
➔ Compensate with retrieved examples

PAELLA Model



Experimental Protocol

- Encoder: Multilingual CLIP
- Decoder: XGLM-2.9B
- Training data:
  - 566K captions sampled from COCO-35
- Evaluation: XM-3600
  - 3600 geographically-diverse images
  - 36 languages with 100 captions per image
  - 5 low-resource languages (L5):
    - Bengali, Cusco Quechua, Maori, Swahili, Telugu



Example training images from COCO

Examples evaluation images from XM3600

Results

	Data	Trained θ	L36	L5
PaLI	12B	17B	53.6	-
Lg $_{\text{COCO-35}}$	19M	2.6B	15.0	12.5
mBLIP: BLOOMZ-7B	135M	800M	23.4	6.7
BB+CC $_{\textit{COCO-35 + CC-35}}$	135M	800M	28.5	22.4
mBLIP: mT0-XL	489M	124M	28.3	7.9
PAELLA	**566K**	**30M**	26.2	20.7

PAELLA is competitive against models with 35-863x more training data, and 4-87x more trained parameters

Zero-shot Multilingual Transfer

- **$\text{PAELLA}_{\text{mono}}$** is a variant trained on 566K examples in English COCO
- Outperforms **Lg** trained on 19.8M examples in the machine translated COCO-35 dataset

	Data	Trained Θ	L36	L5
Lg: Thapliyal et al. $_{\text{COCO-35}}$	19M	2.6B	15.0	12.5
$\text{PAELLA}_{\text{mono}}$	566K$_{\text{en}}$	30M	15.5	12.1

Qualitative Example

类似图片显示:

ऐसी ही तस्वीरें दिखाती हैं:

Imágenes similares muestran:

Similar images show:

the owl is perched outside in front of the people
an owl sitting a top a table during the daytime
an owl is sitting on a perch at a camp site
the fuzzy owl is sitting on a tree branch

A caption I can generate to describe this image in **english** is:

PAELLA

en: *"an owl sitting on top of a tree"*

es: *"un búho sentado en una rama de un árbol"*
(an owl sitting on a tree branch)

hi: *"एक उल्लू एक पेड़ की टहनी पर बैठा है"*
(an owl is sitting on a tree branch)

zh: *"一只 猫头鹰 站在 树上"*
(an owl standing in a tree)

NoRAG

en: *"a large black and white picture of a bird"*

es: *"un pájaro posado en la parte superior de un edificio"*
(a bird perched on the top of a building)

hi: *"एक पेड़ के पास खड़ा एक पक्षी"*
(a bird standing near a tree)

zh: *"一只 长颈鹿 坐在 树枝 上"*
(a giraffe sitting on a branch)

Q: Do you even train?

LMCap: Few-shot Multilingual Image Captioning by Retrieval Augmented Language Model Prompting

Findings of ACL 2023



R. Ramos B. Martins D. Elliott

Socratic Models

- Enable models to “communicate” with each other through their output labels, prompting, and ranking

$$f^3_{\text{VLM}}(f^2_{\text{LM}}(f^1_{\text{VLM}}(\text{image})))$$

Zeng et al. (2023). Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. ICLR.

Socratic Models

- Enable models to “communicate” with each other through their output labels, prompting, and ranking

$$f^3_{\mathrm{VLM}}(f^2_{\mathrm{LM}}(f^1_{\mathrm{VLM}}(\text{image})))$$

detect things

Zeng et al. (2023). Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. ICLR.

Socratic Models

- Enable models to "communicate" with each other through their output labels, prompting, and ranking

$$f^3_{\mathrm{VLM}}(f^2_{\mathrm{LM}}(f^1_{\mathrm{VLM}}(\mathrm{image})))$$

detect things

generate captions

I am an intelligent image captioning bot. This image is a {img_type}. There {num_people}. I think this photo was taken at a {place1}, {place2}, or {place3}. I think there might be a {object1}, {object2}, {object3},... in this {img_type}. A creative short caption I can generate to describe this image is:

Zeng et al. (2023). Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. ICLR.

Socratic Models

- Enable models to “communicate” with each other through their output labels, prompting, and ranking

$$f^3_{\mathrm{VLM}}(f^2_{\mathrm{LM}}(f^1_{\mathrm{VLM}}(\text{image})))$$

detect things

generate captions

re-rank

I am an intelligent image captioning bot. This image is a {img_type}. There {num_people}. I think this photo was taken at a {place1}, {place2}, or {place3}. I think there might be a {object1}, {object2}, {object3},... in this {img_type}. A creative short caption I can generate to describe this image is:

Zeng et al. (2023). Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. ICLR.

Socratic Models

- Enable models to “communicate” with each other through their output labels, prompting, and ranking

$$f^3_{\mathrm{VLM}}(f^2_{\mathrm{LM}}(f^1_{\mathrm{VLM}}(\text{image})))$$

detect things

generate captions

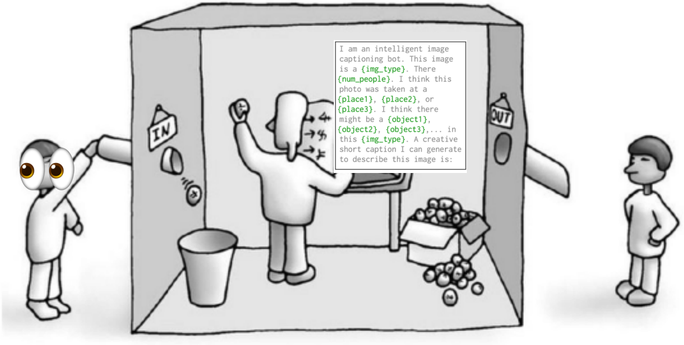
re-rank

I am an intelligent image captioning bot. This image is a {img_type}. There {num_people}. I think this photo was taken at a {place1}, {place2}, or {place3}. I think there might be a {object1}, {object2}, {object3},... in this {img_type}. A creative short caption I can generate to describe this image is:

SM (ours): This image shows an inviting dining space with plenty of natural light.

ClipCap: A wooden table sitting in front of a window.

Zeng et al. (2023). Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. ICLR.



What does it mean to only understand symbols as defined by other symbols?

Multilingual Captioning with Retrieval Augmentation

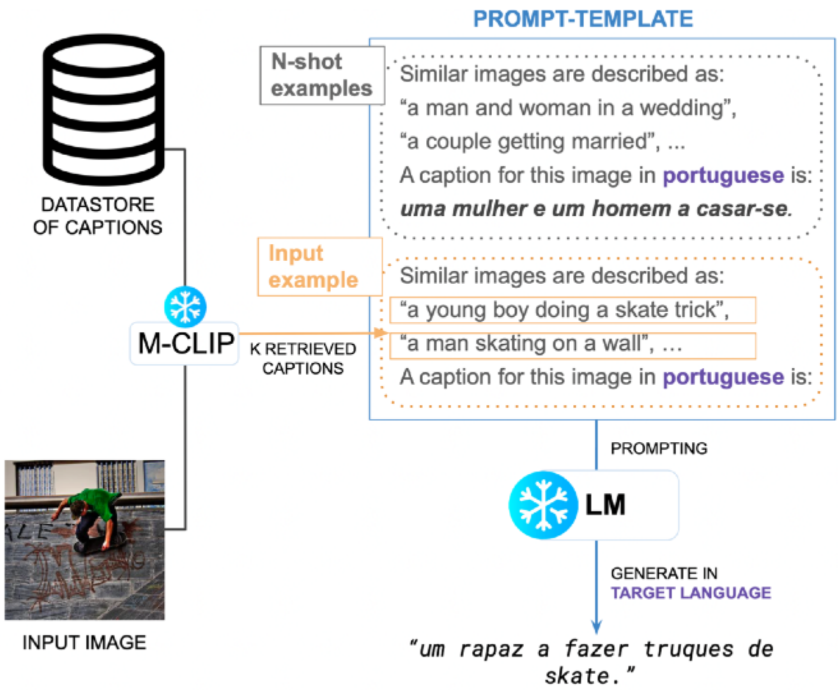
$$f^3_{\mathrm{VLM}}(f^2_{\mathrm{LM}}(f^1_{\mathrm{VLM}}(\text{image})))$$

detect things

generate captions

re-rank

- Prompt with N-shot examples of transforming captions into the target language



Multilingual Captioning with Retrieval Augmentation

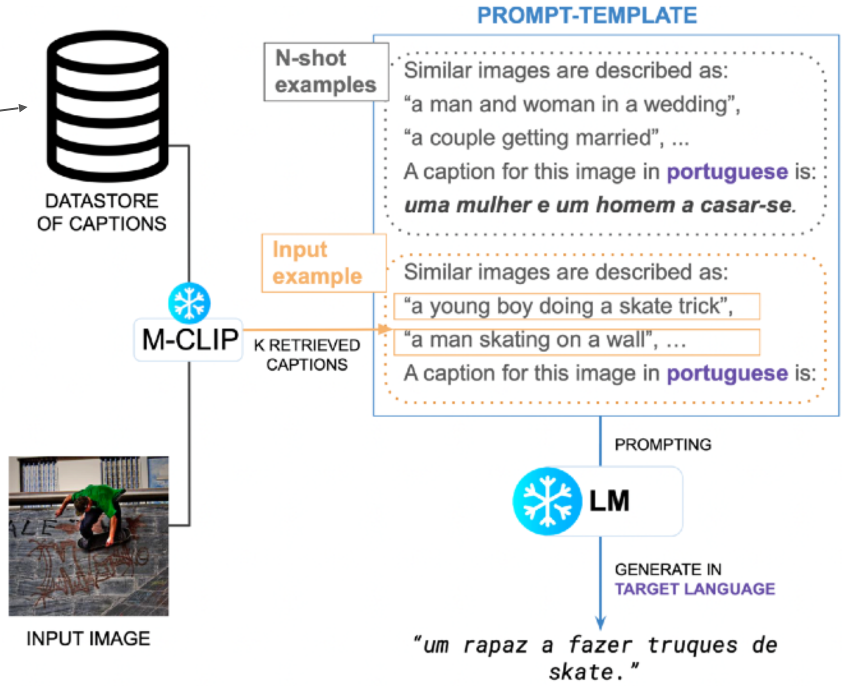
$$f^3_{\mathrm{VLM}}(f^2_{\mathrm{LM}}(f^1_{\mathrm{VLM}}(\mathrm{image})))$$

detect things

generate captions

re-rank

- Prompt with N-shot examples of transforming captions into the target language



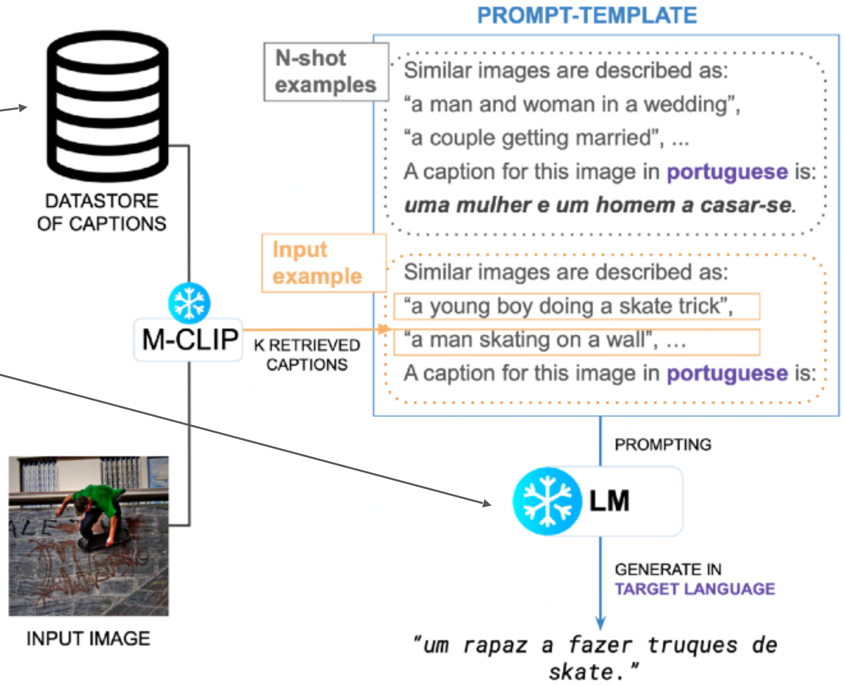
Multilingual Captioning with Retrieval Augmentation

$$f^3_{\text{VLM}}(f^2_{\text{LM}}(f^1_{\text{VLM}}(\text{image})))$$

detect things

generate captions

re-rank



- Prompt with N-shot examples of transforming captions into the target language

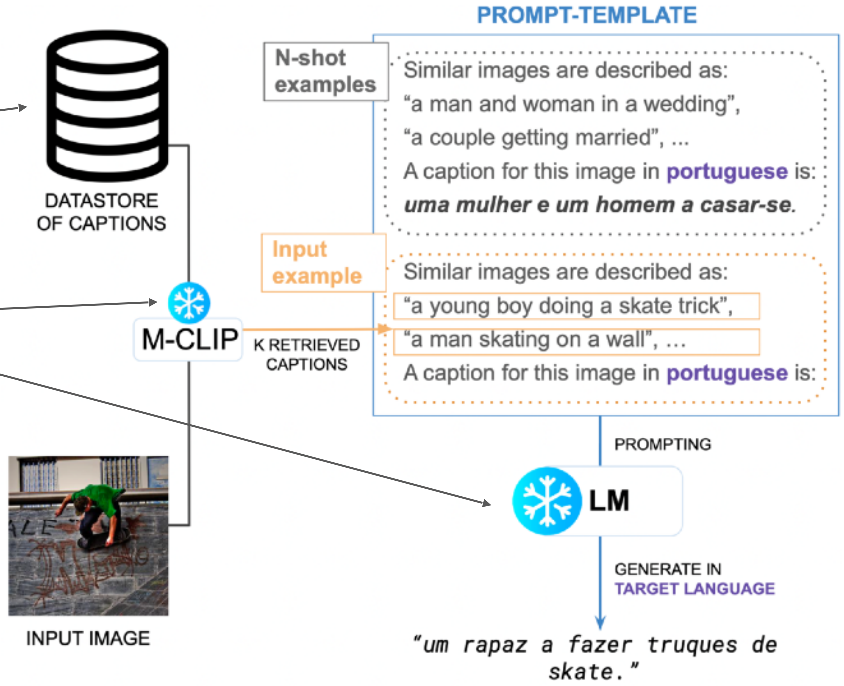
Multilingual Captioning with Retrieval Augmentation

$$f^{3}_{\mathrm{VLM}}(f^{2}_{\mathrm{LM}}(f^{1}_{\mathrm{VLM}}(\text{image})))$$

detect things

generate captions

re-rank



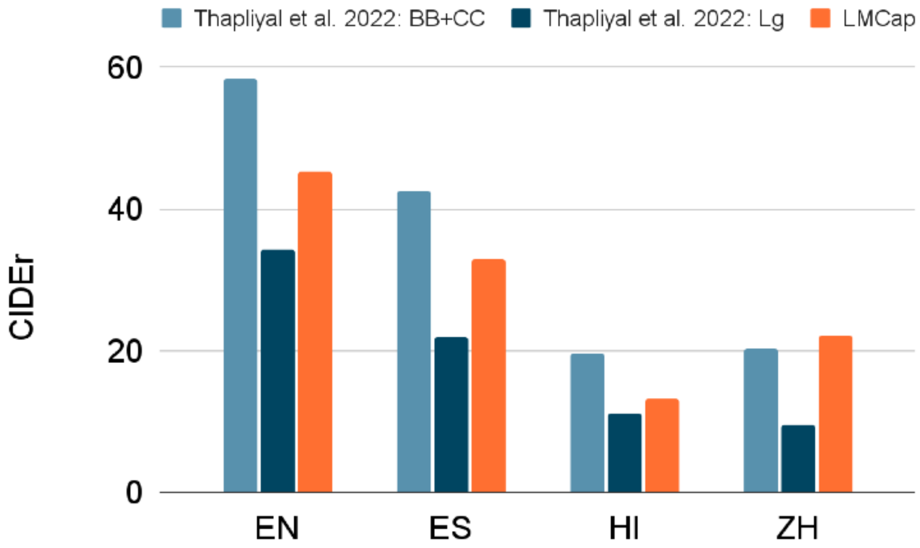
- Prompt with N-shot examples of transforming captions into the target language

Experimental Setup

- XGLM Language Model 564M - 7.6B params
- Multilingual CLIP (LAION)
- Experiments on XM3600 (Thapliyal et al. 2022)
    - 100 images in 36 languages
- **No training or fine-tuning on any captioning data.**

Thapliyal et al. (2022). Crossmodal-3600: A massively multilingual multimodal evaluation dataset. EMNLP.

Results



Params	RAM	en	es	hi	zh
564M	6G	0.411	0.094	0.030	0.146
1.7B	12G	0.637	0.143	0.066	0.272
2.9B	16G	0.767	0.454	0.334	0.584
7.5B	22G	**0.787**	**0.489**	**0.365**	**0.644**

Competitive performance compared to supervised models

Need at least 2.9B parameter decoder for multilingual generation

Qualitative Example

Retrieved Examples

two people and a kid skiing along a trail

an adult and two children are cross country skiing

two men and a little boy are skiing on a snowy spot

two adults on skis with a child on skis between them

More examples in the paper

Qualitative Example

Retrieved Examples

two people and a kid skiing along a trail

an adult and two children are cross country skiing

two men and a little boy are skiing on a snowy spot

two adults on skis with a child on skis between them

Generated Captions

ENG: two people and a kid skiing along a trail

ESP: dos hombres y un niño esquiando en una pista de nieve

ZHO: 两个大人和一个小男孩在雪地上滑雪

More examples in the paper

Q: How does all of this work?

Understanding Retrieval Robustness for Retrieval-augmented Image Captioning

ACL 2024

W. Li

J. Li

R. Ramos

R. Tang

D. Elliott

Revisiting Swanson Soup

- In Ramos et al. CVPR 2023, we observed the power of in-context learning and retrieval-augmentation
- But what is happening here?
- How is the model using the retrieved captions?

VizWiz

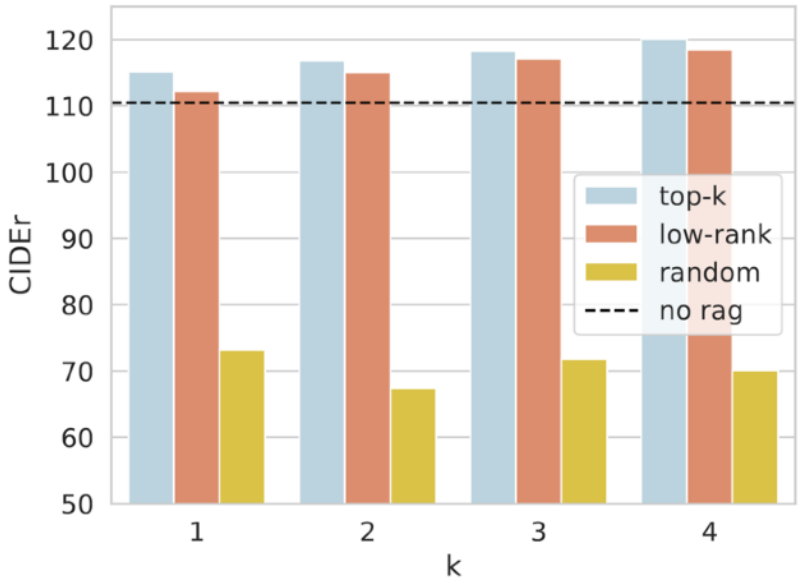
- a can of swanson fat free chicken broth
- a can of swanson brand chicken broth with less sodium
- a 14,5 ounce can of swanson branded chicken broth
- a can of swanson chicken broth on a table

Generated caption:

a can of swanson brand chicken broth on a table

Measuring Robustness

- Is SmallCap sensitive to the *quality* of the retrieved captions?
  - Top-ranked items
  - Random items
  - Lower-ranked items



Question: If the model is so affected by random captions, then is it more like a paraphrasing model that ignores the visual content?

Majority Token Analysis

- Given a list of K retrieved captions, we can create an ordered list of the frequency that each unique token appears in the captions:

$$C := \{C_{T_1}, C_{T_2}, \ldots, C_{T_U}\}$$

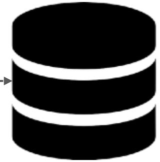
$$C_{T_i} = \sum_{k=1}^{K} \mathbf{1}_{T_i \in R_n}$$

- **Majority Token**: If token T_i appears at least $K/2$ times, then we define it as majority token in the retrieved captions:

$$M_T := \{C_{T_i}, C_{T_j}, \ldots\} \quad \text{s.t.} \;\; C_{T_u} \geq K/2 \; \forall \; U$$

Majority Tokens Example

- **Majority Token**: $$M_T := \{C_{T_i}, C_{T_j}, \ldots\} \quad \text{s.t.} \;\; C_{T_u} \geq K/2 \; \forall \; U$$



R_1: Three people skiing through a forest

R_2: An older woman in a wheelchair holding a white teddy

R_3: A man and a woman sit holding a teddy bear

Majority Tokens: “teddy”, “bear”, “woman”

Known Good / Known Bad Captions

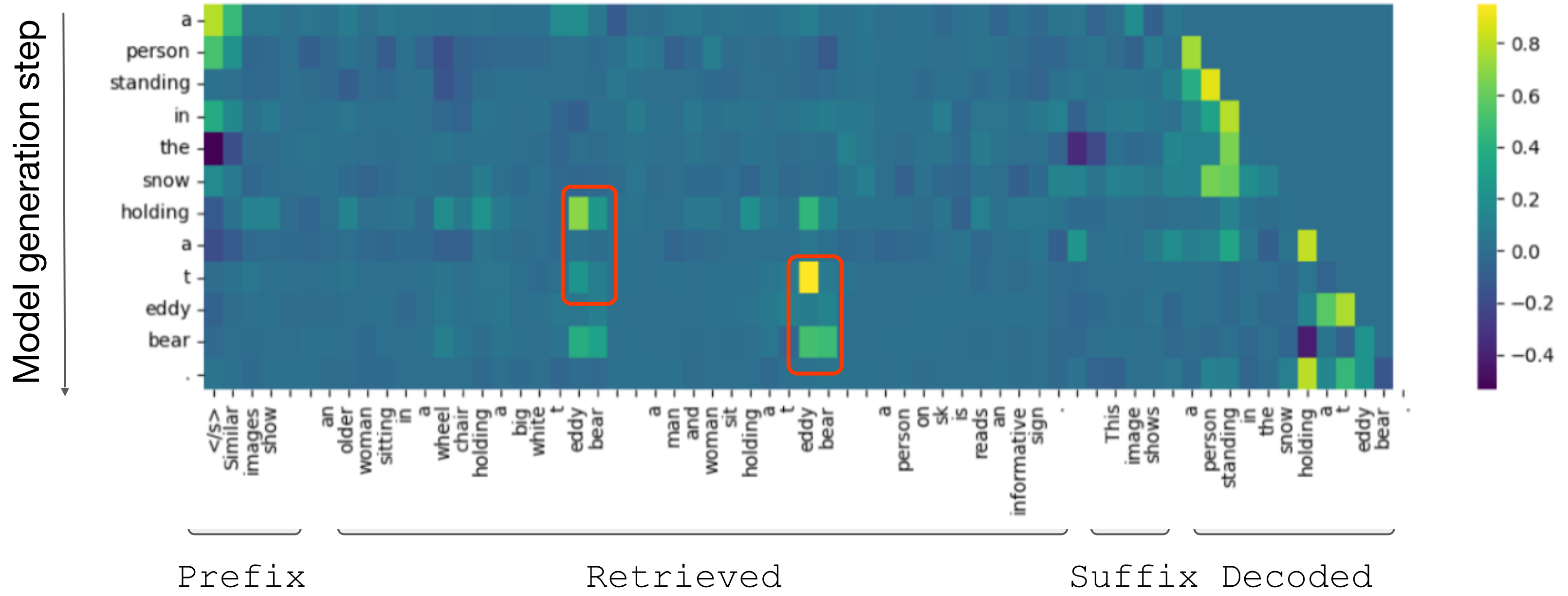
- With Majority Tokens, we can force an experimental setup with known *good* or known *bad* retrieved captions
- Force an asymmetry:
  - 2 Good captions ~ 1 Bad caption ⇒ *useful* majority tokens?
  - 1 Good caption ~ 2 Bad captions ⇒ *harmful* majority tokens?
- **Good**: high-ranked caption **Bad**: random caption

Known Good / Known Bad Captions

- With Majority Tokens, we can force an experimental setup with known *good* or known *bad* retrieved captions
- Force an asymmetry:
  - 2 Good captions ~ 1 Bad caption ⇒ *useful* majority tokens?
  - 1 Good caption ~ 2 Bad captions ⇒ *harmful* majority tokens?
- **Good**: high-ranked caption **Bad**: random caption
- Results
  - 2 Good ~ 1 Bad: 86% of generated captions contain a majority token
  - 1 Good ~ 2 Bad: 21%

Integrated Gradients Input Attribution

- Which input tokens are most/least important in the model output?

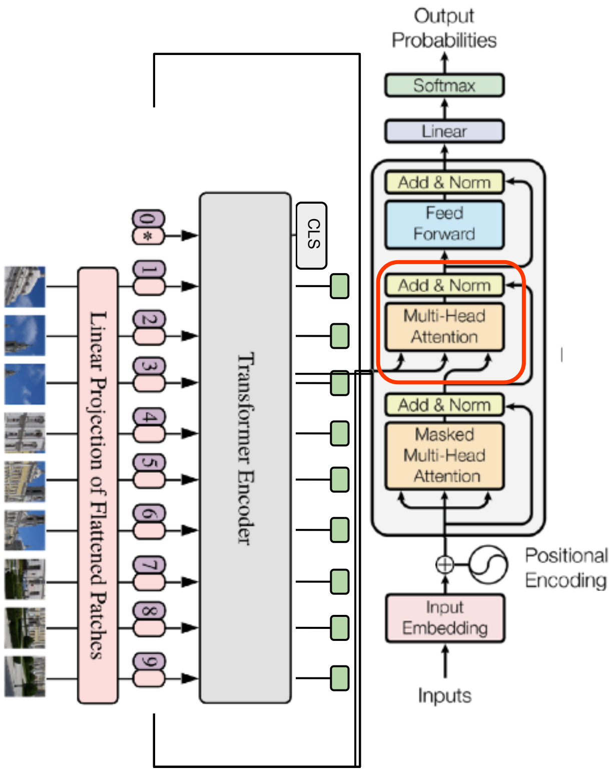


Sundararajan et al. ICML 2017. Axiomatic Attribution for Deep Networks.

Self- and Cross-Attention Analysis

- What can we learn about SmallCap by inspecting what it attends to in the textual and the visual inputs?
- Track the location of the **maximally**-attended inputs

$$\mathbb{1}[I_n(i,j)] = \begin{cases} 1 & \text{if } \arg\max_z Att(j,z)_i \in S_n \\ 0 & \text{otherwise} \end{cases}$$



Self-Attention Analysis

<S> ← BOS

Similar images show ← Prefix

a man working some levers at a train yard
a train engineer preparing the engine of a train
a train being worked on in a train manufacturer
a man wearing a safety vest standing by a train.
← Retrieved

This image shows ← Suffix

a person working … ← Generated

Self-Attention Analysis

<S>

BOS

Similar images show

Prefix

a man working some levers at a train yard
a train engineer preparing the engine of a train
a train being worked on in a train manufacturer
a man wearing a safety vest standing by a train.

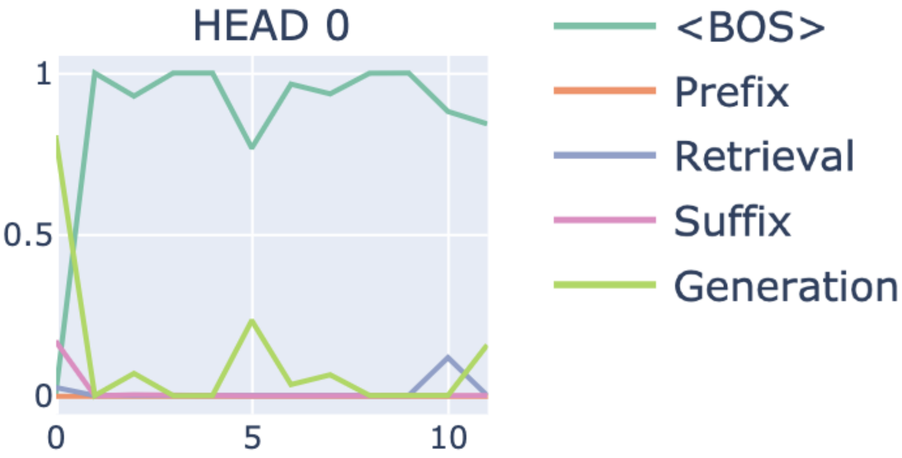
Retrieved

This image shows

Suffix

a person working …

Generated



Self-Attention Analysis

<S> — BOS

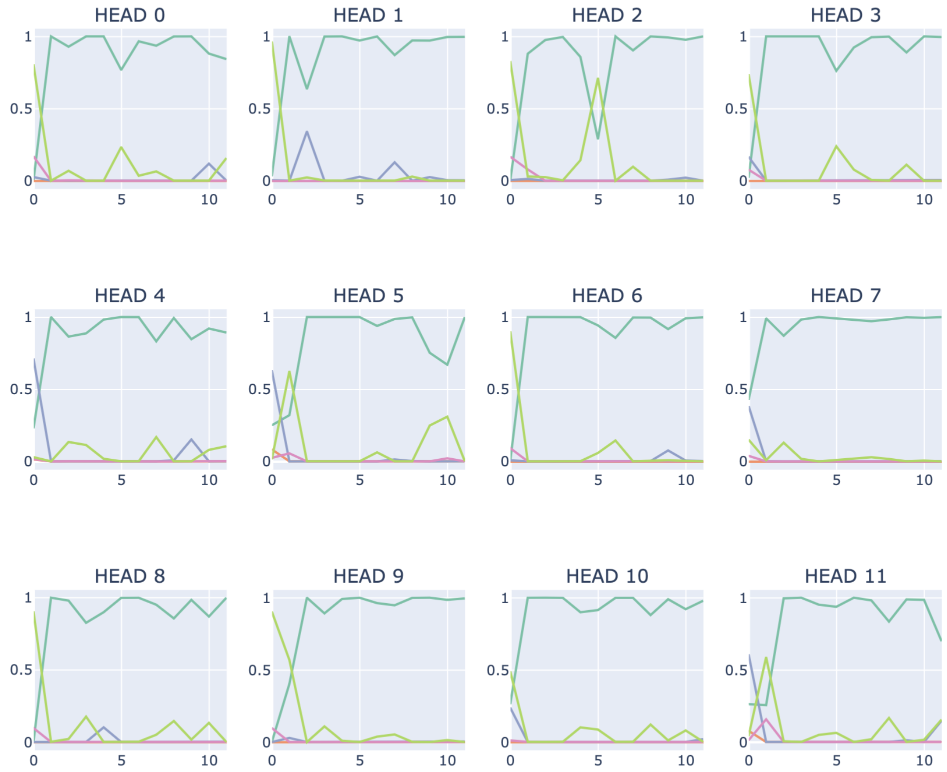
Similar images show — Prefix

a man working some levers at a train yard
a train engineer preparing the engine of a train
a train being worked on in a train manufacturer
a man wearing a safety vest standing by a train. — Retrieved

This image shows — Suffix

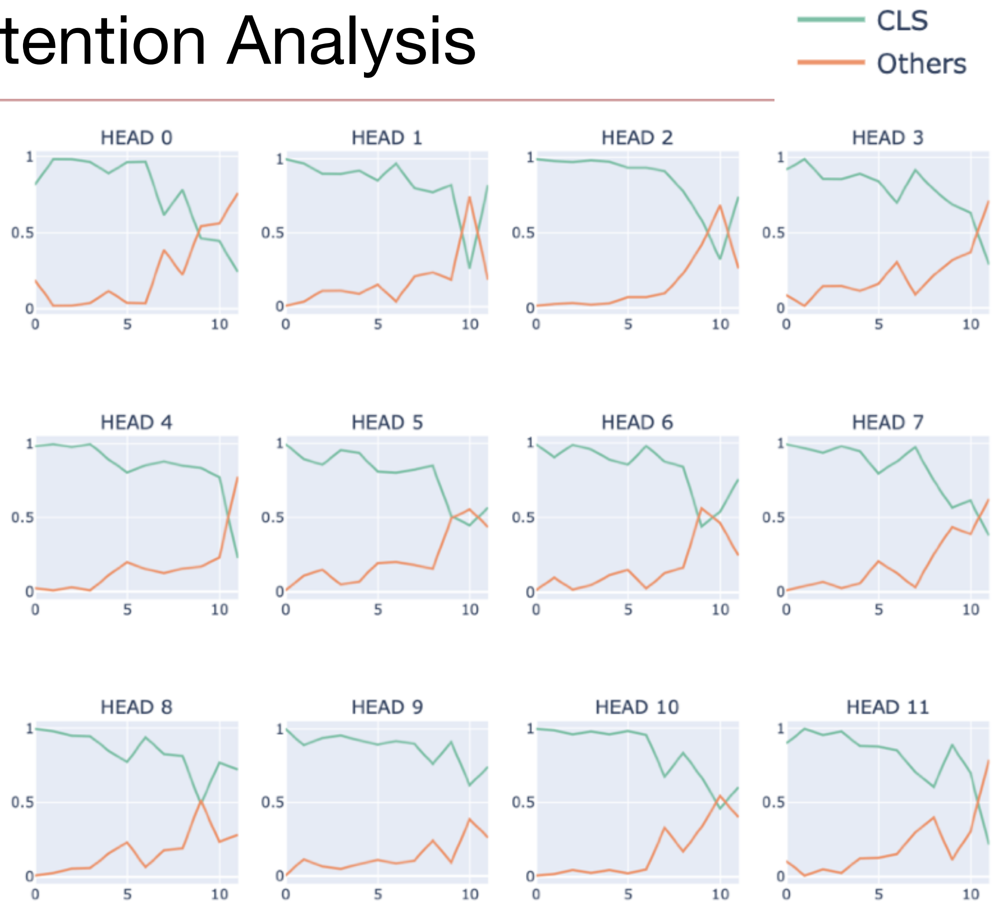
a person working … — Generated

Self-attention is maximally attending to the BOS token



Cross-Attention Analysis

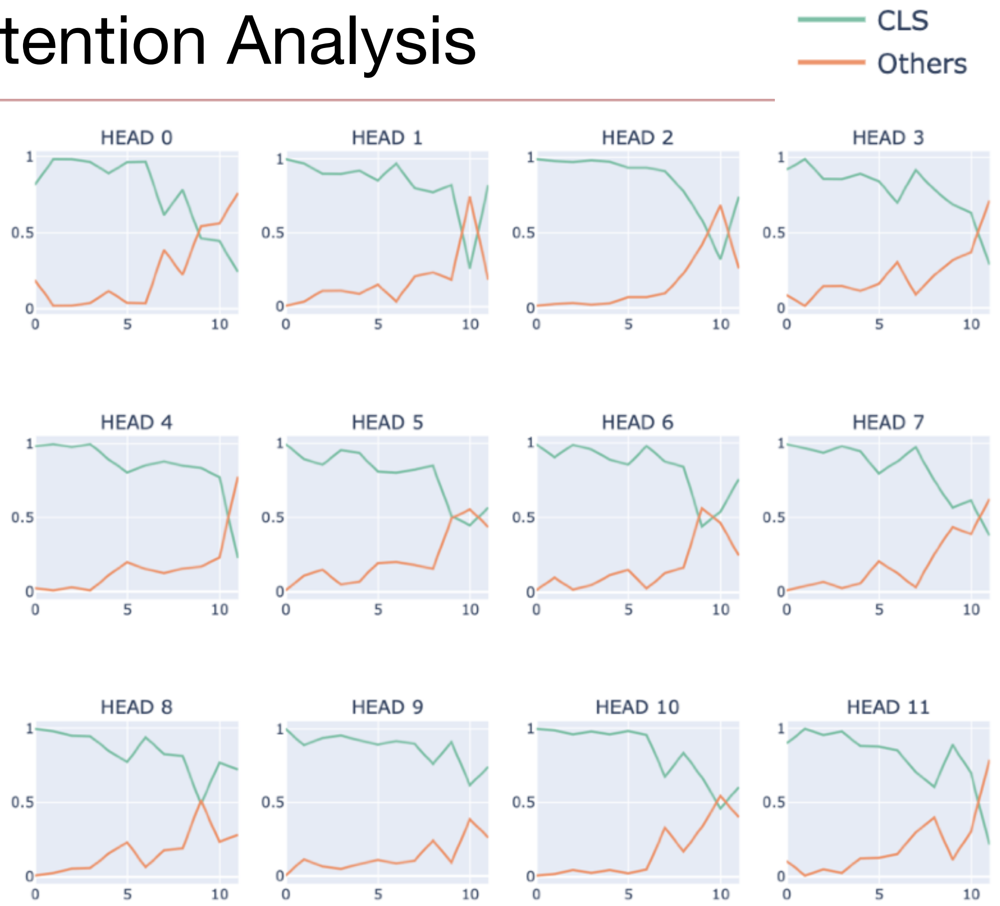
The cross-attention layers focus on the “summary” **image CLS** embedding



Cross-Attention Analysis

The cross-attention layers focus on the “summary” **image CLS** embedding

Cross-attention to **image patches** only emerges at the final layers of the LM



Improving Robustness

- Given that the model appears to be strongly guided by retrieved captions, can we train the model to be less reliant on this?
  - Yes! We can create less-perfect retrieval lists during training

- **Sample-K**: randomly choose k/N retrieved captions
- **C-Sample-K**: only use the most relevant caption, and k-1 randomly sampled captions in the prompt

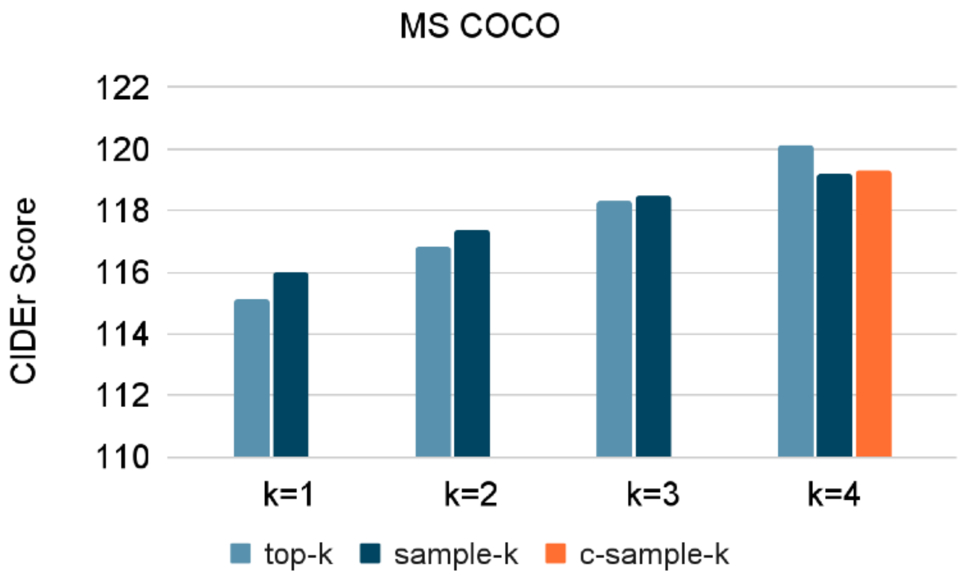
Hoang et al. 2022. Improving robustness of retrieval augmented translation via shuffling of suggestions.

Experimental Protocol

- Encoder: CLIP ViT-B/32
- Decoder: OPT-125M
- Training data: MS COCO
- Evaluation with CIDEr
    - MS COCO
    - VizWiz
    - NoCaps

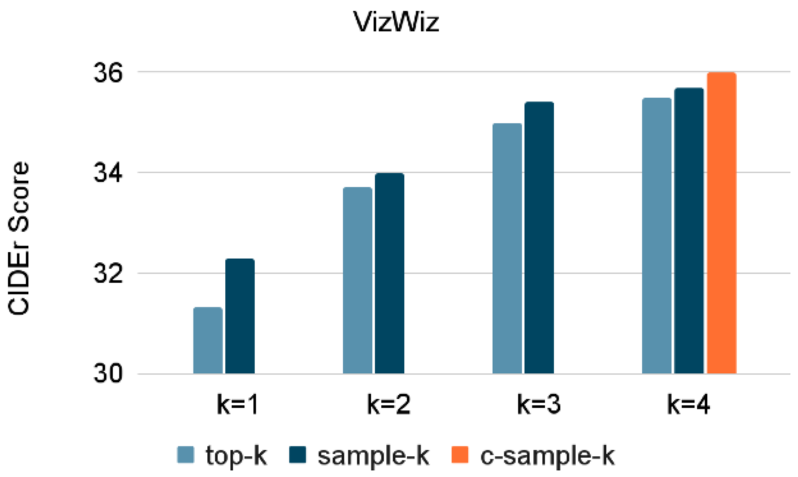
Example evaluation images from the NoCaps dataset

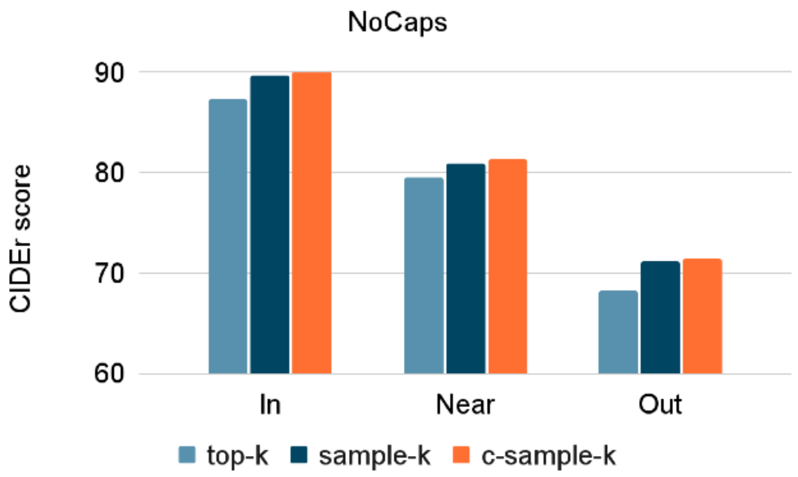
In-Domain Results



Improved performance with smaller retrieval sets

Out-of-Domain Results





Improvements in two out-of-domain datasets

Qualitative Examples

- a man posing with a surfboard on an elevator
- a woman sitting on a bench next to a man in a hat
- a greyhound dog lying on an unmade bed
- a pink teddy bear and a brown teddy bear sitting on wooden rods

Sample-k a person riding a **horse** on top of a beach

Top-k a person sitting on a **bench** on a beach

- a train with the numbers 60016 is heading down the tracks
- a black and white photo of two people holding hands in a city on a rainy day
- this youngster has a boogie board to ride the smaller waves
- a wooden entertainment center containing a television set

Sample-k a close up of a **fire hydrant** on a sidewalk

Top-k a close up of a **person** on a sidewalk

Wrap-up

Open Questions

- How many of these observations apply to visual prefix models?
  - I think we will still observe the problems associated with majority tokens
- What is the best way to construct N-shot examples for mRAG?
  - Demonstrate the diversity of the tasks / target languages / visual inputs
- When will we have *usable* multimodal ICL for multimodal RAG?
  - We have been trying to make progress on this with ImageChain

IMAGECHAIN: Advancing Sequential Image-to-Text Reasoning in Multimodal Large Language Models

Danae Sánchez Villegas* Ingo Ziegler* Desmond Elliott

Final Conclusions

- Retrieval-augmentation is a powerful approach to building lightweight image captioning models that can easily adapt to new domains
    - Improve lightweight trained models
    - Improve zero-training models
    - Enable zero-shot multilingual transfer
- Open questions about how to make RAG-based models more robust and reliable in practice

References

- R. Ramos, B. Martins, and D. Elliott. **Few-shot Multilingual Image Captioning by Retrieval Augmented Language Model Prompting**. ACL 2023.
- R. Ramos, B. Martins, D. Elliott, and Y. Kementchedjhieva. **SmallCap: Lightweight Image Captioning Prompted with Retrieval Augmentation**. CVPR 2023.
- R. Ramos, E. Bugliarello, B. Martins, and D. Elliott. **PAELLA: Parameter-Efficient Lightweight Language-Agnostic Captioning Model**. NAACL 2024.
- W. Li, J. Li, R. Ramos, R. Tang, and D. Elliott. **Understanding Retrieval Robustness for Retrieval-augmented Image Captioning**. ACL 2024.
- D. S. Villegas, I. Ziegler, and D. Elliott. **ImageChain: Advancing Sequential Image-to-Text Reasoning in Multimodal Large Language Models**.