ECCV
EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024

The Hard Positive Truth about Vision-Language Compositionality


Amita Kamath


Cheng-Yu Hsieh


Kai-Wei Chang


Ranjay Krishna

W UCLA Ai2

Vision-Language

What is Compositionality?

"The meaning of the whole is a function of the meaning of its parts."

- Recognizing the effect of word swaps and replacements on sentence meaning

… in the context of an image

- a.k.a. Fine-grained Vision-Language Understanding

Background

Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality

Thrush et al, CVPR 2022

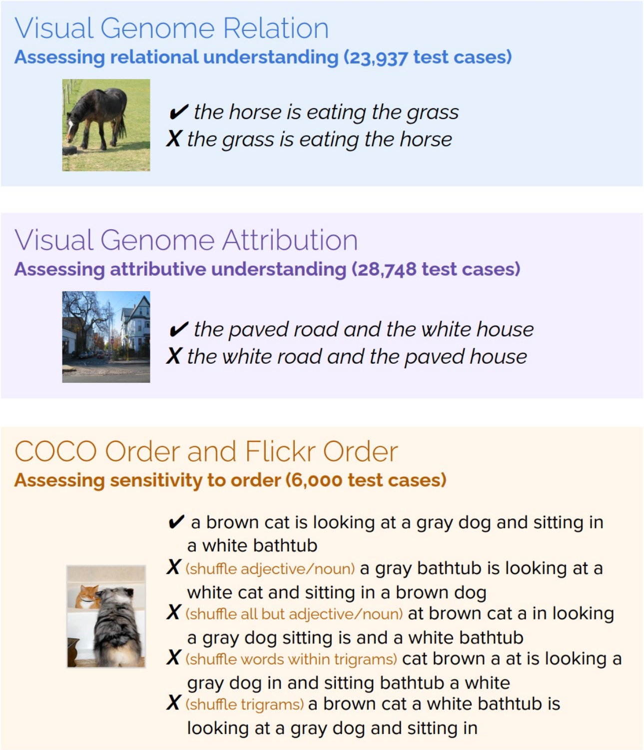
a mug in some grass

some grass in a mug

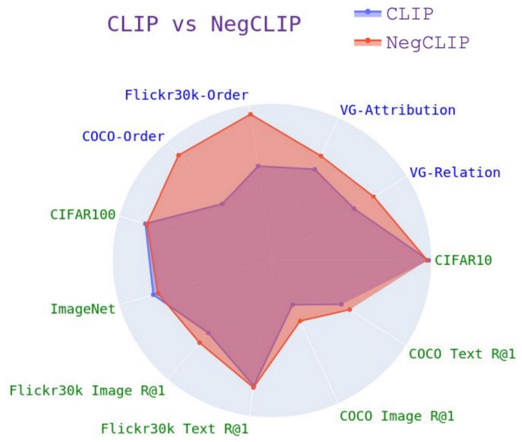
Background

When and Why VL Models Behave like Bags-of-Words, and What to Do about it

Yuksekgonul et al, ICLR 2023



Finetune



Background

Paper	Venue	Perturbation	Finetune?
Winoground	CVPR 2022 (Oral)	word order	
VL-Checklist	EMNLP 2022	replacements	
When-and-Why	ICLR 2023 (Oral)	word order	✓
CREPE	CVPR 2023 (Spotlight)	word order replacements negations	
SVLC	CVPR 2023	replacements	✓
DAC	NeurIPS 2023 (spotlight)	replacements	✓
What's Up	EMNLP 2023	replacements	✓
Text encoders…	EMNLP 2023	word order	
SugarCREPE	NeurIPS 2023	word order replacements additions	✓
COLA	NeurIPS 2023 D&B	replacements	✓

But…

the **paved road** and the white house ✓

↓

the white **road** and the **paved** house ✗

crepe on a skillet ✓

↓

boats on a skillet ✗

So…

Goal: Teach models to understand

that ~~how~~ word order / word replacements

change ~~impact~~ meaning, always

Which isn't true (and isn't compositionality)

We introduce Hard Positives

Hard Negative: semantics-altering change to the original caption

Hard Positive: semantics-retaining change to the original caption

the **paved road** and the white house ✓

the white **road** and the **paved** house ✗

the white house and the **paved road** ✓

Image i

Existing work

		Captions	CLIP	Hard Negative Finetuned
Original Caption	c	brown grass	0.236	0.152
Hard Negative	c_N	blue grass	0.240	0.143

Existing work

Image i

		Captions	CLIP	Hard Negative Finetuned	Ours
Original Caption	c	brown grass	0.236	0.152	0.240
Hard Negative	c_N	blue grass	0.240	0.143	0.231
Hard Positive	c_P	chestnut grass	0.249	0.134	0.241

Our work

Hard Positive Benchmarks

	REPLACE	SWAP
Image i		
Original Caption c	fabric on **black** table	the **black** cat and the carpeted floor
Hard Negative c_N	fabric on white table	the carpeted cat and the **black** floor
Hard Positive c_P	fabric on **ebony** table	the carpeted floor and the **black** cat
	x 27,443	x 28,748

Evaluation

	Model	REPLACE Orig. Test Acc.	REPLACE Aug. Test Acc.	SWAP Orig. Test Acc.	SWAP Aug. Test Acc.	REPLACE Brittleness (↓)	SWAP Brittleness(↓)
(a)	CLIP ViT-B/32	61.6	46.8 (-14.9)	60.5	49.6 (-10.9)	23.2	21.7
(b)	NegCLIP	68.6	52.1 (-16.6)	70.9	56.7 (-14.2)	21.5	26.4
	CREPE-Swap	63.5	50.4 (-13.1)	70.6	56.7 (-13.9)	**19.8**	26.0
	CREPE-Replace	73.7	53.9 (-19.8)	71.1	57.7 (-13.4)	23.9	25.4
	SVLC	76.6	44.5 (-32.1)	72.4	**61.6** (-10.9)	39.9	**20.8**
	SVLC+Pos	64.3	45.0 (-19.3)	56.5	45.4 (-11.1)	29.8	22.8
	DAC-LLM	87.6	48.9 (-38.7)	72.0	61.1 (-10.9)	40.1	21.6
	DAC-SAM	86.9	**55.9** (-31.0)	69.5	56.5 (-13.0)	32.5	25.6

Up to 39% drop in
reported performance!

Findings

- Models are oversensitive, and hard negative finetuning makes them even more so
    - → HNFT doesn’t help models understand *when* perturbations matter
- Oversensitivity transfers across perturbation types
- HNFT lowers scores of the original captions too
    - → hurts use cases like caption evaluation

Improving model performance

- Programmatically generate hard positives from COCO
- Finetune CLIP on hard negatives *and* hard positives

Improving model performance

	Model	REPLACE Orig. Test Acc.	REPLACE Aug. Test Acc.	SWAP Orig. Test Acc.	SWAP Aug. Test Acc.	REPLACE Brittleness (↓)	SWAP Brittleness(↓)
(a)	CLIP ViT-B/32	61.6	46.8 (-14.9)	60.5	49.6 (-10.9)	23.2	21.7
(b)	NegCLIP	68.6	52.1 (-16.6)	70.9	56.7 (-14.2)	21.5	26.4
	CREPE-Swap	63.5	50.4 (-13.1)	70.6	56.7 (-13.9)	**19.8**	26.0
	CREPE-Replace	73.7	53.9 (-19.8)	71.1	57.7 (-13.4)	23.9	25.4
	SVLC	76.6	44.5 (-32.1)	72.4	**61.6** (-10.9)	39.9	**20.8**
	SVLC+Pos	64.3	45.0 (-19.3)	56.5	45.4 (-11.1)	29.8	22.8
	DAC-LLM	87.6	48.9 (-38.7)	72.0	61.1 (-10.9)	40.1	21.6
	DAC-SAM	86.9	**55.9** (-31.0)	69.5	56.5 (-13.0)	32.5	25.6
(c)	Our HN	73.9	55.7 (-18.2)	74.3	60.5 (-13.8)	21.0	25.1
	Our HP+HN	69.0	**58.0** (-11.0)	73.2	**61.1** (-12.1)	**16.9**	**22.9**
(d)	Our HP+HN (Swap-only)	63.9	51.6 (-12.3)	73.0	**61.9** (-11.2)	18.6	**21.2**
	Our HP+HN (Replace-only)	70.9	**59.0** (-11.9)	69.7	55.6 (-14.1)	**17.8**	26.5
	Random Chance	50.0	33.3	50.0	33.3	33.3	33.3
	Human Estimate	97	97	100	100	0	0

Findings

- Adding hard positives to finetuning improves model performance
- Performance on standard benchmarks ✓
- Oversensitivity transfers across perturbations, but improved invariance does not

Check the paper for further experiments targeting different variants of CLIP, and changing the ratio between hard positives and hard negatives!

Hard Positives: an important new aspect of VL compositionality

VL models aren't compositional, and hard-negative finetuning makes them oversensitive

Our new model performs well on both hard negatives and hard positives!

Image i

Existing work

		Captions	CLIP	Hard Negative Finetuned	Ours
Original Caption	c	brown grass	0.236	0.152	0.240
Hard Negative	c_N	blue grass	0.240	0.143	0.231
Hard Positive	c_P	chestnut grass	0.249	0.134	0.241

Our work

Thank you!