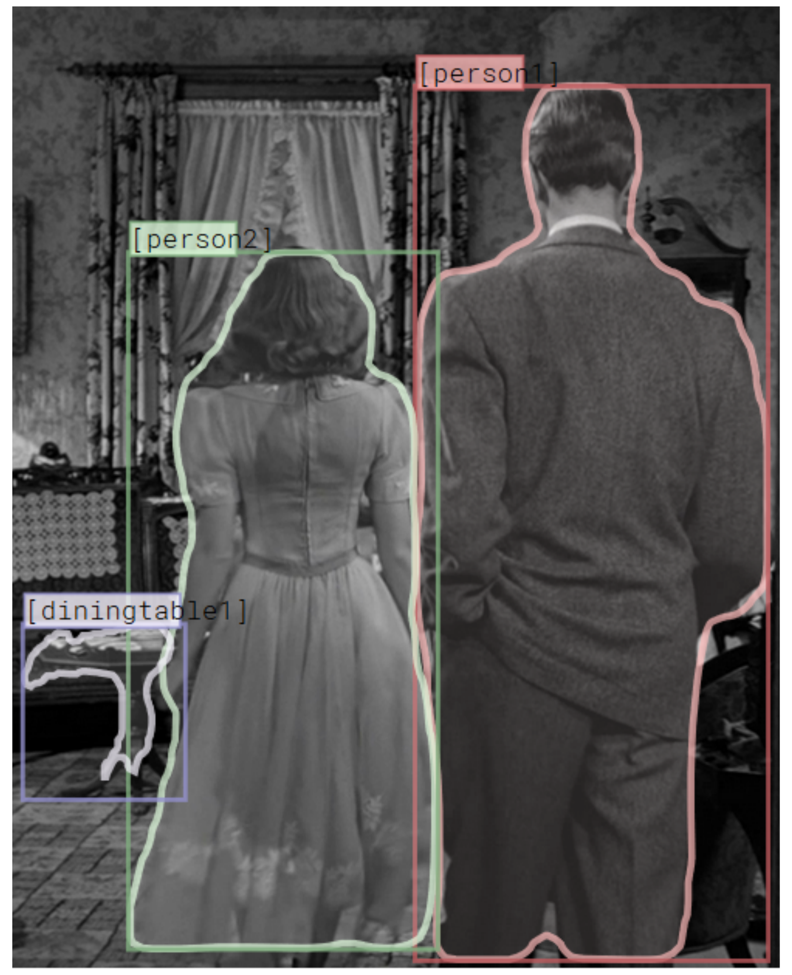
A Case study of the Shortcut Effects in Visual Commonsense Reasoning

Keren Ye, Adriana Kovashka
Department of Computer Science, University of Pittsburgh

Introduction

- Nature of supervised training
  - ➢ Methods are rewarded for finding any connection between inputs and outputs
- An example from VCR dataset
  - ➢ What does [person1] think of [person2]'s dress?



Correct answer: ***[person1]*** thinks ***[person2]*** looks stunning in her ***dress***.

Incorrect #1: She does not approve.
Incorrect #2: [person2] is a girl and girls like to wear makeup.
Incorrect #3: [person1] is confused and annoyed by [person2] following her in the store.

The correct option has the most overlap with the question

- Shortcuts
  - ➢ ***DEFINITION***: A way of achieving the correct answer by simply matching repeated references to the same entities in the question and answer options.
  - ➢ Mainly present in the ***multi-choice VQA*** tasks, which requires choosing an answer from multiple options best responding to the question-image pair
    - ❖ E.g., VCR (Zellers et al. 2019), MovieQA (Tapaswi et al. 2016), SocialIQ (Zadeh et al. 2019)

Contributions

- Point out the detrimental shortcuts in multi-choice VQA
- Quantify the impact of shortcuts on SOTA models
- Propose a curriculum masking technique for robust training

Approach

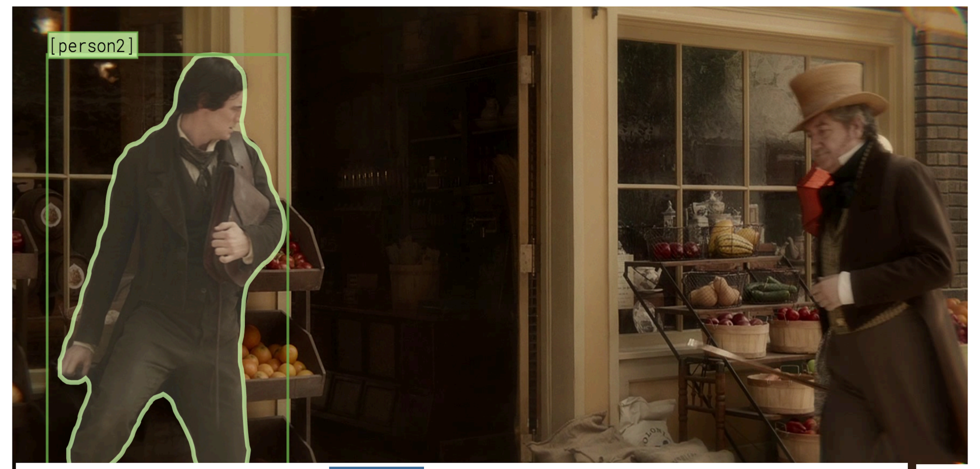
- Quantifying the shortcut effects (in VCR)
  - ➢ ***Intuition***: highlighting shortcuts, testing models' capability of utilizing comprehensive features
    - ❖ If a model relies on shortcuts, will observe ***performance drop*** in generalized settings
  - ➢ Tested four models
    - ❖ R2C (Zellers et al. 2019), HGL (Yu et al. 2019), TAB-VCR (Lin et al. 2019), B2T2 (Alberti et al. 2019)
  - ➢ Two methods to highlight misleading shortcuts
    - Rule-based modification
      - ✓ More realistic, less inflated
      - ✓ Measure precisely how much different methods rely on person tag shortcuts
    - Adversarial modification

$$\mathrm{argmax}_{i\in[1,|a|)]}[-\mathcal{C}(v,q,a)\log\mathcal{P}(v,q,\Psi(a,i);\theta)-(1-\mathcal{C}(v,q,a))\log(1-\mathcal{P}(v,q,\Psi(a,i);\theta))]$$

  - ❖ What words cause performance to drop the most when masked; models rely on ***content-free hints***
- Robust training with curriculum masking
  - ➢ Masking - randomly hide information to force the model to squeeze more. A tradeoff between:
    - ❖ Masking to increase robustness
    - ❖ Maintaining the required information
  - ➢ Curriculum masking
    - ❖ Slowly ***decays*** the amount of masking that is applied

Experiments

Underline – ground truth; **bold** – R2C's choice
R2C made incorrect choices on the trivially modified options



Q: Where is [2] going ?
A0 [2] is going into the store .
A1 [2] is getting into a carriage .
A2 [1] is going to the bathroom .
A3 [1] is going outside to play after the conversation with [2] is over .

Modified by rule:
A0 He is going into the store .
A1 [2] is getting into a carriage .
A2 [2] is going to the bathroom .
A3 [1] is going outside to play after the conversation with [2] is over .

Modified by an adversarial model:
A0 [MASK] is going into the store .
A1 [2] is getting into a [MASK] .
A2 [MASK] is going to the bathroom .
A3 [1] is [MASK] outside to play after the conversation with [2] is over .

Questions regarding	Count	Avg. Perf. Drop on Q→A	Avg. Perf. Drop on QA→R
E.g., *Where is [2] going ?* (Rule-Singular)	16,154	-5%	-6%
E.g., *What are [1,2] feeling ?* (Rule-Plural)	3,657	-2%	-1%

Token x	p(mask x)	p(mask x| exist x)	Token x	p(mask x)	p(mask x| exist x)
#PERSON	25.71%	27.84%	not	1.29%	24.36%
.	3.82%	3.79%	she	1.20%	12.86%
he	2.53%	12.09%	yes	0.86%	22.47%
is	1.56%	2.78%	the	0.82%	2.97%
they	1.54%	11.70%	a	0.80%	3.06%

Remove a shortcut	Avg. Perf. Drop on Q→A	Avg. Perf. Drop on QA→R
Adv-Top1	-19%	-23%

Method	Q→A			
	Std Val	Rule-Singular	Rule-Plural	AdvTop-1
Baseline	68.5	63.3	65.3	37.0
Masking	69.3	63.9	66.0	48.8
Curriculum Masking	**69.9**	**65.9**	**66.8**	**54.5**