Image Retrieval from Contextual Descriptions

Benno Krojer
Mila/McGill

Vaibhav Adlakha
Mila/McGill

Vibhav Vineet
Microsoft Research

Yash Goyal
Samsung SAIT AI Lab

Edoardo Ponti
Mila/McGill

Siva Reddy
Mila/McGill &
Facebook CIFAR AI Chair

An example description:
No bridesmaid visible at all.

Image 2

Does the description fit?

Mila

No bridesmaid visible at all.

Image 1

Image 2

With previous context not so much anymore

No bridesmaid visible at all.

Image 4

What about this one?

No bridesmaid visible at all.

Image 4

Image 5

With next context not a good fit

No bridesmaid visible at all.

Image 3

Image 1

Image 2

Image 4

Image 5

This small task example shows:

Language is often very contextual

This small task example shows:
Language is often very contextual

The ImageCoDe challenge:
Retrieve an image from 10 minimally contrastive images, based on a description

The girl in blue is to the left of the girl in the middle with the purple shoes. The girl in blue is not obscured in any way.

contextual description

retrieve

Mila

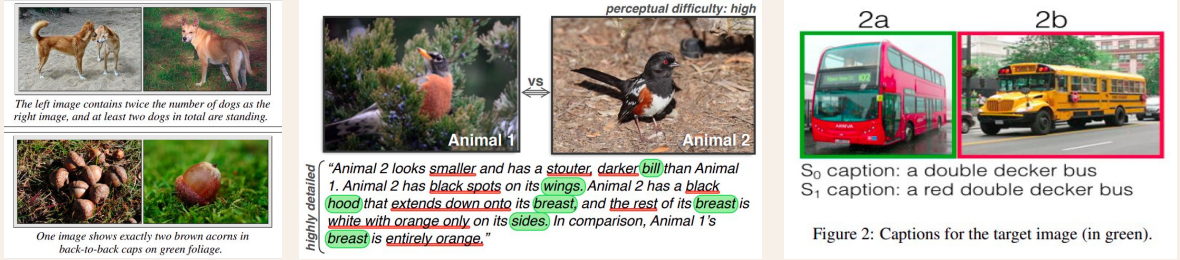
Vision & Language history: tasks

Vision & Language history: tasks

single image (VQA, ReferItGame,…)

Vision & Language history: tasks

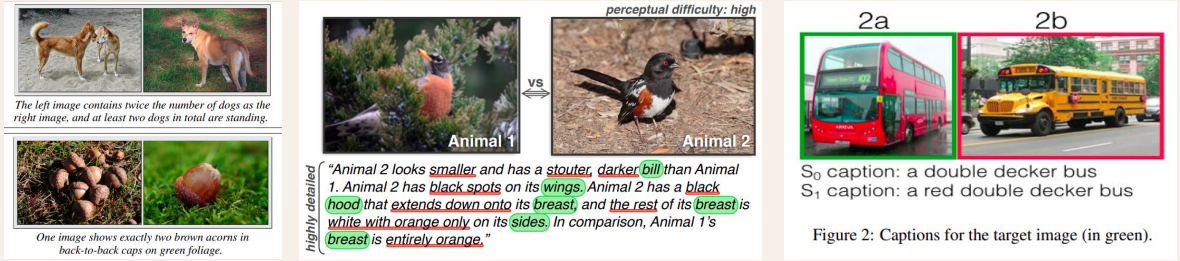
single image (VQA, ReferItGame,…)



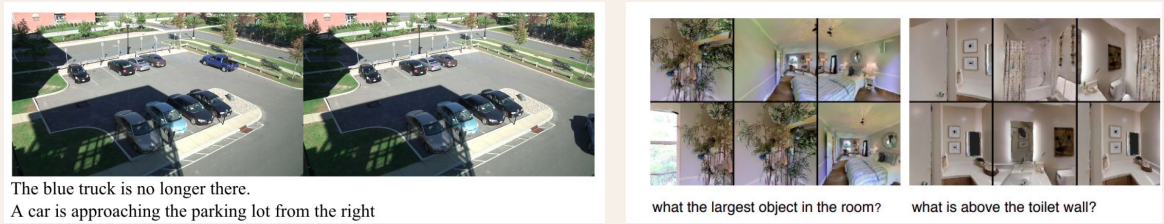
several images & complex long-form or pragmatic language (NLVR2,Neural Naturalist, Cohn-Gordon,…)

Vision & Language history: tasks

single image (VQA, ReferItGame,…)



several images & complex long-form or pragmatic language (NLVR2,Neural Naturalist, Cohn-Gordon,…)

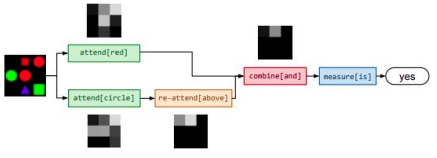


videos as a source of (similar) images (Spot-the-diff, ISVQA,...)

Vision & Language history: models

Vision & Language history: tasks

- previously often specific architectures for specific tasks (FiLM, Neural Module Networks, pragmatics with RSA,...)



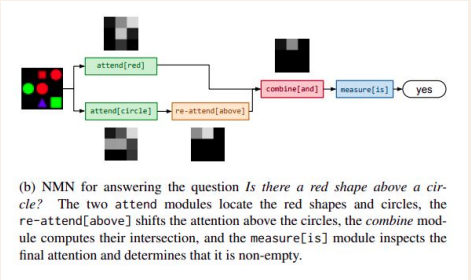
(b) NMN for answering the question *Is there a red shape above a circle?* The two `attend` modules locate the red shapes and circles, the `re-attend[above]` shifts the attention above the circles, the *combine* module computes their intersection, and the `measure[is]` module inspects the final attention and determines that it is non-empty.

NMMs

Vision & Language history: tasks

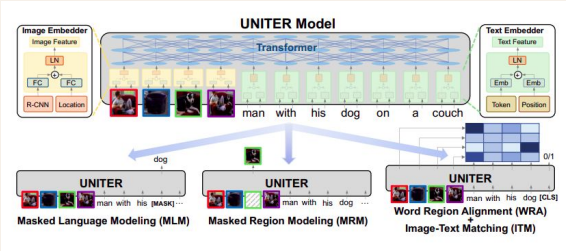
- previously often specific architectures for specific tasks (FiLM, Neural Module Networks, pragmatics with RSA,...)

- recent success of Transformers since 2019 (VilBERT, UNITER,...)
  → learning general purpose multi-modal representations for downstream tasks via cross-encoding



(b) NMN for answering the question *Is there a red shape above a circle?* The two `attend` modules locate the red shapes and circles, the `re-attend[above]` shifts the attention above the circles, the *combine* module computes their intersection, and the `measure[is]` module inspects the final attention and determines that it is non-empty.

NMMs



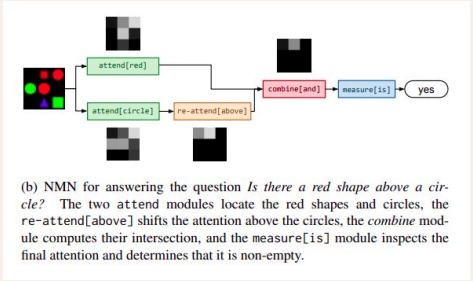
UNITER

Vision & Language history: tasks

- previously often specific architectures for specific tasks (FiLM, Neural Module Networks, pragmatics with RSA,...)

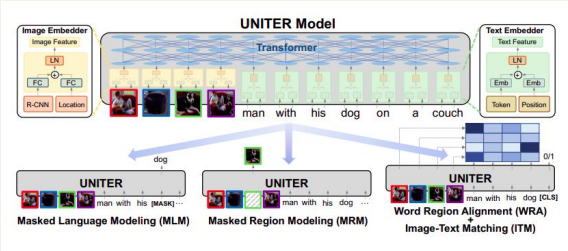
- recent success of Transformers since 2019 (VilBERT, UNITER,...)
  → learning general purpose multi-modal representations for downstream tasks via cross-encoding

- CLIP (Radford et al., 2021) has been widely adopted for its broad zero-shot capabilities
  → contrastive objective with two separate encoders

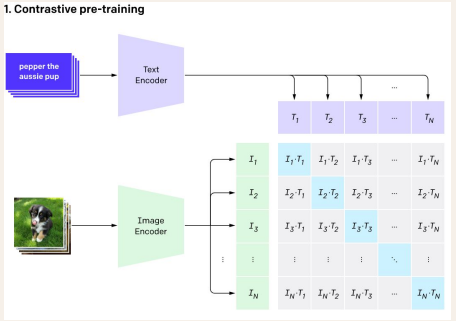


(b) NMN for answering the question *Is there a red shape above a circle?* The two `attend` modules locate the red shapes and circles, the `re-attend[above]` shifts the attention above the circles, the *combine* module computes their intersection, and the `measure[is]` module inspects the final attention and determines that it is non-empty.

NMMs



UNITER



CLIP

What are shortcomings of these models? Which tasks expose them?

What are shortcomings of these models? Which tasks expose them?

- intuition:
  - vision: encoding is the same for minimally contrastive images?
  - language: can only process short simple text?
  - CLIP’s encoders do not interact

What are shortcomings of these models? Which tasks expose them?

- intuition:
  - vision: encoding is the same for minimally contrastive images?
  - language: can only process short simple text?
  - CLIP's encoders do not interact

→ ImageCoDe: long, nuanced and contextual/pragmatic descriptions to distinguish an image among a set of 10 minimally contrastive images

What are shortcomings of these models? Which tasks expose them?

- intuition:
  - vision: encoding is the same for minimally contrastive images?
  - language: can only process short simple text?
  - CLIP's encoders do not interact

→ ImageCoDe: long, nuanced and contextual/pragmatic descriptions to distinguish an image among a set of 10 minimally contrastive images

Sounds simple but gives rise to many interesting phenomena once images are sufficiently similar!

Creating the Dataset

Step 1:

Minimally contrastive images from videos and Open Images based on heuristics and distance of CLIP visual encodings

dist = 0.17

dist = 0.2

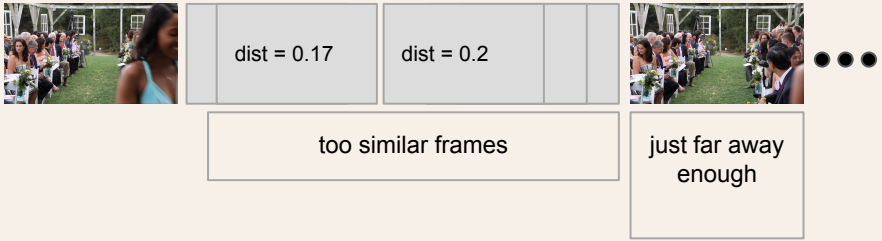
too similar frames

just far away enough

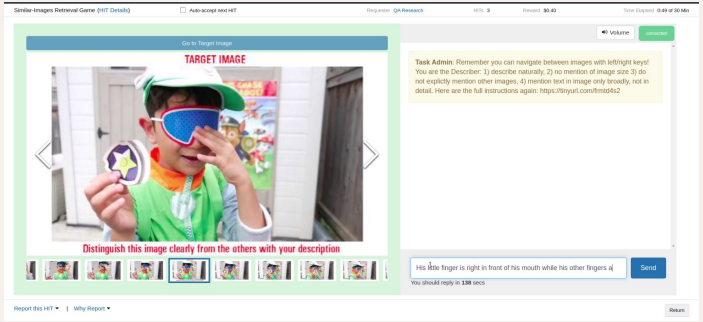
9 nearest neighbours

Mila

Step 1:
Minimally contrastive images from videos and Open Images based on heuristics and distance of CLIP visual encodings



Step 2 & 3:
Crowdsource descriptions and retrieval



- **describing** only relevant differences with no mentions of other images/order (description can function as stand-alone caption)
- retaining valid descriptions via 1-3 human **retrievals**
- avoiding **annotator bias** in descriptions

Mila

End result: The ImageCoDe Challenge

- 94,020 images (resulting in 9,402 image sets)
- 21,202 descriptions
- ~80% video-based (Video-Storytelling, YouCook, MSR-VTT), ~20% Open Images

Challenging phenomena in ImageCoDe

Challenges in ImageCoDe

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.

Table 2: Distribution of challenging phenomena in IMAGECODE based on 200 (or 1000 if underlined) manually annotated examples.

	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

Table 4: Comparison of the text statistics of IMAGECODE with other vision-and-language datasets.

There is <u>no hand visible</u> and almost all of the cardboard box in the bottom right is visible.

CONTEXT, NEGATION & VISIBILITY

NUANCES

The person's palm is towards us and touching the left bottom corner of the cake. There is a small amount of dark space between the right bottom corner of the photo and the edge of the cake.

Modeling: Incorporating Context

Experiments with 3 types of VL models

- CLIP (bi-encoder)
- ViLBERT (cross-encoder)
- UNITER (single-stream-encoder)

Experiments with 3 types of VL models

- CLIP (bi-encoder)
- ViLBERT (cross-encoder)
- UNITER (single-stream-encoder)

Adding context to the models:

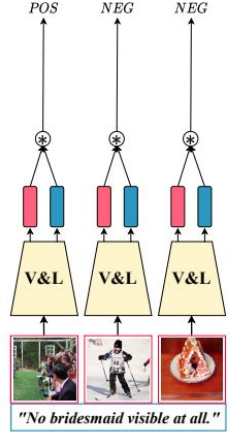
1. hard negative in batch
2. context of other images
3. temporal context

Performance: zero-shot

Model	Accuracy ViLBERT/UNITER/**CLIP**
zero-shot	19.3 / 19.8 / **22.4**

Performance: fine-tuning random batch

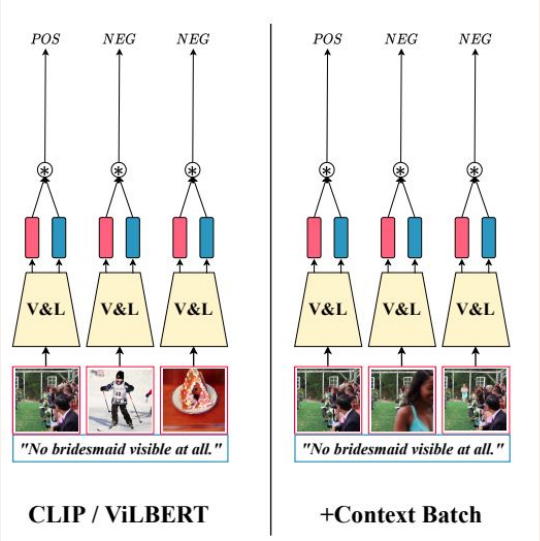
Model	Accuracy ViLBERT/UNITER/**CLIP**
zero-shot	19.3 / 19.8 / **22.4**
random training batches	20.9 / 21.9 / **24.3**



CLIP / ViLBERT

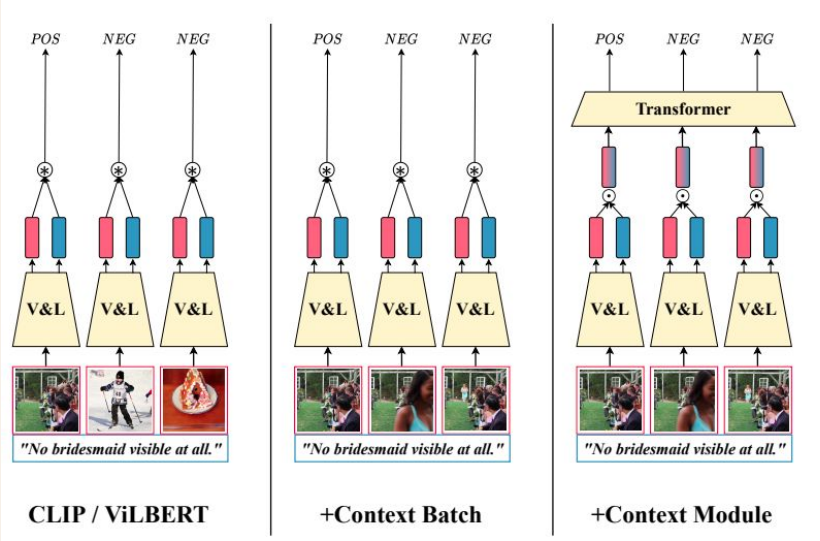
Performance: fine-tuning hard negatives

Model	Accuracy ViLBERT/UNITER/**CLIP**
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**



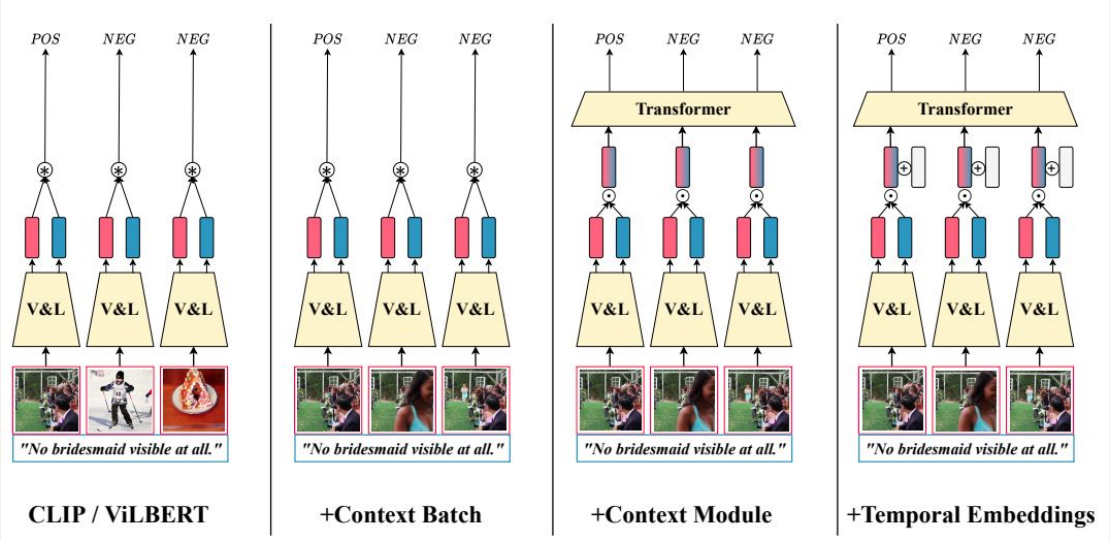
Performance: adding Context Module

Model	Accuracy ViLBERT/UNITER/**CLIP**
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**



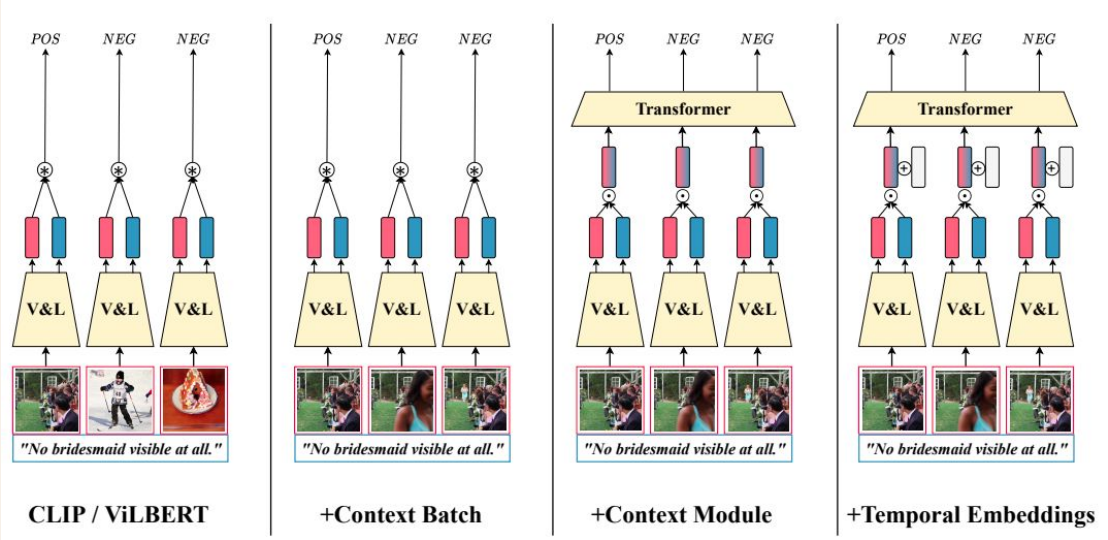
Performance: adding temporality

Model	Accuracy ViLBERT/UNITER/**CLIP**
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**



Performance: adding temporality

Model	Accuracy ViLBERT/UNITER/**CLIP**
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**



only ~30% Accuracy compared to human performance of 91%!

What causes the low performance?

- **biggest factor: video-based image sets**
  → best model achieves 60% on Open Images but only 22% on videos

What causes the low performance?

- **biggest factor: video-based image sets**
  → best model achieves 60% on Open Images but only 22% on videos

- 4 phenomena lead to large performance drops:
  nuances, negation, visibility/occlusion, context

What causes the low performance?

- **biggest factor: video-based image sets**
  → best model achieves 60% on Open Images but only 22% on videos

- 4 phenomena lead to large performance drops:
  nuances, negation, visibility/occlusion, context

- description length

There are many tasks already.
Why another one?

Several factors:

- very high similarity (requires “System 2 reasoning” from humans)
- stand-alone contextual captions with only relevant differences
- diverse domains
- 10 images

→ diverse phenomena, pragmatics/implicatures/ambiguities, reasoning, linguistically complex descriptions

- Finally: empirically we see a large gap to humans!

Takeaways

- A new challenge for the VL community and more broadly a challenge in contextual language understanding
  → Suitable for exploring pragmatic models

- Adding the necessary context to models helps only a little

We invite the community to contribute to our leaderboard!
mcgill-nlp.github.io/imagecode