Image Retrieval from Contextual Descriptions

Benno Krojer[1], Vaibhav Adlakha[1], Vibhav Vineet[2], Yash Goyal[3], Edoardo Ponti[1], Siva Reddy[1]
Mila/McGill University[1] Microsoft Research[2] Samsung SAIT AI Lab[3]

The *ImageCoDe* Challenge:

Retrieve image from 10 minimally contrastive images based on a contextual description

Description:
"*No bridesmaid visible at all*"

Visual **context** and **attention to nuances** are necessary to identify the target image* (green), by cross-referencing the portions of images with bridesmaids (red boxes)

Description:
"*There is no hand visible and almost all of the cardboard box in the bottom right is visible.*"

Pragmatic reasoning is necessary to identify the target image* (green), since the description is semantically false for the target (hands are visible)

Motivation

Create task that shows shortcomings of vision-and-language models
- pragmatics (implicatures, ambiguity)
- requiring "System 2 reasoning"
- temporality
- long-form & complex syntax
- nuances (minimally contrastive)

Step 1: collect minimally contrastive images from 3 video datasets and Open Images via heuristics and visual distance

Step 2 & 3: crowdsource **descriptions** and verify with addition 1-3 human **retrievals**

The Dataset

94K images (9.4K image sets)

21K descriptions

80% video-based

ImageCoDe Statistics and Phenomena

Phenomenon	all %	videos %	static %	Example from ImageCoDe	Definition
Context	47.3	**57.3**	6.6	Figure 2	Visual context or pragmatic inference required.
Temporal	15.0	**18.5**	4.1	*A smiling boy just **begins to** look towards the dog.*	Temporal markers (e.g., *after*) and verbs (e.g., *starts*)
Quantities	48.5	47.7	**51.0**	*There is an **equal amount** of yellow and white between **both** hands.*	—
Spatial Relations	70.5	**72.2**	65.3	*The cloud on **top left side** of box only has **half** of it showing.*	—
Negation	17.9	**20.7**	6.1	*The spoon is at the top right corner, it is **not** moving any of the food.*	—
Visibility / Occlusion	45.5	**54.5**	8.6	*The flowers the woman in the teal strapless dress is carrying are **completely obscured** by the man in the black shirt's head.*	An entity is covered or partially outside of the image.
Nuances	26.3	**31.6**	5.1	*There is the **slightest of openings** to see the end of the bridge through the obstruction.*	Description grounded on small patch of pixels or very non-salient aspects.
Co-reference	41.5	**42.4**	38.8	*The cloud on top left side of box only has half of **it** showing.*	—
Meta Properties	12.0	**13.9**	6.1	***Bright shot** of a girl and boy standing up straight. Her eyes are closed.*	Blurriness, brightness, overlays, and transitions of frames.

	ours	NLVR2	Spot-the-diff
Average length	23.3	15.3	10.6
Word types	6,916	6,602	2,282
Average tree depth	5.1	4.8	4.3
Average sentences	1.6	1.0	1.0

Language statistics compared to other vision-and-language datasets [1,2]

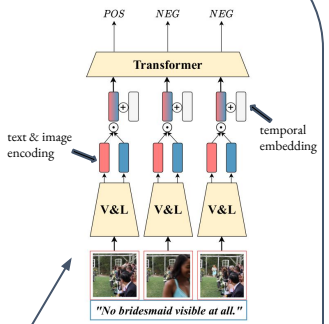
Dataset	After §3.1	After §3.3
MSR-VTT	11,643	8,045
Video-Storytelling	11,459	8,153
YouCook	894	588
Open Images	4,845	4,416

Number of images from each image source before and human retrieval verification

Modeling: Integrating Context

Model	Accuracy ViLBERT/UNITER/**CLIP** [3,4,5]
zero-shot	19.3 / 19.8 / **22.4**
random training batches	20.9 / 21.9 / **24.3**
hard negative training batches	20.9 / 24.8 / **28.4**
+ Context Module	22.3 / 24.4 / **27.7**
+ Temporal embedding	24.5 / 25.7 / **29.9** large gap to human 91%

Test Accuracy on ImageCoDe (Top-1 Retrieval Accuracy)



Best performing model
- CLIP backbone (V&L)
- Context Module on top to attend/compare all images
- temporal embedding

Takeaways

- new task in VL and pragmatic language understanding with large gap to humans
- videos more challenging and interesting than static images
- Adding necessary context to models helps only a little

Check out our project page and leaderboard!
https://mcgill-nlp.github.io/imagecode/

*For brevity we only show a subset of the 10 images for these examples

References:
[1] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A Corpus for Reasoning about Natural Language Grounded in Photographs
[2] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to Describe Differences Between Pairs of Similar Images
[3] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks.
[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation Learning
[5] Alec Radford, et al. Learning transferable visual models from natural language supervision.