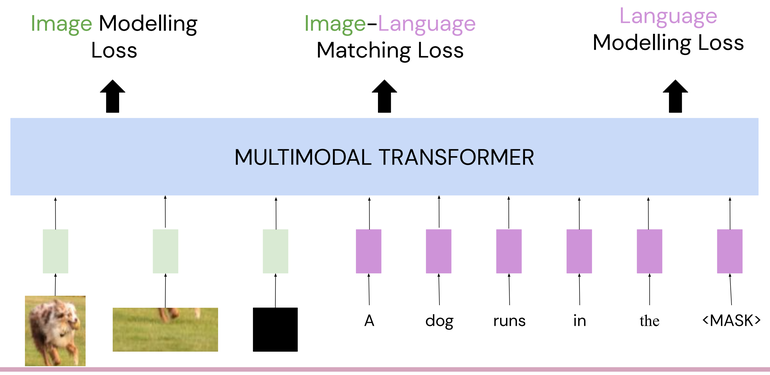
Probing Image-Language Transformers for Verb Understanding

Lisa Anne Hendricks and Aida Nematzadeh
DeepMind

Paper at https://arxiv.org/pdf/2106.09141.pdf

Do SOTA multimodal transformers (MMTs) have fine-grained verb understanding?

MMTs are SOTA on most language-vision tasks. We are interested in shedding light on the quality of their pretrained representations, and in particular, if they have a good understanding of verbs.



Why verbs?

Concrete nouns are **consistent** and **easily observable.**

Verbs are less so, as they capture **relations** → require more structured understanding.

The noun *apple* across multiple images.

The verb *eat* across multiple images.

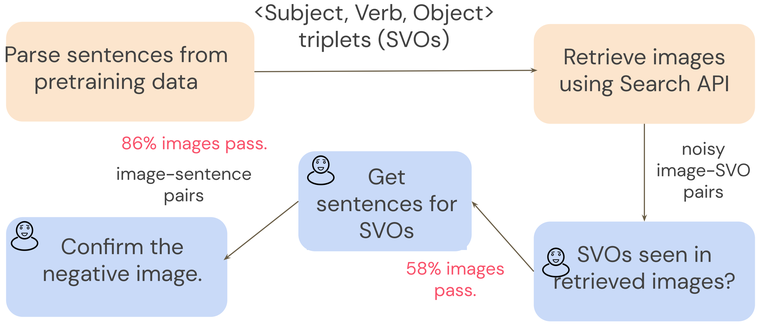
SVO-Probes

What does image retrieval test?

A person is ***riding*** a horse

Do not need to understand "riding" to sort images!

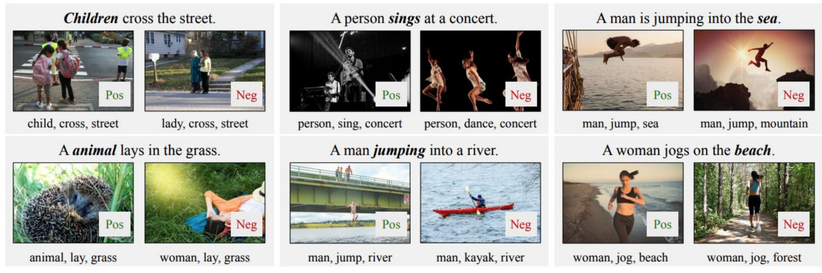
Data Collection Pipeline



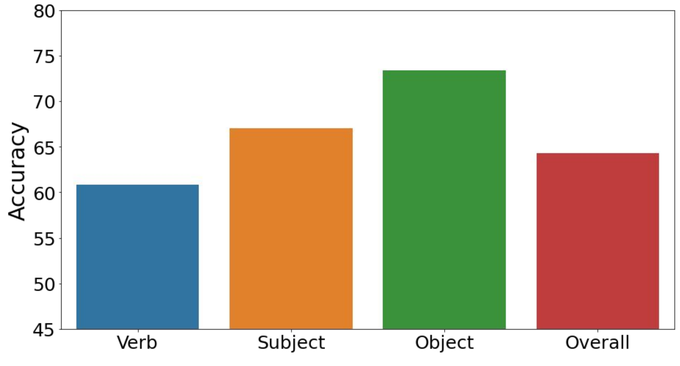
Comparison to Other Datasets

	# Verbs	Probes noun understanding	Probes verb understanding	# Sentences
Flickr	n/a	?	?	5k
FOIL	0	✔	✘	64k
HICO	117	✔	✔	0
SVO-Probes	421	✔	✔	48k

Examples

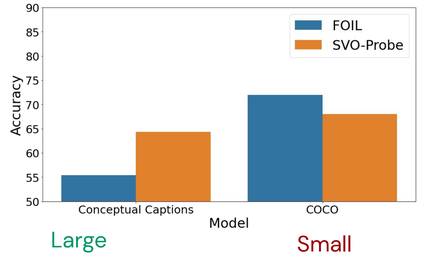


Verb Understanding in MMTs



- Performance on verbs consistently worst
- Controlled for: frequency of words at train time, similarity of positive and negative words, similarity of test image features to train image features and verbs are consistently harder than subjects and object
- Models particularly struggle with classifying negative examples

Impact of Pretraining Dataset

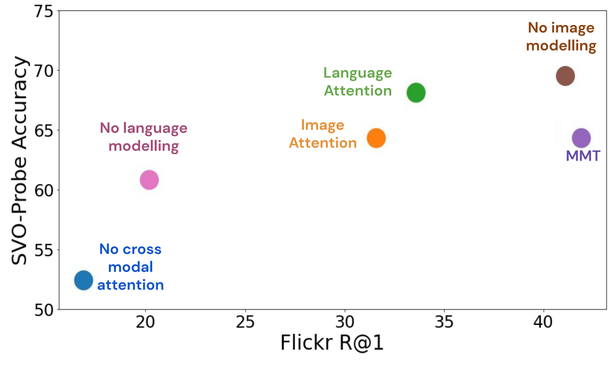


Large
Noisy
Domain matches SVO-Probes

Small
Clean
Domain mismatch from SVO-Probes

Models pretrained on COCO do better than models trained on Conceptual Captions. We hypothesize this is because COCO is cleaner (images match text better) and MMTs are not robust to noise.

SVO-Probes vs. Image Retrieval Performance



Models with weaker image modelling perform better on SVO-Probes than full MMT model.