中国科学院软件研究所学术年会暨重点实验室科技活动周 2025 第十届
学术论文

VEglue: 通过对象对齐联合擦除测试视觉蕴涵系统

Zhiyuan Chang[1]，Mingyang Li[1]*，Junjie Wang[1], Cheng Li[1]，Wang Qing[1]*
[1]智能博弈重点实验室

ACM Transactions on Software Engineering and Methodology (TOSEM)

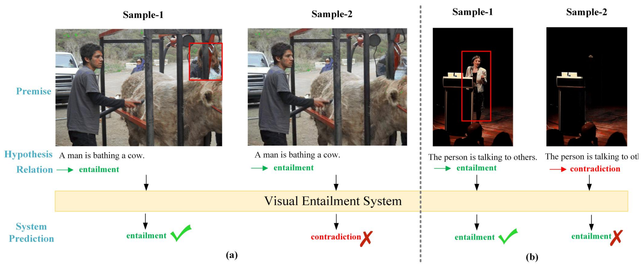
* 通讯地址：{mingyang2017, wq}@iscas.ac.cn

Introduction

- **Visual Entailment:** A multimodal reasoning task that determines if a sentence logically follows from an image, playing a key role in applications like fake news detection and visual question answering.
- **Metamorphic Testing:** A testing approach that checks whether a system behaves as expected when inputs are transformed in specific ways. It does this by designing metamorphic relations and generating new test cases to verify that the expected testing oracles still hold.

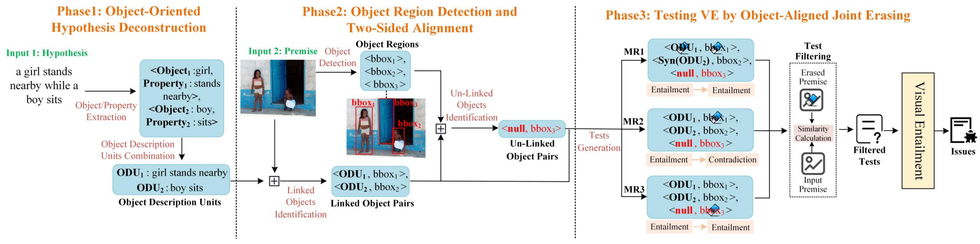
Motivation

- **Motivation-1**: Visual entailment models require inputs from multiple modalities, and these modalities are interrelated. Their performance may be affected by external disturbances.
- **Motivation-2**: Visual entailment models may sometimes infer the relationship between the premise and hypothesis based on incorrect understanding of objects (object confusion).



Approach

Approach name: **VEglue**



- **Phase 1:** Decompose the hypothesis into multiple object descriptions.
- **Phase 2:** Detect object regions and link objects between the hypothesis and premise.
- **Phase 3**: Design three metamorphic relations using joint object alignment and erasure to construct a test set for robustness evaluation of visual entailment models.

Experiment

- Dataset
  - SNLI-VE and e-SNLI-VE: Benchmark datasets for visual entailment tasks, containing 17,901 and 17,572 test samples, respectively.
- Baseline
  - TextFlint: A testing method for textual entailment tasks that changes the expected result to "contradiction" by replacing words in the hypothesis with their antonyms.
  - CAT: A translation model testing method based on the metamorphic testing framework, which perturbs the input by replacing words while keeping the expected entailment relation unchanged.
  - FIP: Perturbs images by introducing effects such as noise, blur, and weather, thereby preserving the original semantic content of the image and keeping the expected entailment relation unchanged.
- Result

Table 2. The quality, diversity and the issue detection capability of the tests generated by VEglue and baselines

VE Systems	Metric	VEglue	TextFlint	CAT	FIP			
					Impulse	*Zoom*	*Snow*	*Contrast*
OFA-VE	*INUM*	**9695**	318	3085	2110	2838	2155	1951
	IFR	**47.4%**	32.7%	17.4%	11.9%	16.0%	12.2%	11.0%
ALBEF-VE	*INUM*	**11293**	389	3191	2305	3015	2455	2128
	IFR	**55.2%**	40.0%	18.0%	13.0%	17.0%	13.9%	12.0%
LLaVA	*INUM*	**13841**	474	5585	4592	4771	4504	4451
	IFR	**69.4%**	48.7%	31.5%	25.9%	26.9%	25.5%	25.1%
GPT-4V*	*INUM*	**340**	66	64	29	46	33	36
	IFR	**37.8%**	22.1%	21.5%	9.6%	15.4%	11.0%	12.0%
-	*VTR*	96.4%	**98.3%**	80.3%	91.3%	62.7%	89.3%	76.4%
	CI	94.5%-97.9%	96.9%-99.3%	77.2%-84.2%	88.5%-93.5%	58.5%-66.9%	85.9%-91.5%	73.2%-80.6%
	DS	**13.7%**	4.9%	1.8%	6.6%	10.3%	8.1%	4.8%

Table 4. The accuracy of original/retrained models on two datasets

	SNLI-VE				e-SNLI-VE			
	VEglue	*TextFlint*	*CAT*	*FIP*	VEglue	*TextFlint*	*CAT*	*FIP*
Original OFA-VE	51.1%	64.2%	80.4%	84.5%	54.1%	70.4%	84.8%	87.9%
Retrained OFA-VE	99.7%	97.0%	87.1%	89.7%	99.9%	97.4%	87.5%	92.2%
Original ALBEF-VE	43.5%	56.9%	79.7%	85.2%	46.2%	63.1%	84.3%	86.8%
Retrained ALBEF-VE	99.4%	96.4%	86.8%	89.4%	99.0%	97.0%	87.1%	90.3%

Experiment Result: In terms of issue detection efficiency, the VEglue increases Issue Finding Rate (IFR) by 18.4% to 68.6% compared to state-of-the-art testing methods. Furthermore, by retraining with test cases generated by VEglue, the average accuracy of the retrained model improves by 50.8% over the original model.