Bongard-HOI: Benchmarking Few-Shot Visual Reasoning for Human-Object Interactions

Huaizu Jiang*, Xiaojian Ma*, Weili Nie, Zhiding Yu, Yuke Zhu, Anima Anandkumar

We've created a **new quest** for visual intelligence

- Images in set A depict the **same** human-object interaction (HOI);
- Images in set B **DON'T** depict this HOI;
- Can you tell whether the two **query images** depict that HOI?

What makes Bongard-HOI challenging?

- **Few-shot learning**: A problem in Bongard-HOI is **2**-way **6**-shot – the learner needs to figure out the true HOI concept with **extremely** few samples and **binary** labels, while performs reasoning with it.
- **Context-dependent reasoning**: An image in Bongard-HOI can be interpreted with different HOI, depending on the current problem "context".
- **Hard negatives**: We make images in a problem share the **same object**. Therefore, merely recognizing the object won't be helpful.

- **Real-world images**: Compared to counterparts using synthetic images (ex. Bongard-LOGO [2]), natural images impose challenges on perceptual uncertainty.

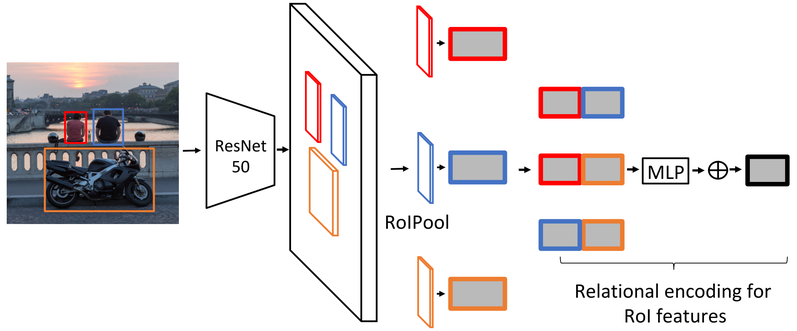
Possible models for Bongard-HOI

Few-shot learning for Bongard-HOI

- We can formulate Bongard-HOI as a few-shot learning problem. Each problem is a few-shot instance with $N = 2$ classes and $2M$ samples, and the model learns from a training set $\mathcal{S} = \mathcal{P} \cup \mathcal{N} = \{(I_1^P, 1), \ldots, (I_M^P, 1), (I_1^N, 0), \ldots, (I_M^N, 0)\}$ and is evaluated on a query image (I_q, y_q). Each example include an image I and a class label y. In our case, we set $M = 6$.
- **Non-episodic** methods: they map all the images in a few-shot instance, *i.e.* $\cup_{i=1}^{2M+1} I_i$ to the query label y_q. We evaluated two of them: **CNN-Baseline** and **WReN**.
- **Meta-learning** method: they learn to train a classifier using $2M$ samples and evaluate their trained classifier on the query (I_q, y_q). We evaluate four mete-learning methods in our experiments: **ProtoNet**, **MetaOptNet**, **ANIL** and **Meta-Baseline.**

Image encoding with relational network

- Since Bongard-HOI requires the learner to identify the compositional concepts from the images, it is a common practice to help the model reason with relations by using **relational inductive bias**. Here we use Relational Network [3] to encode the pairwise relations of the objects detected in an image.



Generalization tests

- We introduce four generalization tests to investigate whether a learner can transfer its few-shot learning skill to novel HOI concepts:

Main results

	bbox	pre-train	test set: seen act., seen obj.	test set: seen act., unseen obj.	test set: unseen act., seen obj.	test set: unseen act., unseen obj.	avg.
CNN-Baseline [33]	-	scratch	50.03	49.89	49.77	50.01	49.92
WReN-BP [2,33]	-	IN	50.31	49.72	49.97	49.01	49.75
ProtoNet* [43]	det	IN	-	-	-	-	-
ProtoNet [43]	gt	IN	58.90	58.77	57.11	58.34	58.28
MetaOptNet# [23]	det	IN	-	-	-	-	-
MetaOptNet [23]	gt	IN	58.60	58.28	58.39	56.59	57.97
ANIL [35]	det	IN	50.18	50.13	49.81	48.83	49.74
ANIL [35]	gt	IN	52.73	50.11	49.55	48.19	50.15
Meta-Baseline [6]	det	scratch	54.61	53.79	54.58	53.94	54.23
Meta-Baseline [6]	det	MoCoV2	55.23	54.54	54.32	53.11	54.30
Meta-Baseline [6]	det	IN	56.45	56.02	55.60	55.21	55.82
Meta-Baseline [6]	gt	IN	58.82	58.75	58.56	57.04	58.30
HOITrans [54] (oracle)	-	-	59.50	64.38	63.10	62.87	62.46
Human (Amateur)	-	-	87.21	90.01	93.61	94.85	91.42

- Most of the models perform worse on challenging test sets with unseen actions or objects.
- Meta-learning methods are generally better than non-episodic method, but still largely fall behind amateur human testers.
- Even a "oracle" model, which is trained to detect HOIs from images and has seen all the objects and actions, cannot perform well on our tasks.

What can we learn from the results on Bongard-HOI so far?

- **We need holistic perception and reasoning**. Models that have only good perception (HOITrans) are likely to fail. Rather, an ideal learner needs to integrate visual perception in natural scenes and detailed cognitive reasoning.
- **Pre-training improves performances**. We can see from above that pre-training is very helpful. Compared to no pre-training, using either manual labels or self-supervision leads to a performance boost.
- **Visual perception matters in Bongard-HOI**. Natural scenes rich visual stimuli while bring challenges to reliable perception. In our case, the detected bounding boxes can be noisy and therefore hurt the reasoning.

References

[1] Bongard, M. M. (1970). Pattern Recognition
[2] Nie, W. et al. (2020) Bongard-LOGO: A New Benchmark for Human-Level Concept Learning and Reasoning
[3] Santoro, A et al. (2017), A simple neural network module for relational reasoning

Paper

Code