Improving Users' Mental Model with Attention-directed Counterfactual Edits

Kamran Alipour[1], Arijit Ray[2], Xiao Lin[2], Michael Cogswell[2], Jurgen Schulze[1], Yi Yao[2], and Giedrius Burachas[2]

[1]University of California San Diego
[2]SRI International

June 25, 2021

Abstract

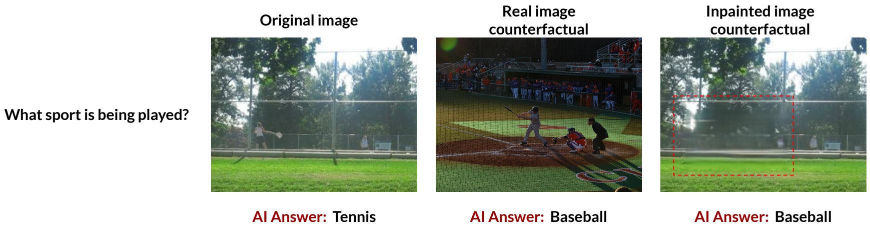
In the domain of Visual Question Answering (VQA), studies have shown improvement in users' mental model of the VQA system when they are exposed to examples of how these systems answer certain Image-Question (IQ) pairs. In this work, we show that showing controlled counterfactual image-question examples are more effective at improving the mental model of users as compared to simply showing random examples. We compare a generative approach and a retrieval-based approach to show counterfactual examples. We use recent advances in generative adversarial networks (GANs) to generate counterfactual images by deleting and inpainting certain regions of interest in the image. We then expose users to changes in the VQA system's answer on those altered images. To select the region of interest for inpainting, we experiment with using both human-annotated attention maps and a fully automatic method that uses the VQA system's attention values. Finally, we test the user's mental model by asking them to predict the model's performance on a test counterfactual image. We note an overall improvement in users' accuracy to predict answer change when shown counterfactual explanations. While realistic retrieved counterfactuals obviously are the most effective at improving the mental model, we show that a generative approach can also be equally effective.

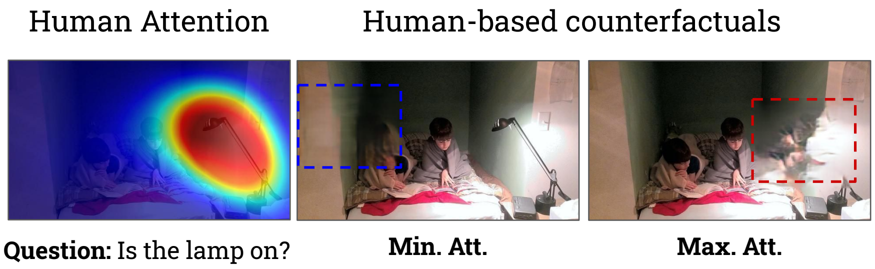
Hosted file

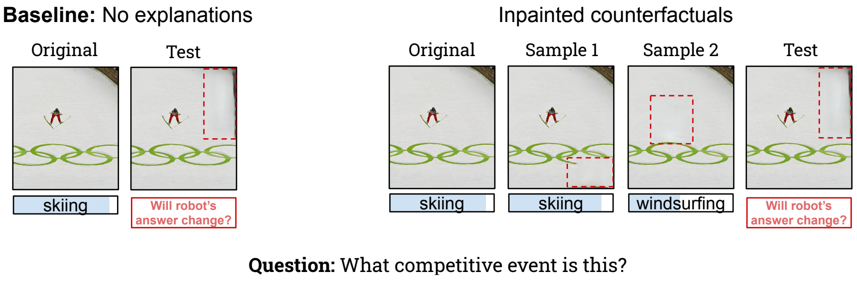
`main_text.zip` available at https://authorea.com/users/422041/articles/527829-improving-users-mental-model-with-attention-directed-counterfactual-edits

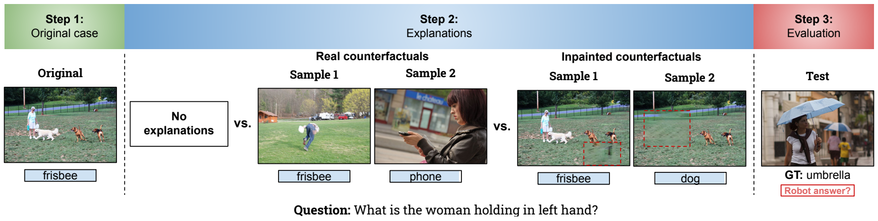
Hosted file

`main.pdf` available at https://authorea.com/users/422041/articles/527829-improving-users-mental-model-with-attention-directed-counterfactual-edits


Original image
Real image counterfactual
Inpainted image counterfactual
What sport is being played?
AI Answer: Tennis
AI Answer: Baseball
AI Answer: Baseball


Human Attention
Human-based counterfactuals
Question: Is the lamp on?
Min. Att.
Max. Att.


Baseline: No explanations
Original
Test
skiing
Will robot's answer change?
Inpainted counterfactuals
Original
Sample 1
Sample 2
Test
skiing
skiing
windsurfing
Will robot's answer change?
Question: What competitive event is this?


Step 1:
Original case
Step 2:
Explanations
Step 3:
Evaluation
Original
frisbee
No explanations
vs.
Real counterfactuals
Sample 1
Sample 2
frisbee
phone
vs.
Inpainted counterfactuals
Sample 1
Sample 2
frisbee
dog
Test
GT: umbrella
Robot answer?
Question: What is the woman holding in left hand?