NatuRel: Advancing Relational Understanding in Vision-Language Models with Natural Language Variations

Ashna Khetan, Isabel Sieh, Laya Iyer | {ashnak, isabelrs, laya}@stanford.edu

CS22N Winter '24

Motivation

Vision-Language Models (VLMs) have demonstrated **remarkable capabilities** across various tasks (e.g. text-image generation and image-text retrieval). However, the recent Winoground benchmark testing **compositional understanding**, indicates that embedding-based VLMs like CLIP, Flava, VisualBERT **perform near chance**.

E.x. Match two images with two captions with two words swapped

"there is [a mug] in [some grass]"
"there is [some grass] in [a mug]"

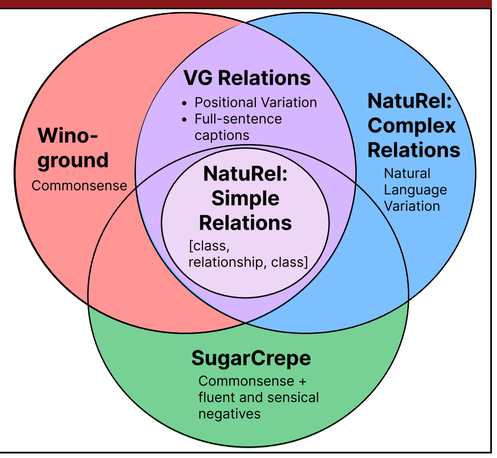
This is particularly evident with **unconventional captioning** (e.g., "the water rests below the sail"). We introduce **NatuRel**, an image-caption dataset with enriched with captions ranging from simple to complex, and **varied in sentence structure** and **natural language expression**. We aim to *bridge gaps in visio-linguistic compositional understanding*.

Goals

- Improve understanding of object positionality by fine-tuning a VLMs, CLIP and SigLIP
- Create a large image dataset with positional information with captions that have more natural language and sentence structure variation
- Understand SigLIP's compositional reasoning capabilities

Benchmarks & Data

The **benchmarks** we used to evaluate our VLMs are Winoground, ARO (VGRelations), SugarCrepe, and our generated dataset NatuRel. The features each of these datasets test for are described in the figure to the right. Winoground has 800 unique images, VGRelations has 1603, and NatuRel has 100k.



References

Yuksekgonul et al. "When and Why Vision-Language Models Behave like Bags-Of-Words...", Zhai et al. "Sigmoid Loss for Language Image Pre-Training", Thrush et al. "Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality", Kuznetsova et al. "The Open Images Dataset V4", Diwan et al. "Why is Winoground Hard? Investigating Failures in Visuolinguistic Compositionality"

NatuRel: Simple & Complex

100,000 images and **660,000 image-caption-pair examples** with **330 unique** (class, relationship, class) triplets, and **~2,500 captions**.
Images and annotations come from OpenImages V4 but captions are generated

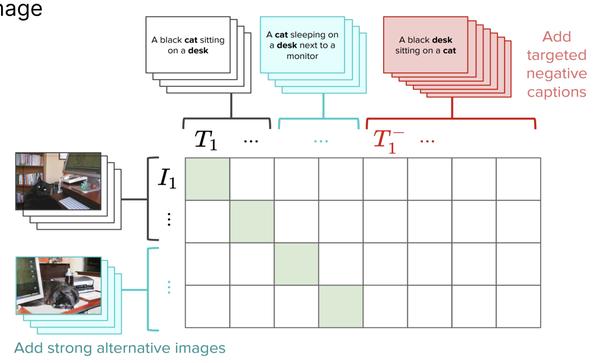
Each example has 8 captions:

1 simple positive
e.g. "cat on desk"
1 simple negative
e.g. "desk on cat"
3 complex positive
e.g. "Atop the table, sits the cat."
3 complex negative
e.g. "The cat is under the table."

Models

NegCLIP

- Contrastive Language Image Pretraining (CLIP) uses **contrastive learning**: model is exposed to positive – similar – and negative – dissimilar – examples from a dataset
- NegCLIP doubles the input matrix that CLIP takes for to allow a **hard negative** (example that is mismatched but closely resembles the correct example) **caption and image**
- We utilize NegCLIP, holding the negative image constant as a sentinel blank image



SigLIP

- Uses **sigmoid loss** (a single image-text loss instead of a global view of losses) → reduces training time by 1/5 to achieve the same accuracy
- Allows for multiple positive and negative captions for a single image

Our Fine-tuned Models

NegCLIP Single Pair: 1 randomly selected complex positive and 1 randomly selected complex negative caption from NatuRel
NegCLIP All Pairs: paired up each positive complex caption with a negative complex, paired the two simple captions, CoCo-Order, Part of Winoground
SigLIP Two Pairs: Two positive and two negative captions from NatuRel train set.

Results

Model	VG-Relations	NatuRel		Winoground (text)	
		Simple	Complex	All	Object
CLIP	0.59	0.47	0.52	0.31	0.36
SigLIP	0.46	0.38	0.49	0.14	0.12
SigLIP-TP	0.53	0.45	0.31	0.14	0.12
NegCLIP	0.81	0.87	0.62	0.31	0.31
NegCLIP-SP	0.65	0.70	0.84	0.19	0.23
NegCLIP-AP	0.80	0.99	0.84	0.28	0.33

Table 1: Results of Various Models on Various Benchmarks

Model	REPLACE			SWAP		ADD	
	Object	Attribute	Relation	Object	Attribute	Object	Attribute
CLIP	0.91	0.80	0.69	0.61	0.63	0.77	0.68
NegCLIP	0.93	0.86	0.76	0.76	0.75	0.89	0.83
NegCLIP-SP	0.89	0.74	0.68	0.60	0.62	0.81	0.66
NegCLIP-AP	0.91	0.83	0.73	0.69	0.72	0.84	0.76

Table 2: Results on SugarCrepe

Analysis

- Our models perform well on **'reversed', less-intuitive spatial** captions, consistent with how we train (see figure to right)
- We refrain from hard negative mining to encourage the model to learn its own **generalizable** patterns
- We perform similar to NegCLIP even **without the use of a negative hard image**, exemplifying the strength of natural language variations
- The disparity between Simple and Complex in NegCLIP reveals that NegCLIP may be taking a **shortcut in the Simple relation tasks**, rather than having a comprehensive understanding (assessed by Complex)

Captions NegCLIP gets correct, while NegCLIP-AP does not	Captions NegCLIP-AP gets correct, while NegCLIP does not
the cow is to the left of the grass	the pasture is to the right of the dog
the cow is to the left of the pasture	the pasture is to the right of the cows
the dog is to the left of the pasture	the grass is to the right of the cow

Future Work

- Have **annotators** review the dataset to remove malformed data
- Experiment during finetuning with **hyperparameter** tuning such as the learning rate and batch size to see how that affect accuracy
- Incorporate negative **image mining** instead of passing a white box
- Include data that **varies attributional information** of objects in an image instead of just the relationship between objects