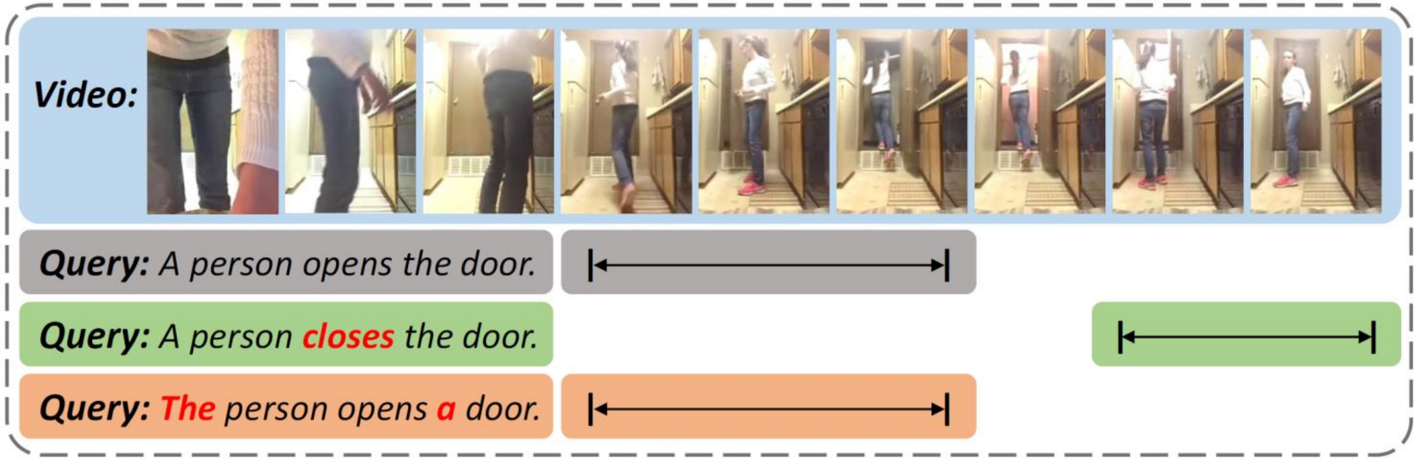
Exploring the Effect of Primitives for Compositional Generalization in Vision-and-Language

Chuanhao Li[1], Zhen Li[1], Chenchen Jing[3], Yunde Jia[2,1], Yuwei Wu[2,1]

[1]Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science & Technology, Beijing Institute of Technology, China
[2]Guangdong Lab of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, China
[3]School of Computer Science, Zhejiang University, Hangzhou, China

{lichuanhao, li.zhen, jiayunde, wuyuwei}@bit.edu.cn
jingchenchen@zju.edu.cn

Motivation



- An indispensable premise for improving **compositional generalization** is to understand the effect of the primitives, including words, image regions, and video frames. Primitives are compositional building blocks mainly involved in V&L tasks and the determinants of sample semantics.
- Existing methods cannot correctly establish the relationship between the primitives and the sample semantics and thus the ground-truth, so they cannot achieve compositional generalization.

Motivation

We present a self-supervised learning based framework that equips existing V&L methods with

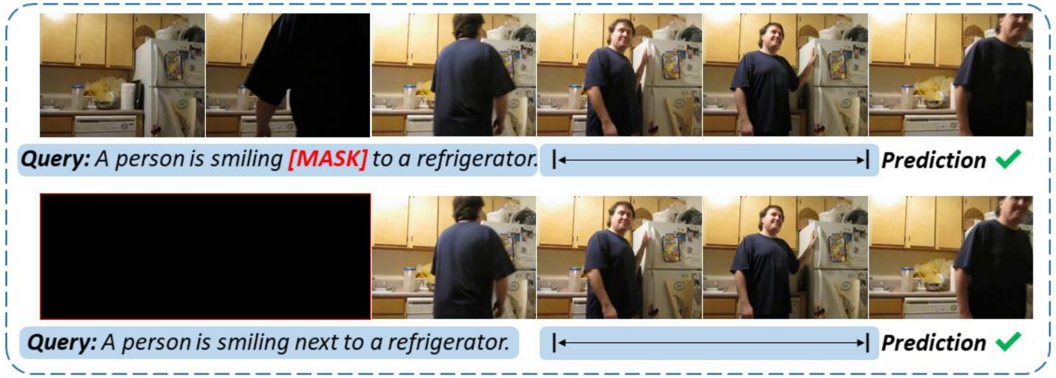
- semantic equivariance
- semantic invariance

by generating numerous labeled training samples, including

- equivariant samples
- invariant samples

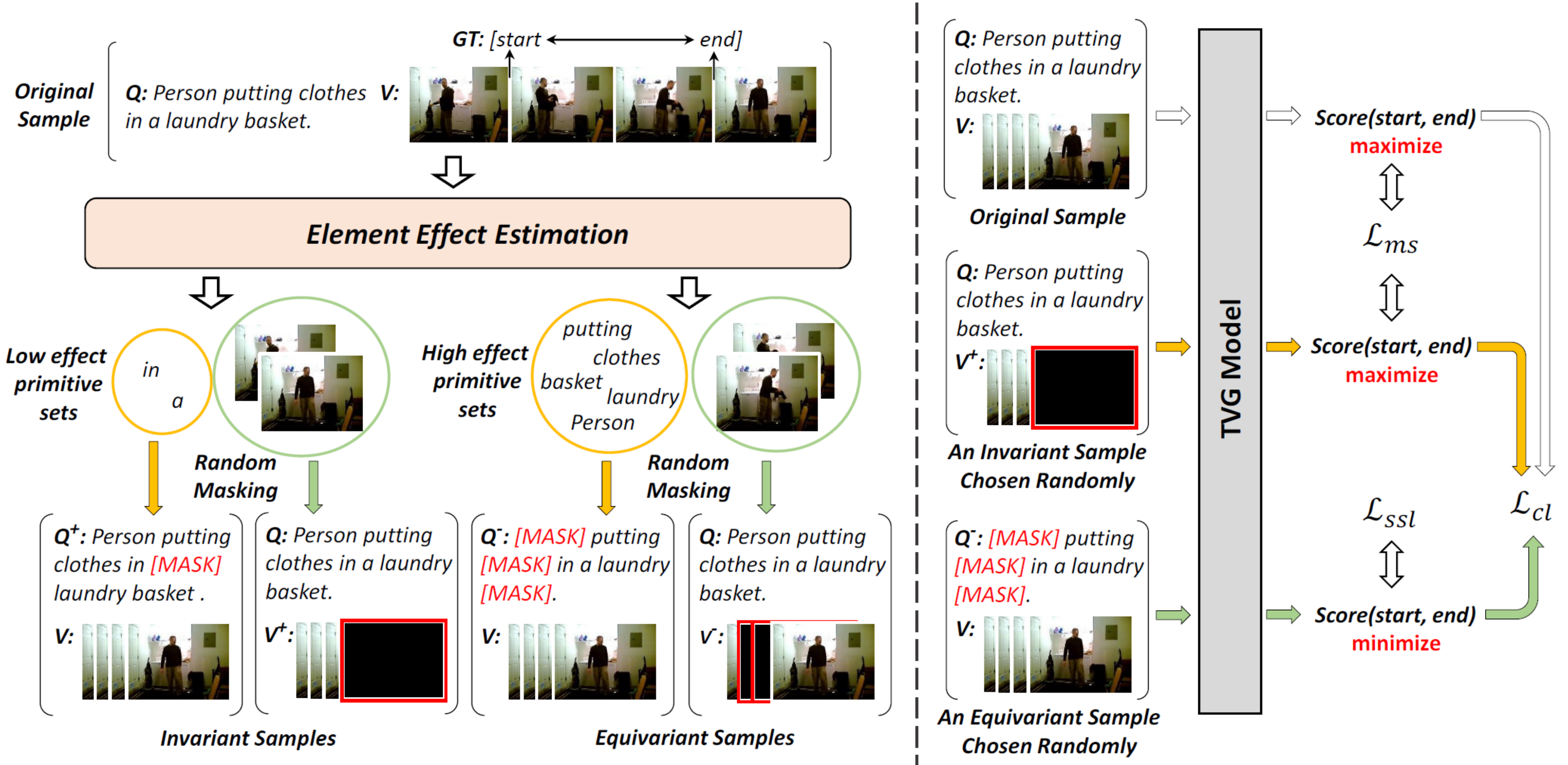
(a) An original example in the context of temporal video grounding.

(b) Equivariant samples generated by masking critical primitives.



(c) Invariant samples generated by masking irrelevant primitives.

Framework



(a) Invariant and equivariant samples generation

(b) Training TVG model with generated samples

Effect Estimation of Primitives

- For words,

  nouns/verbs: α, adjectives/adverbs: β, other words: γ

- For image regions,

  word-region similarities: pre-trained CLIP[1]

- For video frames,

  frame-query similarities: pre-trained TCL[2]

We quantify all of the effect of primitives as numbers in the interval [0, 1].

[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." ICML. 2021.
[2] Yang, Jinyu, et al. "Vision-language pre-training with triple contrastive learning." CVPR. 2022.

Sample Generation

➢ Equivariant samples

Definition: a series of samples that have different semantics from the original samples.

Generation: randomly mask primitives with high effect.

➢ Invariant samples

Definition: a series of samples that have same semantics from the original samples.

Generation: randomly mask primitives with low effect.

Optimization

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

➢ Method-specific Loss

$$\mathcal{L}_{ms} = f(P(V,Q),Y) + \lambda_i f(P(V^i,Q^i),Y),$$

➢ Self-supervised Learning Loss

$$\mathcal{L}_{ssl} = u \cdot P(V^e,Q^e)[g(Y)],$$

➢ Contrastive Learning Loss

$$\mathcal{L}_{cl} = -\log\left(\frac{e^{h(P(V,Q),P(V^i,Q^i))}}{e^{h(P(V,Q),P(V^i,Q^i))} + e^{h(P(V,Q),P(V^e,Q^e))}}\right),$$

where $f(\cdot,\cdot)$ is the loss function used in the selected method, $g(\cdot)$ converts Y to its index in all categories, and $h(\cdot,\cdot)$ is cosine similarity.

Experiments

Table 1. Performance (%) of the state-of-the-art methods on the Charades-CG dataset. The best scores are bold and the second-best scores are underlined.

Type	Method	*Test-Trivial*			*Novel-Composition*			*Novel-Word*		
		R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
Weakly-supervised	WSSL [10]	15.33	5.46	18.31	3.61	1.21	8.26	2.79	0.73	7.92
RL-based	TSP-PRL [36]	39.86	21.07	38.41	16.30	2.04	13.52	14.83	2.61	14.03
Proposal-free	VSLNet [42]	45.91	19.80	41.63	24.25	11.54	31.43	25.60	10.07	30.21
	LGI [27]	49.45	23.80	45.01	29.42	12.73	30.09	26.48	12.47	27.62
	VISA [22]	53.20	26.52	47.11	<u>45.41</u>	22.71	**42.03**⋆	42.35	20.88	40.18
Proposal-based	TMN [23]	18.75	8.16	19.82	8.68	4.07	10.14	9.43	4.96	11.23
	2D-TAN [44]	48.58	26.49	44.27	30.91	12.23	29.75	29.36	13.21	28.47
	2D-TAN* [44]	48.06	27.10	43.72	32.74	15.25	31.50	37.12	18.99	35.04
	2D-TAN + Ours	53.91	31.82	46.84	35.42	17.95	33.07	43.60	25.32	39.32
	MS-2D-TAN* [43]	<u>57.85</u>	<u>37.63</u>	<u>50.51</u>	43.17	<u>23.27</u>	38.06	<u>45.76</u>	<u>27.19</u>	<u>40.80</u>
	MS-2D-TAN + Ours	**58.14**	**37.98**	**50.58**	**46.54**	**25.10**	<u>40.00</u>	**50.36**	**28.78**	**43.15**

* indicates the results from our reimplementation using official released codes.
⋆ indicates that the method can be incorporated into our framework for further improvements.

Table 4. Accuracies (%) of the state-of-the-art methods on the CLEVR and CLOSURE datasets. The HM represents the harmonic mean accuracies.

Method	CLEVR	CLOSURE	HM
MGN-e2e¶ [32]	-	80.9	-
Vector NMN† [4]	98.0	71.3	82.5
Vector NMN†‡ [4]	98.0	94.4	96.2
LG-NMN† [1]	98.9	88.0	93.1
TMN†‡ [38]	97.9	95.4	96.6
NS-VQA†§ [41]	**100**	77.2	87.1
FiLM [30]	97.0	60.1	74.2
MAC [16]	98.5	72.4	83.5
ViLBERT [26]	95.3	51.2	66.6
GLT [6]	99.1	<u>96.1</u>	<u>97.6</u>
GLT* [6]	99.1	95.0	97.0
GLT + Ours	<u>99.1</u>	**98.4**	**98.7**

¶ for methods trained with external correspondence labels.
§ for methods using domain-knowledge for deterministically execution.
† for methods trained with external layout annotations.
‡ for methods using external layout annotations when testing.
* for the results from our reimplementation using official released codes.

Summary

- Understanding the effect of primitives on ground-truth can implicitly improve the compositional generalization capability.
- The presented self-supervised learning framework can equip existing methods with semantic equivariance and semantic invariance.
- The proposed framework is capable of improving not only the compositional capability of existing methods, but also the IID generalization capability of them.

Thanks!