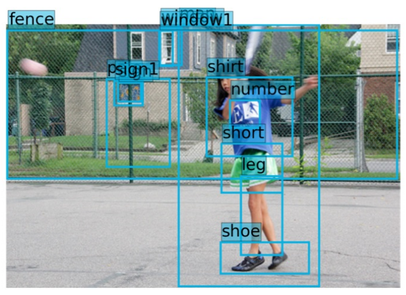
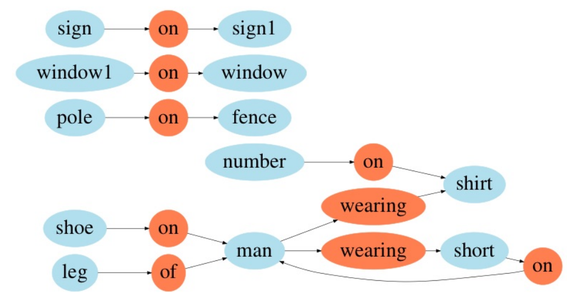
Weakly-Supervised Learning of Visual Relations in Multimodal Pretraining

Emanuele Bugliarello | Aida Nematzadeh | Lisa Anne Hendricks
emanuele@di.ku.dk
Google DeepMind
UNIVERSITY OF COPENHAGEN

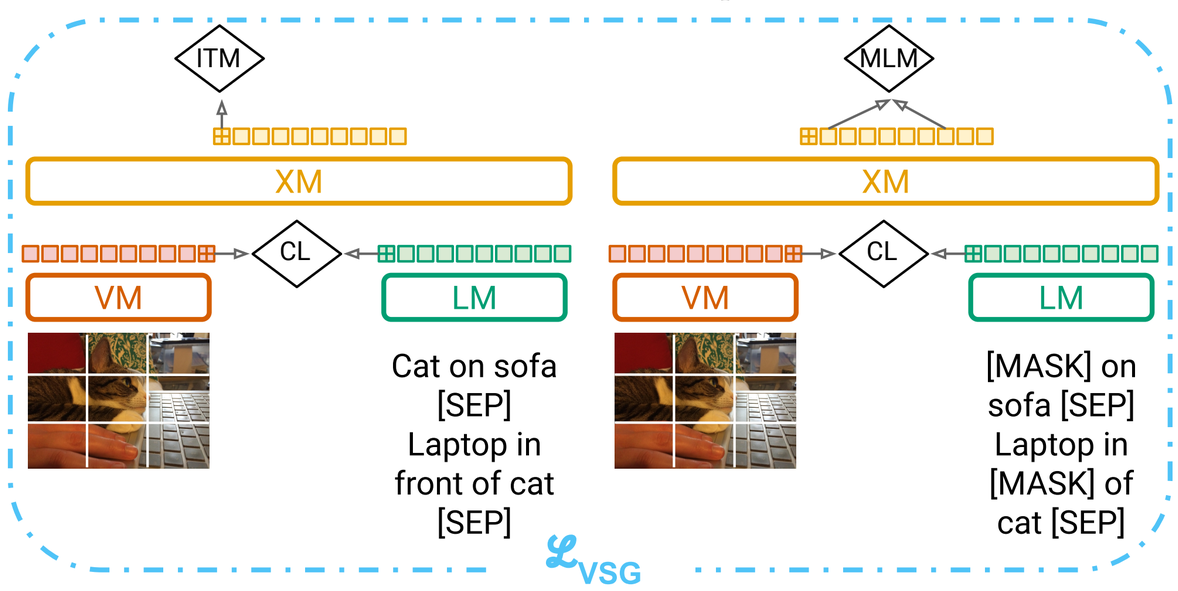
We improve fine-grained understanding in VLMs by modelling the *structure* of visual scenes with a *small* amount of human-annotated **scene graphs**





Learning Visual Relations

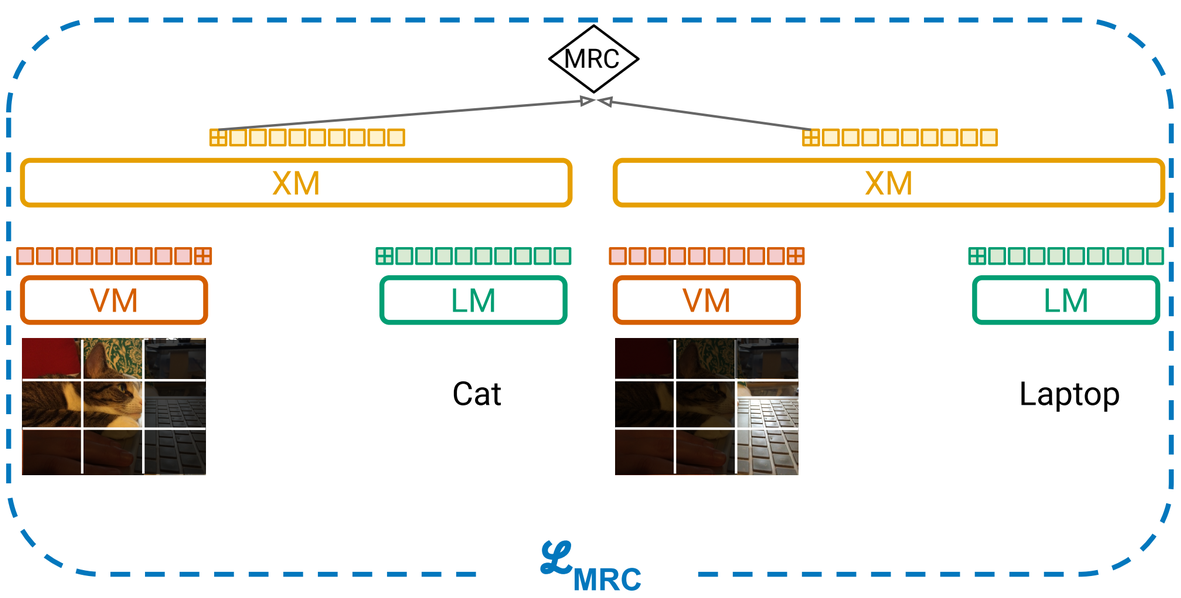
Verbalised Scene Graphs (VSG)



Data-to-text strategy

1. Sample K triplets
2. Sort them based on the subject location
3. Verbalise into a caption: "[CLS] s_1 r_1 o_1 [SEP] ... s_K r_K o_K [SEP]"
4. Apply standard (*e.g.*, ALBEF) image—text losses

Masked Relation Classification (MRC)



Pretraining cross-entropy objective

1. Encode a triplet's Subject and Object independently (by masking their visual contexts)
2. Pool their final cross-modal representations ([CLS] token)
3. Concat pooled representations and map them to V outputs (relation labels) with an MLP

Zero-Shot Evaluation Tasks

Image—Text Retrieval

Coarse-grained
A person is riding a horse.

Dense Fine-grained
A person with long hair and beige sweater is smiling and riding …

Fine-grained SVO-Probes

A woman **lying** with a dog

Fine-grained VSR

Caption: The cow is **ahead of** the person
Label: FALSE

Fine-grained VALSE

pieces	**existence**	**plurality**	**counting**
instruments	*existential quantifiers*	*semantic number*	*balanced, adversarial, small numbers*
caption (blue) / foil (orange)	*There are no animals / animals shown.*	*A small copper vase with some flowers / exactly one flower in it.*	*There are four / six zebras.*
image			

Other pieces:

- **relations**
- **actions**
- **coreference**

Scene Graphs for Fine-grained Understanding

Baselines

- **ALBEF** (coarse-grained)
- **X-VLM** (fine-grained: ALBEF+bbox prediction)

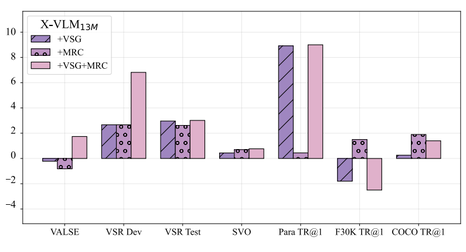
Relation-enhanced (ours)

- **ReALBEF** (ALBEF + VSG + MRC)
- **ReX-VLM** (X-VLM + VSG + MRC)

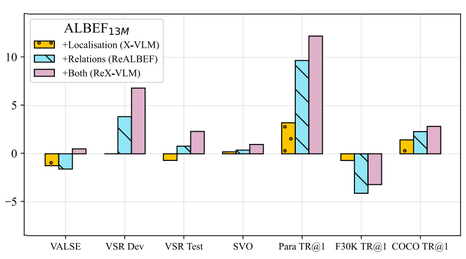
Model Name	Role	VSR Random Dev / Test Acc	VALSE Acc_r	SVO-Probes Acc_r	Stanford Paragraphs IR@1/5	Stanford Paragraphs TR@1/5
$ALBEF_{13M}$	BASELINE	60.4 / 59.4	72.2	86.7	77.1 / 93.7	73.7 / 90.3
$REALBEF_{13M}$	+RELATIONS	64.6 / 61.3	70.4	87.5	86.7 / 97.5	86.5 / 97.2
$X\text{-}VLM_{13M}$	+LOCALISATION	61.1 / 60.5	71.3	87.3	80.3 / 94.9	76.8 / 92.4
$REX\text{-}VLM_{13M}$	+BOTH	**68.4** / **63.5**	**73.3**	**88.1**	**89.3** / **98.0**	**88.8** / **97.7**

- Enhanced visual spatial reasoning capabilities
  - ReX-VLM SOTA on zero-shot VSR: +6.8/3.0 w.r.t. X-VLM
  - ReALBEF gains +3.8/0.8 w.r.t. ALBEF ⇒ modelling relations is helpful for VSR
- Improved fine-grained understanding
  - ReX-VLM gains +1.7 on VALSE and +0.8 on SVO-Probes
- Substantially better fine-grained understanding on dense captions:
  - ReX-VLM gains +9.0/12.0 on Stanford Paragraphs
- Similar trends when pretraining on 3M images but gains are higher on larger 13M Web corpora!
- Competitive coarse-grained retrieval on COCO and Flickr30K

Ablations

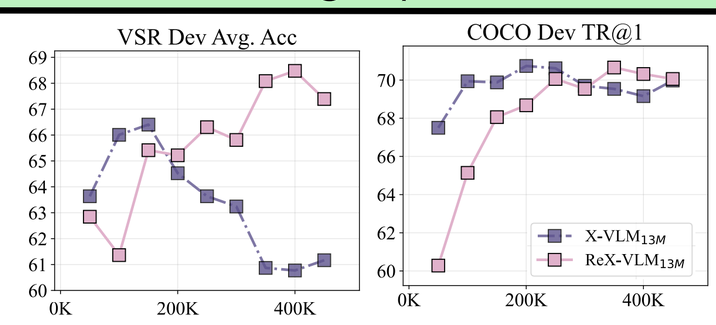


- MRC+VSG is generally best
- VSG enables image–paragraph retrieval (# rels matters)



modelling relations alone is typically *better* than modelling objects at scale

Learning Dynamics



- ReX-VLM requires longer training to achieve peak performance across fine-grained tasks
- When should we stop pretraining?
  - COCO Dev is helpful for coarse-grained
  - Not a single checkpoint for fine-grained

Conclusions

Two new approaches to learn from scene graph data

Improvements on fine-grained tasks

- Small supervised datasets are useful!
- Future work: automatic large-scale data generation

Pretraining checkpoint selection can be important

- When to stop pretraining?
- How to balance performance across tasks?