



SMALLCAP: Lightweight Image Captioning Prompted with Retrieval Augmentation

Rita Ramos, Bruno Martins, Desmond Elliott, Yova Kementchedjieva

INESC-ID, Instituto Superior Tecnico, University of Lisbon Department of Computer Science,
University of Copenhagen, Pioneer Center for AI

2023/CVPR

2024.04.15

이상민

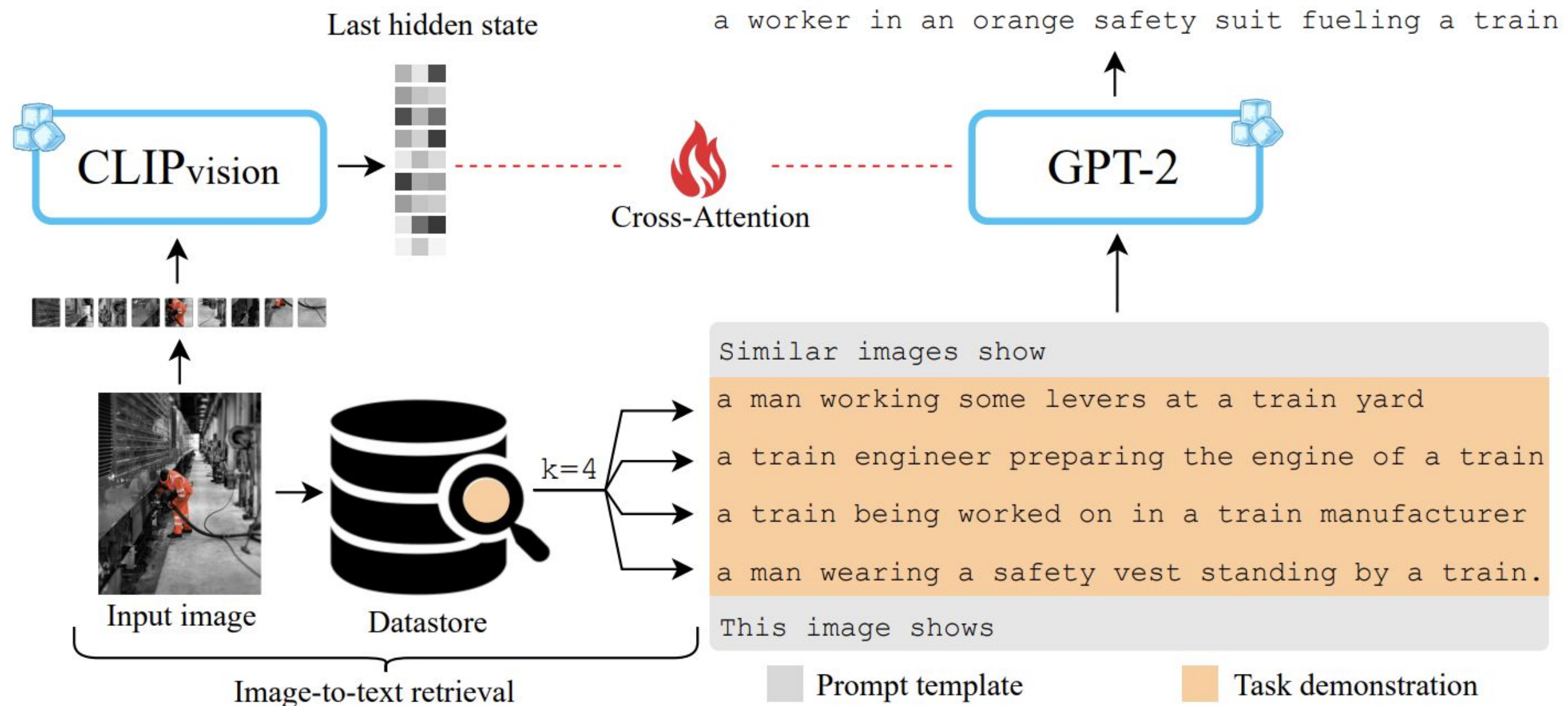
1. Introduction

- 최근 이미지 캡셔닝 분야는 데이터와 모델 사이즈 증가에 집중하고 있다
- 모델 사이즈가 커지면서 모델을 Pre-train, fine-tune할 때 학습이 오래 걸리고 실용적으로 모델을 사용하기 어렵다
- large model의 대안으로 경량 파라미터의 모델들이 연구 되었지만, 유의미한 성능 향상은 얻지 못했다.
- 본 연구에서는 대안으로 SMALLCAP, image captioning model을 새롭게 제안

2. Proposed Approach

- **SMALLCAP**

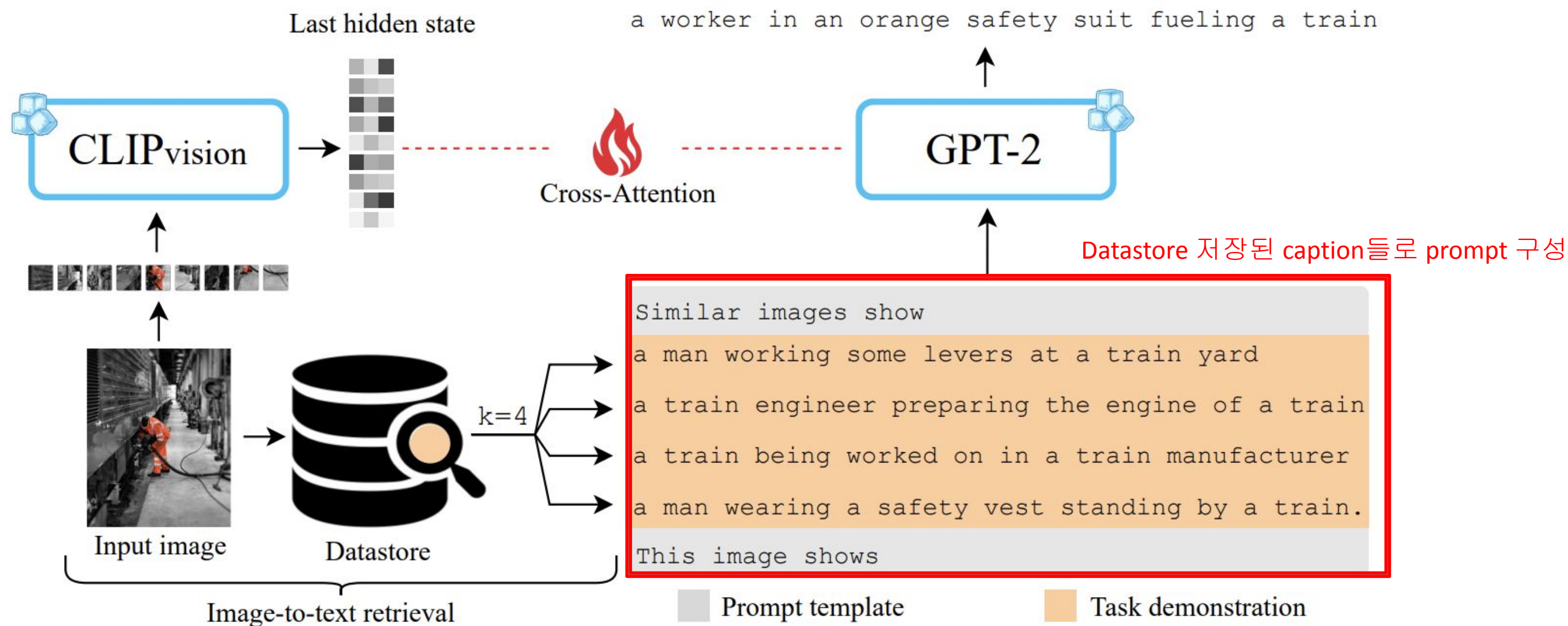
- 사전학습 된 CLIP vision encoder, GPT-2 모델을 freezing해서, 인코더 디코더로 사용
- 입력 이미지와 유사한 caption으로 구성된 prompt를 디코더의 입력으로 사용



(a) Full Model Architecture

2. Proposed Approach

- Prompting with Retrieved Captions

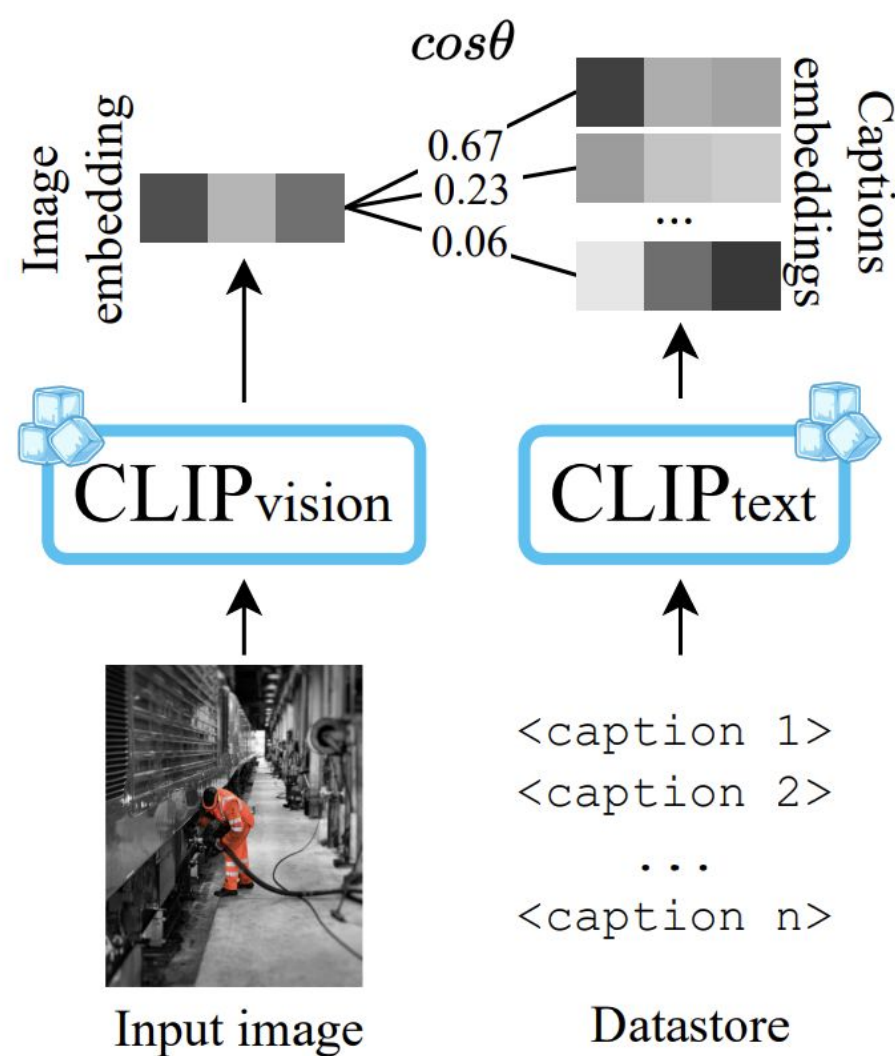


(a) Full Model Architecture

2. Proposed Approach

- **Prompting with Retrieved Captions**

- 입력 이미지와 데이터스토어의 내용을 인코딩
- 이후 코사인 유사도를 기반으로 이미지와 가장 유사한 k개의 텍스트를 datastore에서 검색
- k개의 텍스트는 prompt template에 삽입



Similar images show
{caption1}
...
{captionk}.
This image shows .

[Prompt template]

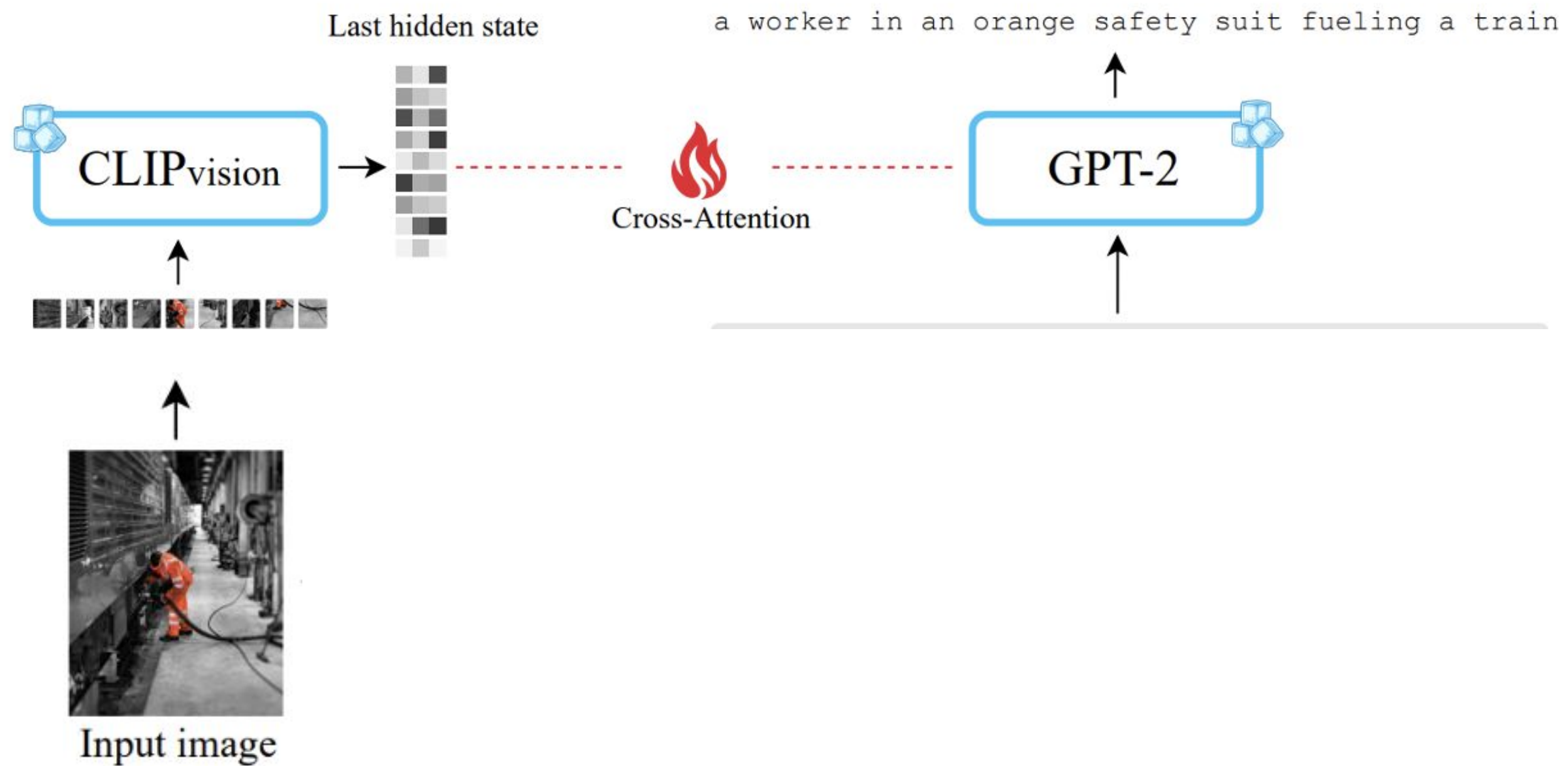


```
Similar images show
a man working some levers at a train yard
a train engineer preparing the engine of a train
a train being worked on in a train manufacturer
a man wearing a safety vest standing by a train.
This image shows
```


2. Proposed Approach

- Image Encoding

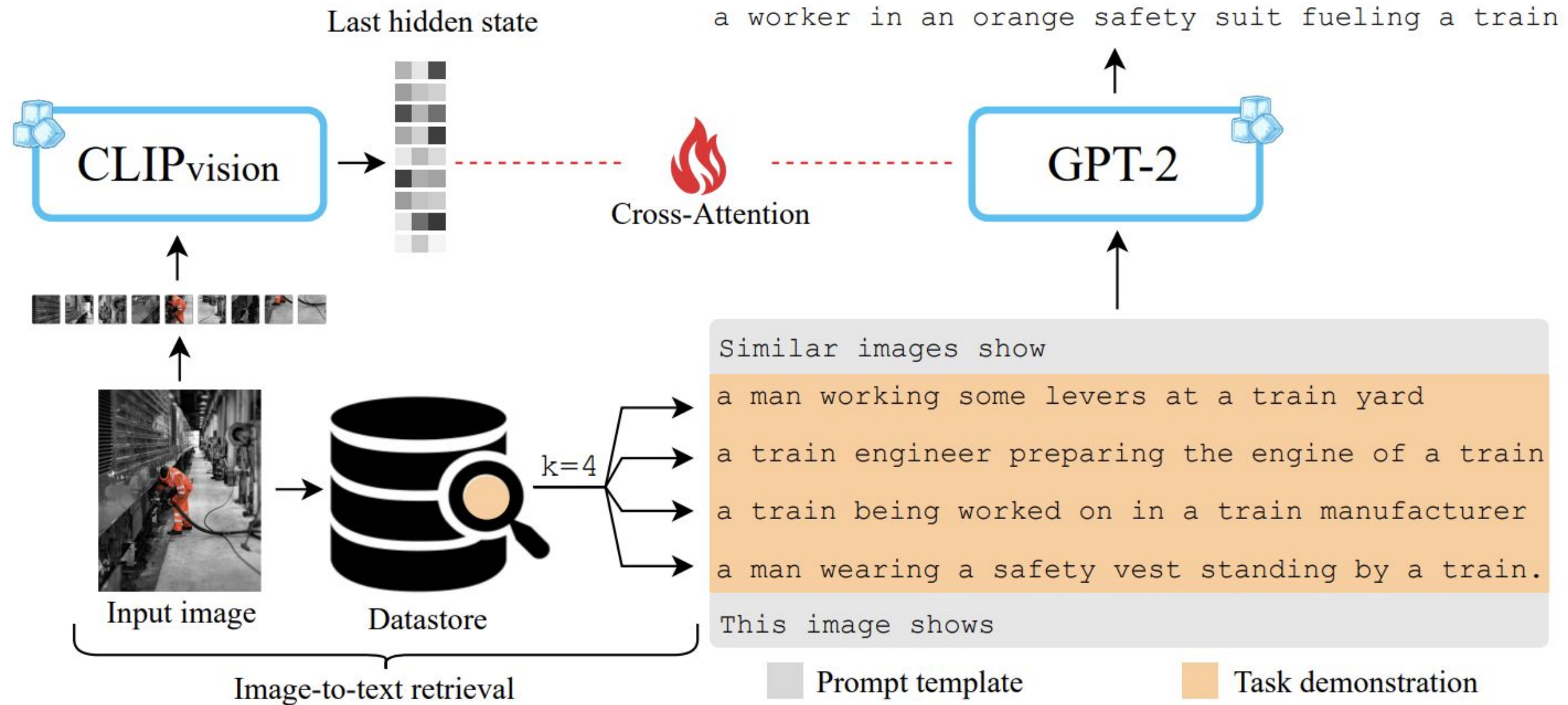
- CLIP vision encoder를 사용해 이미지의 patch들을 각각 embedding의 형태로 변환



2. Proposed Approach

• SMALLCAP

- 최종적으로 디코더는 image features와 prompt를 통해 caption을 생성한다.
- 인코더와 디코더를 freezing했기 때문에 실질적으로 cross-attention layer만 학습



(a) Full Model Architecture

2. Proposed Approach

- 학습 방법

- 전체 모델 구조에서 cross-attention layer만 cross-entropy loss를 최소화 하는 방향으로 학습

$$L_{\theta} = - \sum_{i=1}^M \log P_{\theta}(y_i | y_{<i}, \mathbf{X}, \mathbf{V}; \theta).$$

\mathbf{V} : image feature
 \mathbf{X} : prompt
 θ : cross-attention layer parameters

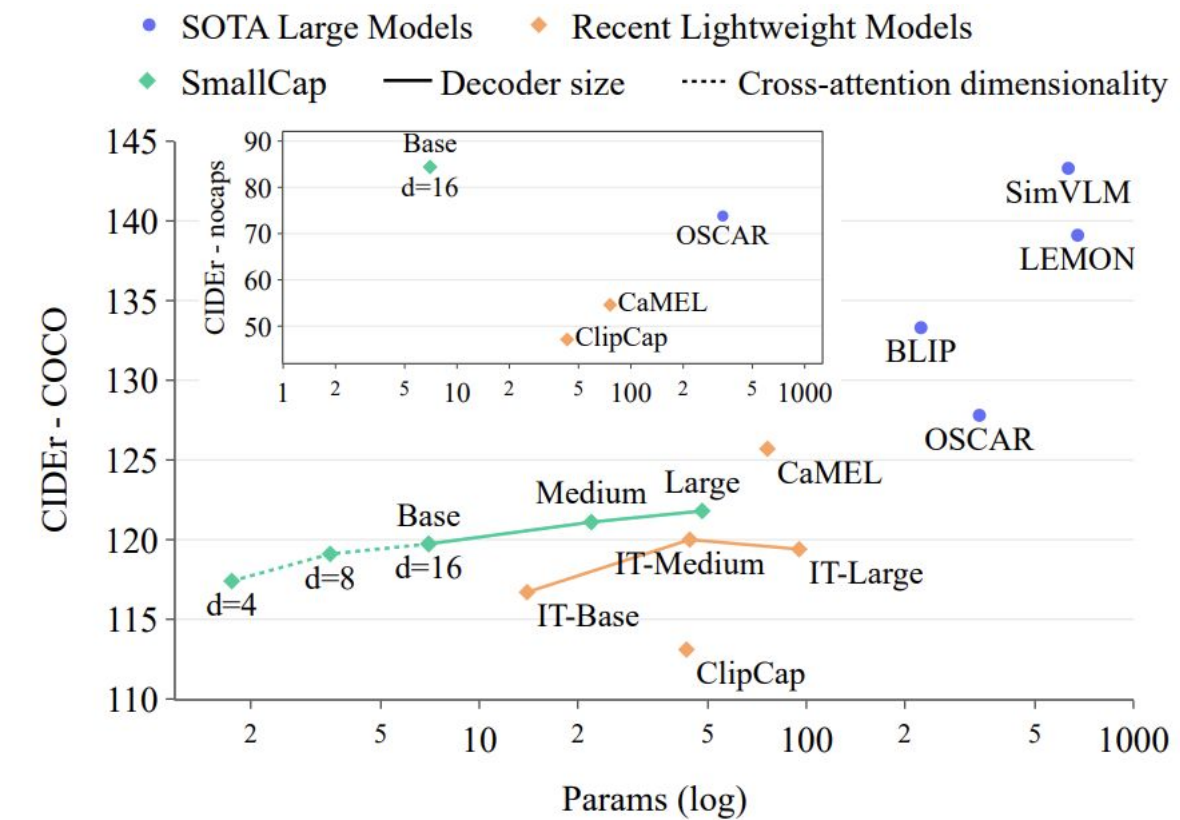
3. Experiments

• Benchmark Results

COCO test 데이터 셋에 대한 성능 결과, 아래 모든 모델은 COCO dataset을 학습 했다.

Model	$ \theta $	B@4	M	CIDEr	S
Large Models with V&L pre-training					
LEMON _{Huge} [11]	675	41.5	30.8	139.1	24.1
SimVLM _{Huge} [42]	632	40.6	33.7	143.3	25.4
OSCAR _{Large} [19]	338	37.4	30.7	127.8	23.5
BLIP _{CapFilt-L} [18]	224	39.7	-	133.3	-
Lightweight-training models					
I-Tuning _{Large} [22]	95	34.8	29.3	119.4	22.4
CaMEL [5]	76	39.1	29.4	125.7	22.2
I-Tuning _{Medium} [22]	44	35.5	28.8	120.0	22.0
ClipCap [25]	43	33.5	27.5	113.1	21.1
I-Tuning _{Base} [22]	14	34.8	28.3	116.7	21.8
SMALLCAP	7	37.0	27.9	119.7	21.3
SMALLCAP _{d=16, Large}	47	37.2	28.3	121.8	21.5
SMALLCAP _{d=16, Med}	22	36.5	28.1	120.7	21.6
SMALLCAP _{d=8, Base}	3.6	36.7	27.8	119.1	21.1
SMALLCAP _{d=4, Base}	1.8	36.0	27.4	117.4	21.0

Table 1. Results on the COCO test set with cross-entropy training.
 $|\theta|$: number of trainable parameters in the model (in millions).



3. Experiments

• Benchmark Results

- Nocaps 데이터의 IN-domain, Near-domain, out-of-domain, entire-domain에서의 CIDEr 점수
- In-domain을 제외한 모든 도메인에서 성능이 가장 높다.

Model	In	Near	Out	Entire
OSCAR _{Large} [◇]	84.8	82.1	73.8	80.9
CaMEL [*]	88.1	79.1	54.6	75.9
ClipCap [*]	74.5	65.6	47.1	63.4
SMALLCAP	83.3	77.1	65.0	75.8
SMALLCAP _{+W+H}	87.9	84.6	84.4	85.0

Table 2. CIDEr results on the nocaps test set. ◇: Results copied from the respective publications. *: Results computed by us. +W+H: datastore with additional Web and Human-labeled data.

SMALLCAP + W+H: datastore에 Web, Human-labeled data를 추가한 모델

3. Experiments

• Qualitative Examples

- Retrieving된 caption 들은 입력 이미지와 유사하고 생성된 caption과도 의미론적으로 비슷하다
- 두번째 이미지를 보면 Closeup of a person 과 같이 이미지와 유사하지 않는 정보에도 robust하게 예측한다

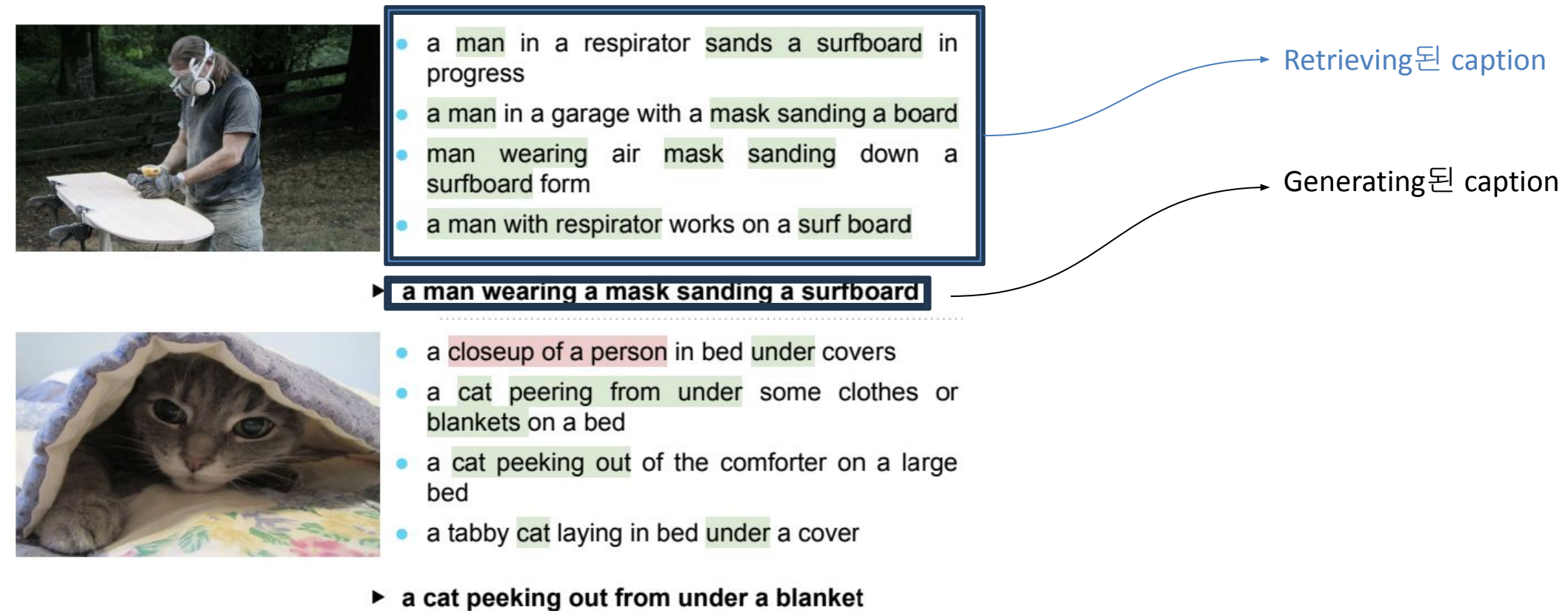


Figure 3. Examples generated by SMALLCAP, together with the retrieved predictions from the COCO datastore. • denotes the retrieved captions, highlighted as green or red to indicate correct and mismatch captions, respectively. ► denotes the generated caption.

3. Experiments

• Qualitative Examples

- coco 데이터로 학습된 모델을 추론 시점에 data store의 구성을 coco에서 새로운 도메인으로 변경 했을 때 새로운 도메인에 적응하는 것을 보여준다
- 입력이미지와 비슷한 도메인의 데이터로 구성된 datastore를 사용하면 처음 보는 이미지에 대해서도 강건한 예측이 가능하다.




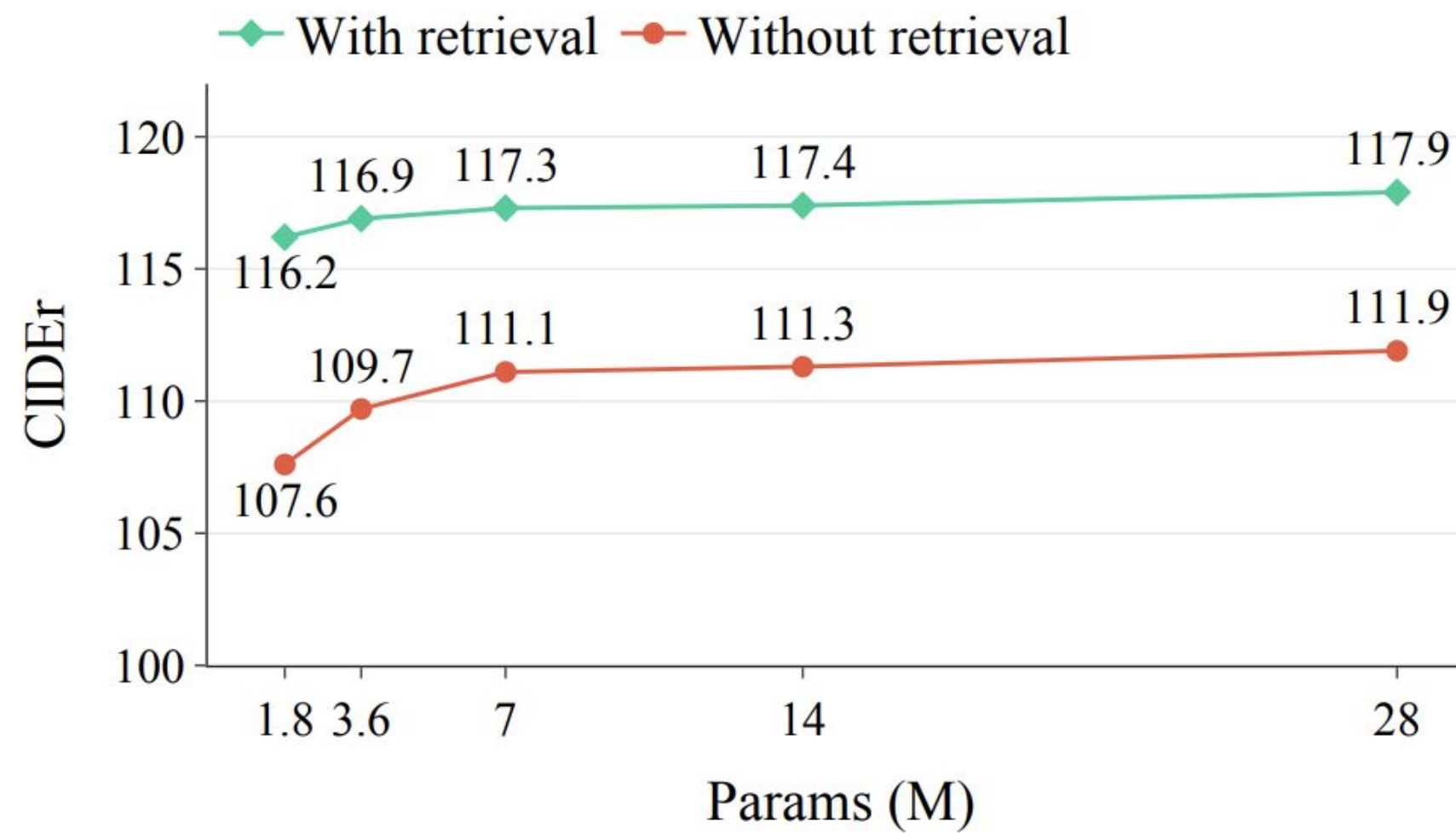
			
	Flick30k	VizWiz	MSR-VTT
기존 datastore 에서 Retrieving된 caption	<ul style="list-style-type: none"> a little girl holding forth a pink ballerina teddy bear a girl is dancing in a pink skirt a young girl is dancing while holding a umbrella future dancers might use their umbrellas in a routine 	<ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> playing on a small laptop and a phone at the same time is not recommended a blue, red, and yellow training at a train station an image of a split screen of variety of images people play demos of the newest nintendo games
모델이 생성한 caption	<ul style="list-style-type: none"> a little girl holding an umbrella in a room 	<ul style="list-style-type: none"> a close up of a plate of food on a table 	<ul style="list-style-type: none"> a bunch of different images of a train station
입력이미지와 같은 도메인에서 Retrieving된 caption	<ul style="list-style-type: none"> a little girl is dressed in a pink ballerina costume a little girl in pink dances with her hands on her hips a little girl in a pink tutu gets ready for ballet dancing, a boy in a spider-man shirt behind a young girl wearing a pink tutu 	<ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> players explore the pokemon universe on a split screen screen cast of an original pokemon game a man screencasts himself playing the original pokemon series pokemon engaging battle in video game
모델이 생성한 caption	<ul style="list-style-type: none"> a little girl in a pink tutu is dancing 	<ul style="list-style-type: none"> a can of swanson brand chicken broth on a table 	<ul style="list-style-type: none"> a bunch of pictures of different pokemon

Figure 5. Examples of captions generated for Flickr30k, VizWiz and MSR-VTT, with retrieval either from COCO or in-domain data. The captions use words retrieved from the in-domain datastores which were rarely seen in the COCO training data (tutu, swanson, pokemon).

3. Experiments

• The Impact of Retrieval

- Retrieval 적용 유무에 따른 SMALLCAP 모델 성능 평가 결과.
- Retrieval을 적용하지 않고 prompt로 “this image show” 만 사용한 결과 성능이 현저히 떨어졌다.



4. Conclusion

- 학습이 빠르고 재학습 없이 다양한 도메인에 적용 될 수 있는 retrieval augmented image caption 모델인 SMALLCAP을 제안
- 학습 파라미터가 다른 모델들에 비해 훨씬 적음에도 높은 성능을 보임

5. Open questions

- datastore를 news dataset으로 변경한다면 SMALLCAP 모델은 Entity가 포함된 news dataset에 대해서도 robust하게 예측할까?