



EMNLP 2020
16th – 20th November



Learning to Contrast the Counterfactual Samples for Robust Visual Question Answering

EMNLP 2020

Zujie Liang, Weitao Jiang, Haifeng Hu, Jiaying Zhu
Sun Yat-Sen University, China







Background – Bias in VQA



EMNLP 2020
16th – 20th November



- Strong superficial linguistic correlations in the training set

Example 1	<p>Train</p> <p>Q+[A] What color is the dog ? [White]</p> <p>Image </p> <p>Training Prior</p> <div><div></div>white <div></div>red <div></div>blue <div></div>green <div></div>yellow <div></div>...</div>	<p>Test</p> <p>Q+[A] What color is the dog ? [Black]</p> <p>Image </p> <p>Models</p> <p>SAN GVQA</p> <p>White Black</p>
	<p>Example 2</p> <p>Q+[A] Is the person wearing shorts ? [No]</p> <p>Image </p> <p>Training Prior</p> <div><div></div>no <div></div>female <div></div>woman <div></div>...</div>	<p>Q+[A] Is the person wearing shorts ? [Yes]</p> <p>Image </p> <p>Models</p> <p>SAN GVQA</p> <p>No Yes</p>

Examples from [Agrawal et al. CVPR 2018]





Background – Bias in VQA



EMNLP 2020
16th – 20th November



- Strong superficial linguistic correlations in the training set
- VQA models easily guess the answer based only on the question

Example 1	<p>Train</p> <p>Q+[A] What color is the dog ? [White]</p> <p>Image </p> <p>Training Prior</p> <div><div></div>white <div></div>red <div></div>blue <div></div>green <div></div>yellow <div></div>...</div>	<p>Test</p> <p>Q+[A] What color is the dog ? [Black]</p> <p>Image </p> <p>Models</p> <p>SAN GVQA</p> <p>White Black</p>
	<p>Example 2</p> <p>Q+[A] Is the person wearing shorts ? [No]</p> <p>Image </p> <p>Training Prior</p> <div><div></div>no <div></div>female <div></div>woman <div></div>...</div>	<p>Q+[A] Is the person wearing shorts ? [Yes]</p> <p>Image </p> <p>Models</p> <p>SAN GVQA</p> <p>No Yes</p>

Examples from [Agrawal et al. CVPR 2018]





Background – Bias in VQA



EMNLP 2020
16th – 20th November



- Strong superficial linguistic correlations in the training set
- VQA models easily guess the answer based only on the question
- Poor robustness and generalization

Example 1	<p>Train</p> <p>Q+[A] What color is the dog ? [White]</p> <p>Image </p> <p>Training Prior</p> <div><div></div>white <div></div>red <div></div>blue <div></div>green <div></div>yellow <div></div>...</div>	<p>Test</p> <p>Q+[A] What color is the dog ? [Black]</p> <p>Image </p> <p>Models</p> <p>SAN GVQA</p> <p>White Black</p>
	<p>Example 2</p> <p>Q+[A] Is the person wearing shorts ? [No]</p> <p>Image </p> <p>Training Prior</p> <div><div></div>no <div></div>female <div></div>woman <div></div>...</div>	<p>Q+[A] Is the person wearing shorts ? [Yes]</p> <p>Image </p> <p>Models</p> <p>SAN GVQA</p> <p>No Yes</p>

Examples from [Agrawal et al. CVPR 2018]




Background – Counterfactual Samples



EMNLP 2020
16th – 20th November

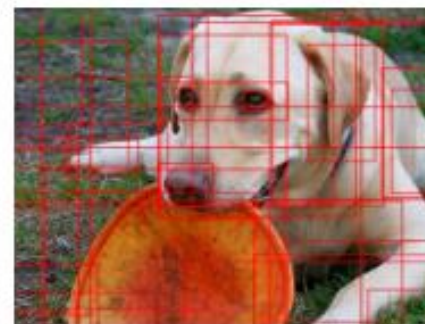


- Augmenting the training data with counterfactual samples for VQA

	Image	Question	Answer	
Original		What color is the man's tie	green	(a)
V-CSS		What color is the man's tie	NOT green	(b)
Q-CSS		What color is the man's [MASK]	NOT green	(c)

Examples from [Chen et al. CVPR 2020]

What color is the frisbee ?

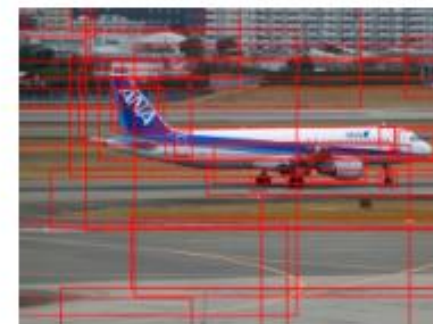


Correct answer(s): orange



Correct answer(s): nil

What is written on the tail ?



Correct answer(s): ANA



Correct answer(s): nil

Examples from [Teney et al. ECCV 2020]

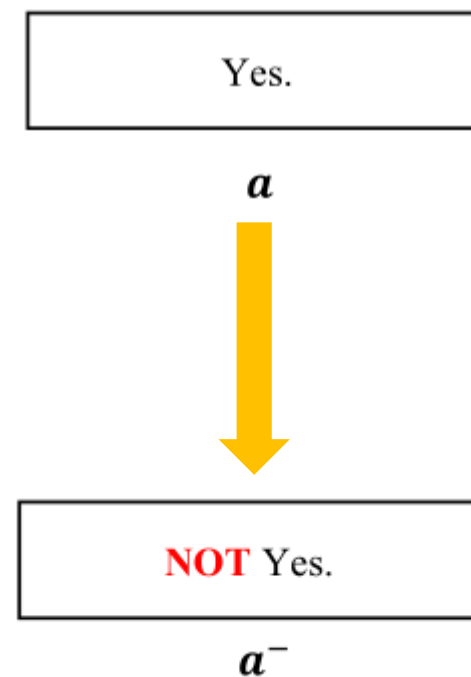
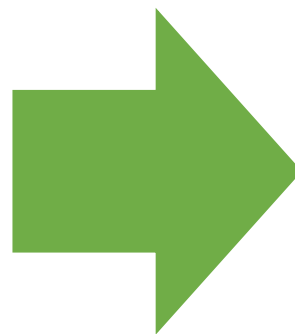
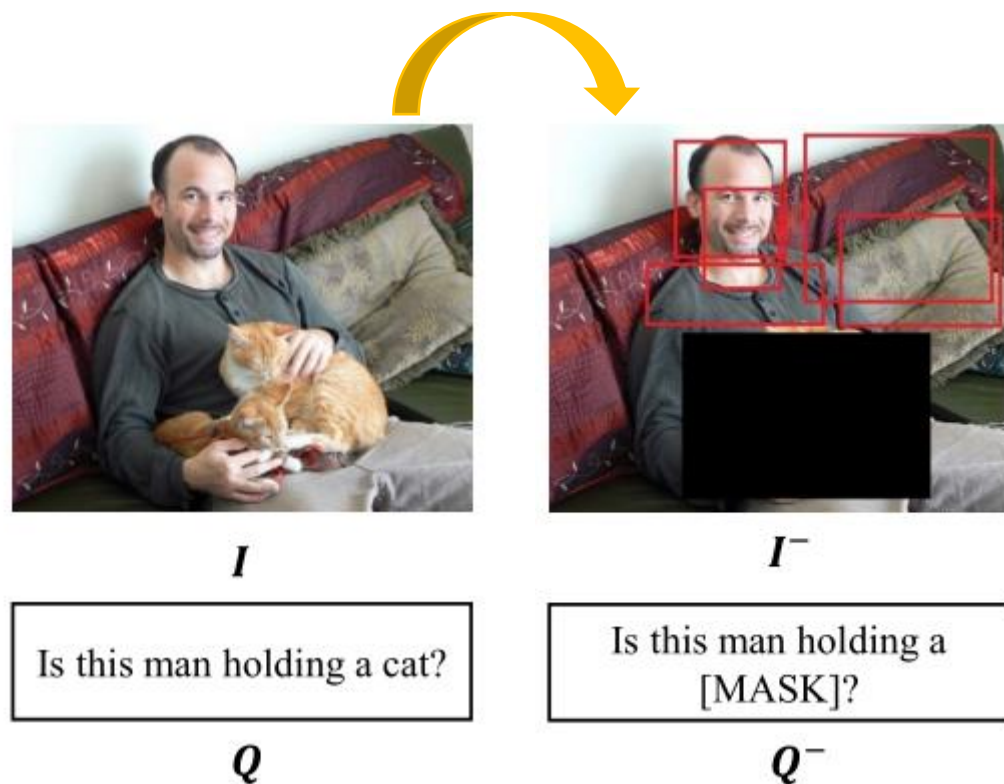
Motivation



EMNLP 2020
16th – 20th November



- Enabling the VQA models to understand: What is the “Cause” for the “Effect”?



“Cause”: the change of input (original -> counterfactual)

“Effect”: the change of answer

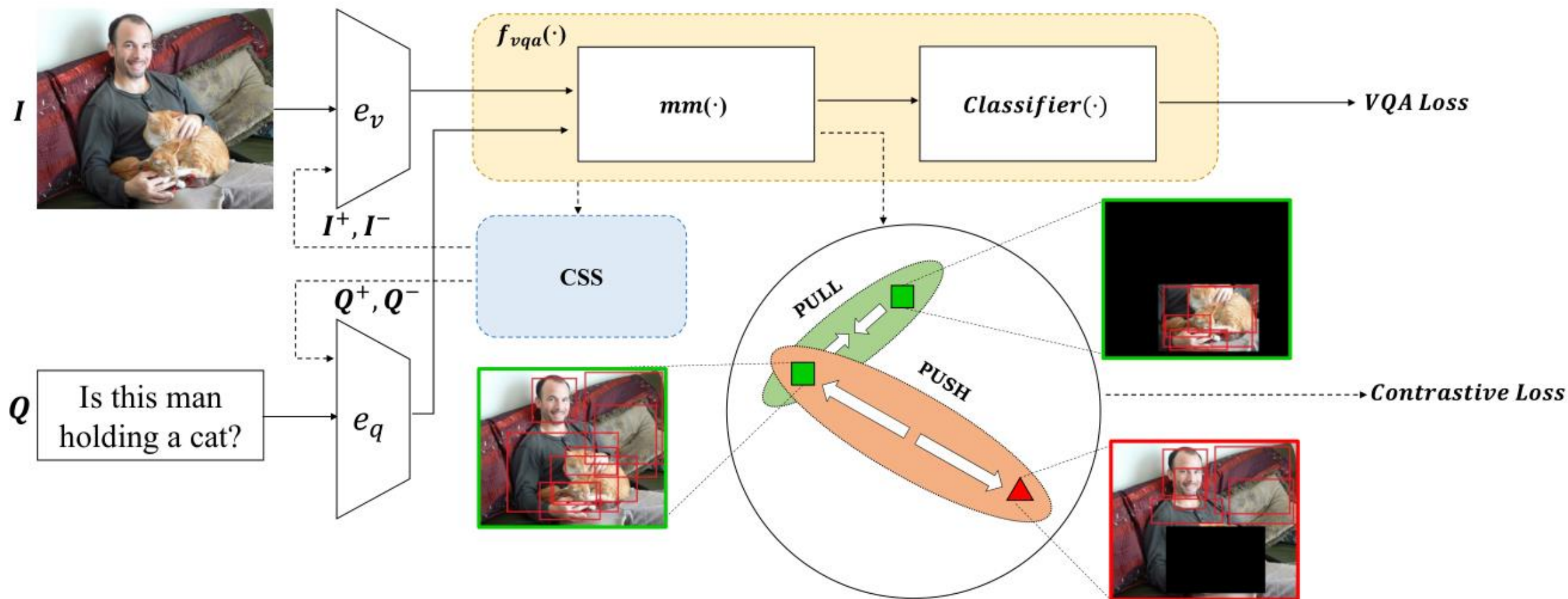
Methodology



EMNLP 2020
16th – 20th November



- Building the causal triplet (I, I^+, I^-) and (Q, Q^+, Q^-)



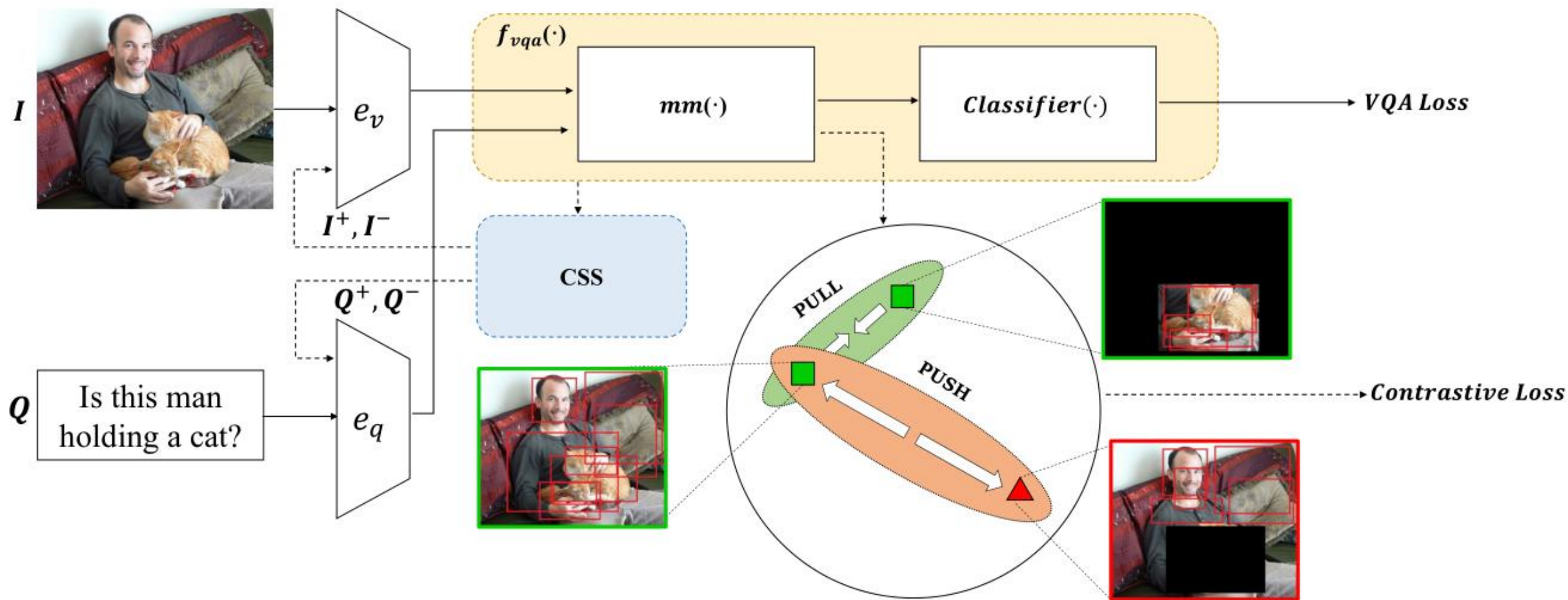
Methodology



EMNLP 2020
16th – 20th November



- Building the causal triplet (I, I^+, I^-) and (Q, Q^+, Q^-)
- Contrastive learning in the multi-modal embedding space



Experiments



EMNLP 2020
16th – 20th November



- State-of-the-art results on VQA-CP v2 and VQA-CP v1

Model	Expl.	VQA-CP v2 <i>test</i>			
		Overall	Y/N	Number	Other
SAN (Yang et al., 2016)		24.96	38.35	11.14	21.74
GVQA (Agrawal et al., 2018)		31.30	57.99	13.68	22.14
Unshuffling (Teney et al., 2020b)		42.39	47.72	14.43	47.24
+CF (Teney et al., 2020a)	HAT	46.00	61.30	15.60	46.00
+CF+GS (Teney et al., 2020a)	HAT	46.80	64.50	15.30	45.90
UpDn (Anderson et al., 2018)		39.74	42.27	11.93	46.05
+AReg (Ramakrishnan et al., 2018)		41.17	65.49	15.48	35.48
+GRL (Grand and Belinkov, 2019)		42.33	59.74	14.78	40.76
+RUBi (Cadene et al., 2019b)		44.23	67.05	17.48	39.61
+LMH (Clark et al., 2019)		52.01	72.58	31.12	46.97
+LMH+CSS* (Chen et al., 2020)		57.74	83.18	47.59	47.19
+LMH+CSS+GS* (Teney et al., 2020a)		57.37	79.71	50.85	47.45
+LMH+CSS+CL(ours)		59.18	86.99	49.89	47.16
+HINT (Selvaraju et al., 2019)	HAT	47.70	70.04	10.68	46.31
+SCR (Wu and Mooney, 2019)	HAT	49.17	71.55	10.72	47.49

Table 1: Performance (%) comparison with SoTA on VQA-CP v2 dataset. *indicates the results of our reimplementation. Expl. denotes the extra annotations that the model has used. HAT is the human attention (Das et al., 2016).

Model	VQA-CP v1 <i>test</i>			
	Overall	Y/N	Number	Other
UpDn (Anderson et al., 2018)	37.87	42.58	14.16	42.71
+AReg (Ramakrishnan et al., 2018)	45.69	77.64	13.21	26.97
+GRL (Grand and Belinkov, 2019)	44.09	75.01	13.40	25.67
+RUBi (Cadene et al., 2019b)	44.81	69.65	14.91	32.13
+LMH (Clark et al., 2019)	55.27	76.47	26.66	45.68
+LMH+CSS* (Chen et al., 2020)	59.63	86.62	28.93	45.12
+LMH+CSS+GS* (Teney et al., 2020a)	58.05	78.50	37.24	46.08
+LMH+CSS+CL(ours)	61.27	88.14	34.43	45.34

Table 2: Performance comparison on VQA-CP v1 *test*. *indicates the results of our reimplementation.

Experiments



EMNLP 2020
16th – 20th November



- Our method works with different architecture

Model	Overall	Y/N	Number	Other
UpDn (Anderson et al., 2018)	39.74	42.27	11.93	46.05
UpDn*	38.85	42.60	11.51	44.38
+CSS*	39.77	42.80	12.55	45.66
+CSS+GS*	40.02	41.97	11.94	46.70
+CSS+MarginCL(ours)	40.15	42.38	12.45	46.57
+CSS+CL(ours)	40.49	42.90	12.44	46.93
LMH (Clark et al., 2019)	52.01	72.58	31.12	46.97
LMH*	52.66	73.47	34.21	46.81
+CSS*	57.74	83.18	47.59	47.19
+CSS+GS*	57.37	79.71	50.85	47.45
+CSS+MarginCL(ours)	58.68	85.54	51.60	46.54
+CSS+CL(ours)	59.18	86.99	49.89	47.16

Table 3: Effectiveness of different supervision of counterfactual samples on different architectures on VQA-CP v2 test. *indicates the results of our reimplementation.

Experiments



EMNLP 2020
16th – 20th November



- Our method generalizes better on the counterfactual samples and factual samples

Model	Original Samples	Factual Samples	Counterfactual Samples
CSS	57.74	46.41	48.96
CSS+GS*	57.37	45.83	50.09
CSS+CL(ours)	59.18	46.73	50.12

Table 4: The VQA performance (%) of the counterfactual samples and factual samples on VQA-CP v2 dataset. *indicates the results of our reimplementation.



EMNLP 2020
16th – 20th November



Thanks for listening!

