Is CLIP Fooled by Optical Illusions?

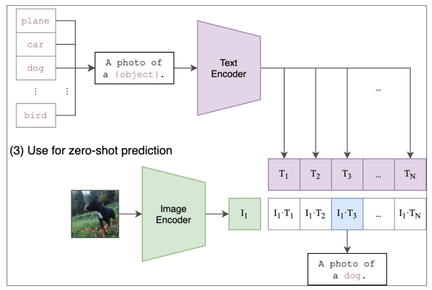
Jerry Ngo, Swami Sankaranarayanan, Phillip Isola
MIT CSAIL

Introduction

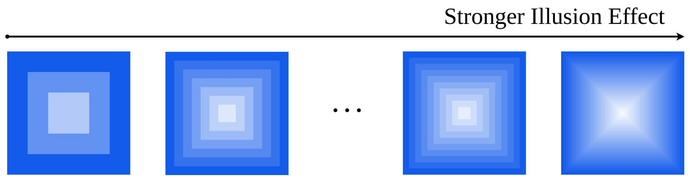
- Recent big machine learning shows impressive generalization performance for various perception tasks
- Optical illusions are the product of human biology, learning, and perception
- **Do machine learning models get fooled by optical illusion?**

Methods

- We measure the effect on the CLIP models

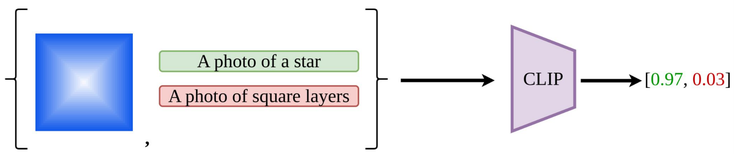


by using images of illusions with different illusion strengths



and classifying them into illusory or non-illusory prompts.

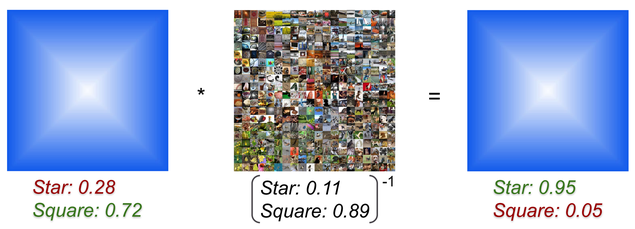
"A photo of a star" - Illusory prompt
"A photo of square layers" - Non-illusory prompt



- We utilized ChatGPT to generate different pair of prompts for each illusion:

"Below are annotations for the same image. There are two sets of annotations. The first set of annotation describes the image, and the second one doesn't, which contrast the first annotation, so we can use them for the classification task:"

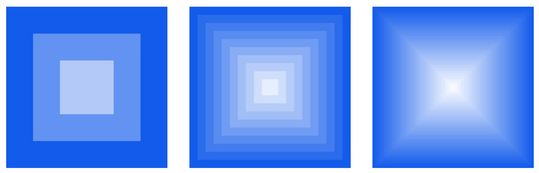
- We calibrate the association by using content-free probabilities



Results

Edge Detection Illusion

Vasarely Illusion

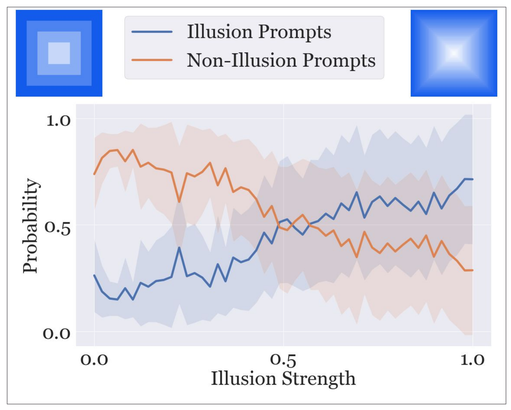


3 squares *10 squares* *50 squares*

- Impose squares with different luminance level
- An invisible X shape gradually appear

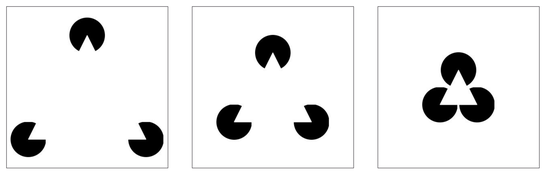
Examples of Illusory prompts
- An image of a bright star on a blue background…
- A photo of a blue background with a prominent star…

Examples of non-illusory prompts
- A picture of a stack of boxes in varying shades of blue
- An image of blue boxes arranged in a stack…



Gestalt Closure

Kanizsa Triangle

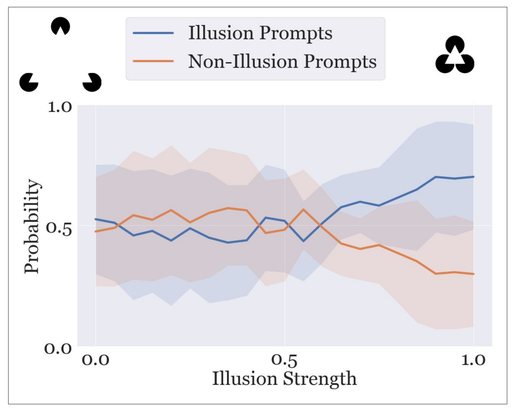


300 pixels *200 pixels* *100 pixels*

- Place three sliced circles in a triangular formation
- Illusory contours evoke the perception of a triangle

Examples of Illusory prompts
- … three circles being obscured by a white triangle
- … three sliced black circles making up a triangle…

Examples of non-illusory prompts
- A photo of three circles equally spaced from each other
- A photo of three circles, one on top, two at the bottom



Lightness Illusion

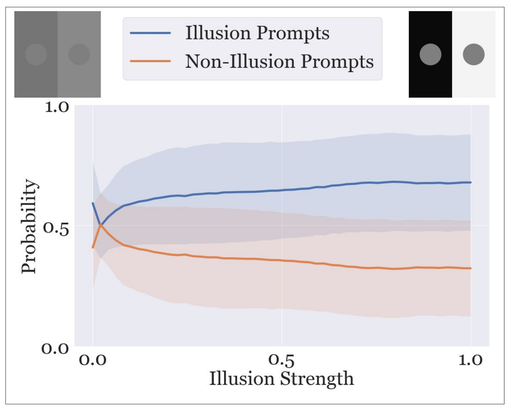
Simultaneous Contrast Illusion

2 lightness (away) *32 lightness* *100 lightness*

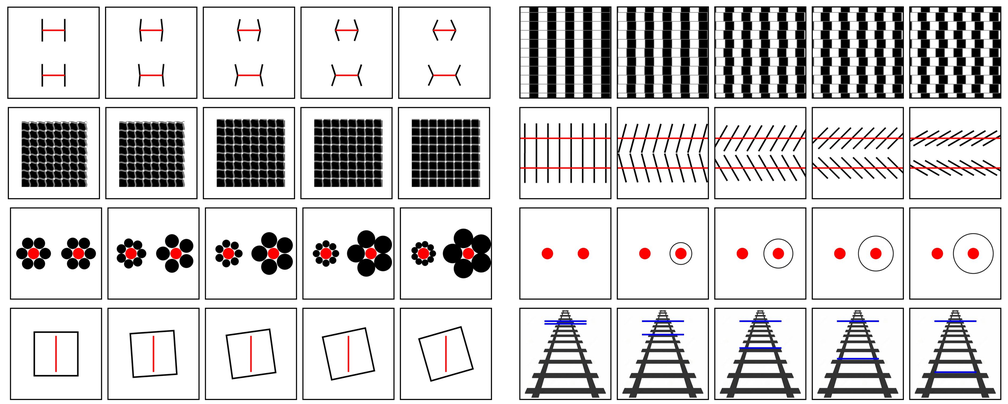
- Two circles have the same level of luminance
- The left circle seems lighter than the right one

Examples of Illusory prompts
- A photo of two circles on a gray background, one circle is lighter/darker than the other

Examples of non-illusory prompts
- A photo of two gray circles of the same size and same shade of gray



Other Illusions that Don't Fool CLIP



Conclusion

- A variety of human optical illusions fools CLIP
- This shows similarities between human biology/perception and machine learning
- This also poses potential vulnerabilities to these models
- Future work on how human associates those prompts which illusion and why CLIP sees these illusion is needed

Acknowledgement

We appreciate Professor Yoon Kim's suggestion of using the calibration method to debias CLIP. He also helps adapt the original language model calibration to image-text calibration.