Gridspace

GRIDSPACE IAP 2024 LECTURE 5
CAN ML MODELS PERCEIVE?

January 8, 2023

Exercises:

- Mistral-7B-v0.1 uses sliding window attention. What's its window length & number of layers? What's its theoretical attention span? Ans: W=4096, L=32, theoretical span 128k
- Can you suggest some reasons that the attention span in practice is lower than theoretical span? How to remedy? Ans: model was not actually exposed to such long texts during training. Train with longer texts may help
- What are some of the pain points of depending on third-party search engines? Ans: $$$, they may queue us/impose quota, their server may not be available, their updates may cause unexpected change on our side

Perception

noun

per·cep·tion /pərˈsepSH(ə)n/

Perception

noun

per·cep·tion /pərˈsepSH(ə)n/

1. *the ability to see, hear, or become aware of something through the senses.*

Visual Perception

noun

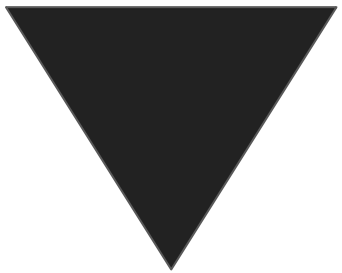
vis·u·al per·cep·tion /ˈviZHəlˈpərˈsepSH(ə)n/

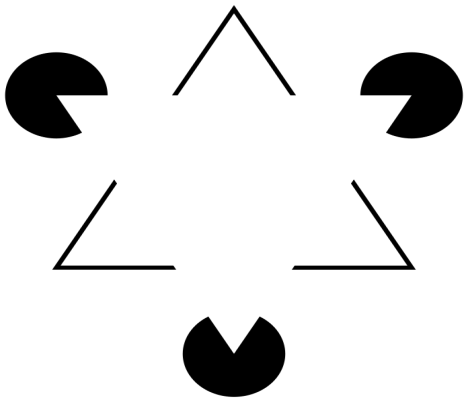
Visual Perception

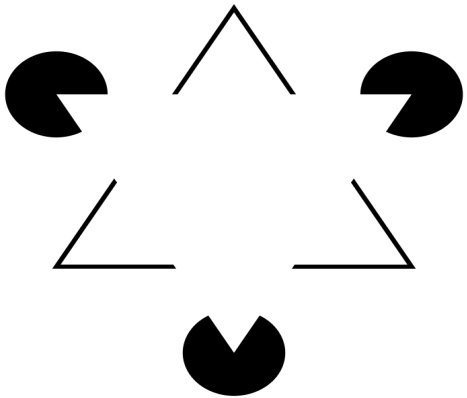
noun

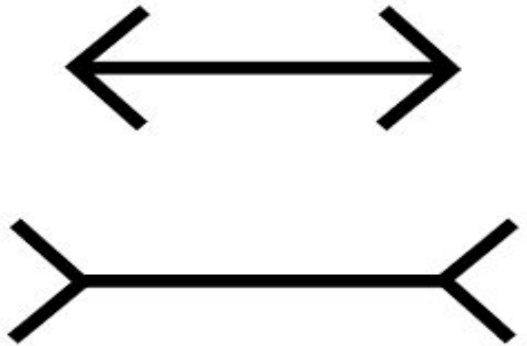
vis·u·al per·cep·tion /ˈviZHəlˈpərˈsepSH(ə)n/

1. *the ability to see, through the sensation of light*









Visual Perception

noun

vis·u·al per·cep·tion /ˈviZHəlˈpərˈsepSH(ə)n/

1. *the ability to see, through the sensation of light*

Visual Perception

noun

vis·u·al per·cep·tion /ˈviZHəlˈpərˈsepSH(ə)n/

~~1. the ability to see, through the sensation of light.~~

2. *the organization, identification, and interpretation of light*

Perception

noun

per·cep·tion /pərˈsepSH(ə)n/

1. *the ability to see, hear, or become aware of something through the senses.*

Perception

noun

per·cep·tion /pərˈsepSH(ə)n/

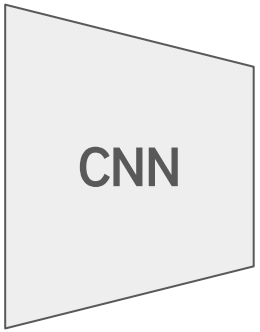
~~1. *the ability to see, hear, or become aware of something through the senses.*~~

2. *the organization, identification, and interpretation of sensory information*

Can ML models perceive?

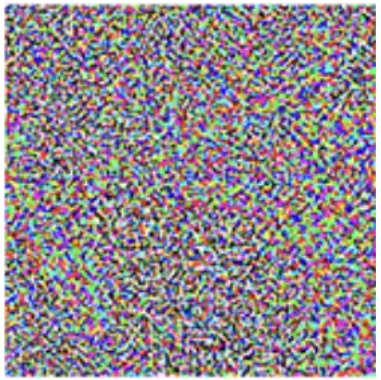
Can ML models organize and interpret information?

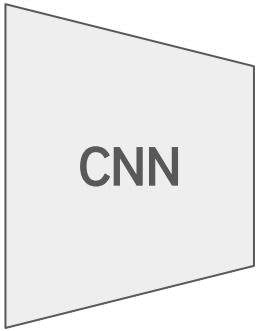
Can ML models organize and interpret information?



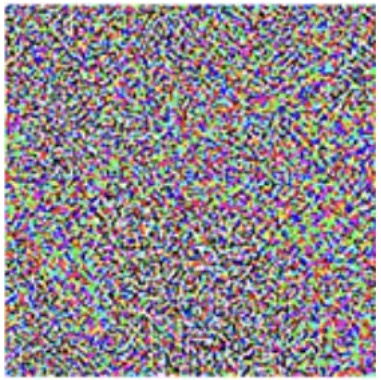
"Cat"

Can ML models organize and interpret information?





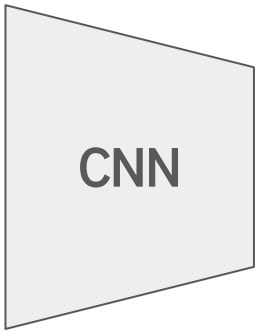
Can ML models organize and interpret information?



CNN

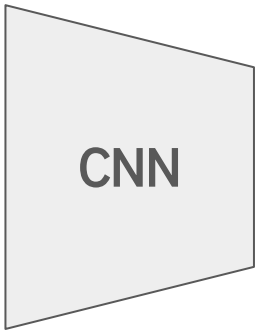
"Avocado"

Can ML models organize and interpret information?



“Avocado”

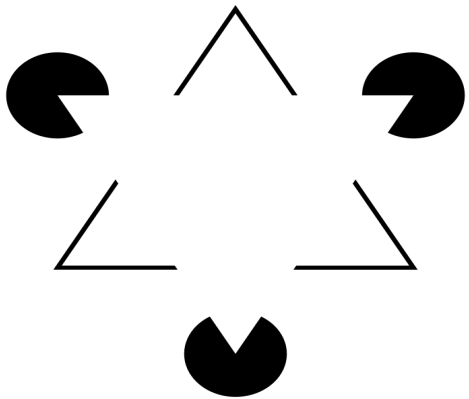
Can ML models organize and interpret information like humans?

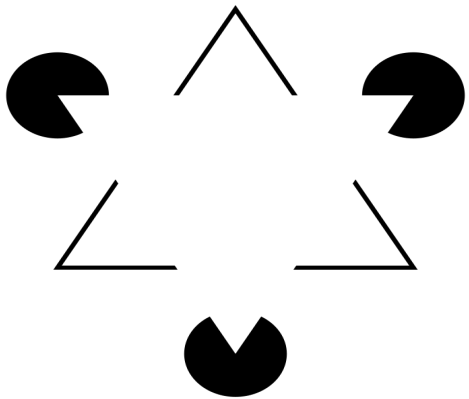


“Avocado”

Overview

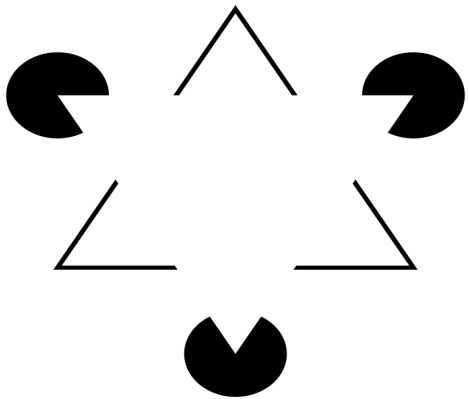
- Do ML models perceive like humans?
- Approach 1: Naive approach
- Approach 2: Psychological approach
- Approach 3: Neuroscience approach
- Improving methods
- Op-ed





Is CLIP Fooled by Optical Illusions?

Jerry Ngo, Swami Sankaranarayanan, & Phillip Isola. (2023).

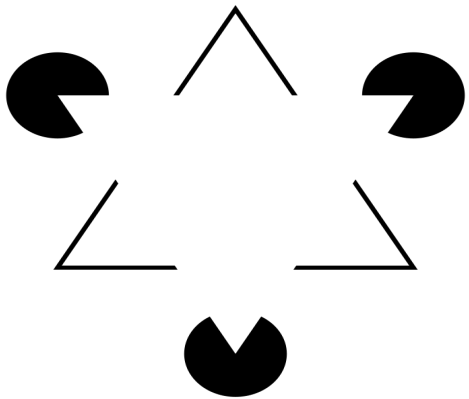


Is CLIP Fooled by Optical Illusions?

Jerry Ngo, Swami Sankaranarayanan, & Phillip Isola. (2023).

Neural Networks Trained on Natural Scenes Exhibit Gestalt Closure

Been Kim, Emily Reif, Martin Wattenberg, Samy Bengio, & Michael C. Mozer. (2020).



Is CLIP Fooled by Optical Illusions?

Jerry Ngo, Swami Sankaranarayanan, & Phillip Isola. (2023).

Neural Networks Trained on Natural Scenes Exhibit Gestalt Closure

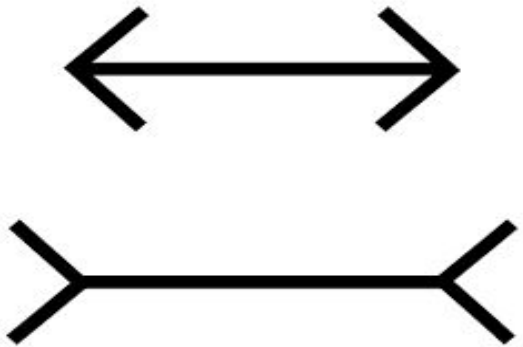
Been Kim, Emily Reif, Martin Wattenberg, Samy Bengio, & Michael C. Mozer. (2020).

Deep convolutional networks do not perceive illusory contours

Baker, N., Kellman, P.J., Erlikhman, G., & Lu, H. (2018).

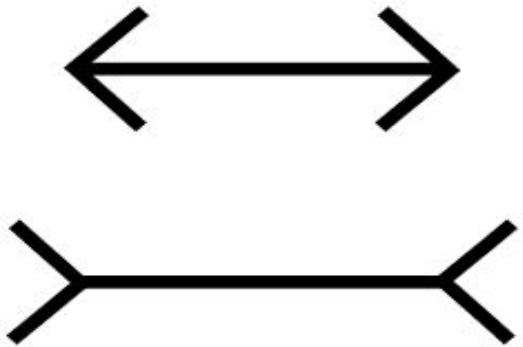
Ask the model! (gemini-pro)

0 : 0



Which line appears longer?

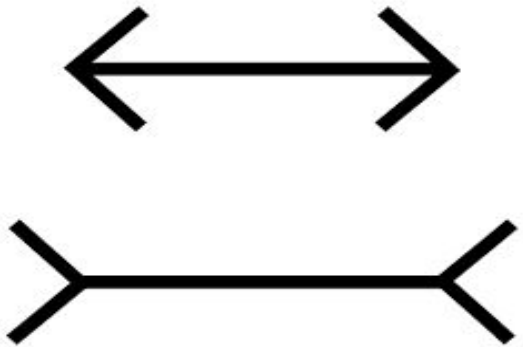
1 : 0



Which line appears longer?

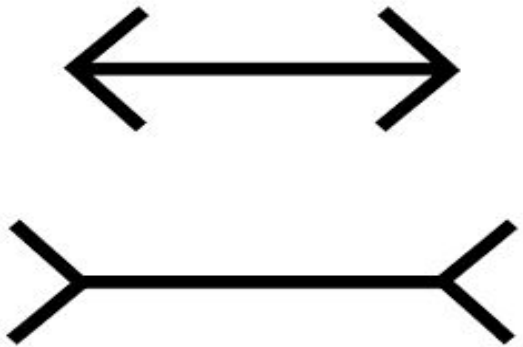
Both lines are the same length.

1 : 0



What is the name of this image?

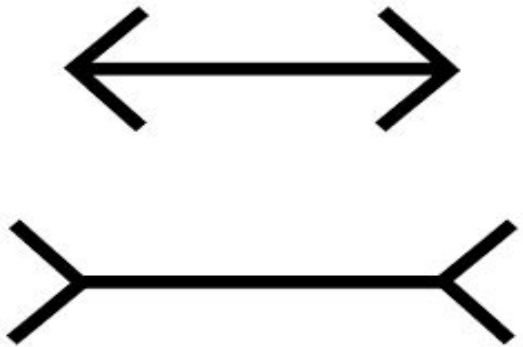
1 : 0



What is the name of this image?

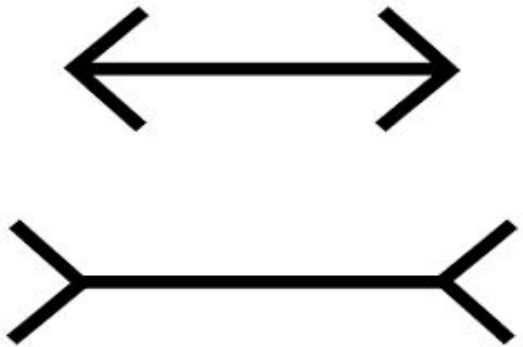
The name of this image is the Muller-Lyer illusion.

1 : 0



What is perceived in the image?

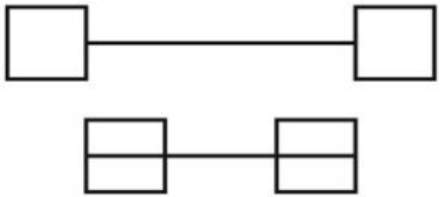
1 : 1



What is perceived in the image?

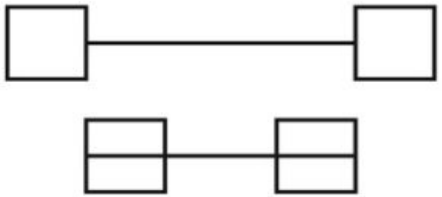
The upper arrow appears longer than the lower arrow. However, they are the same length.

1 : 1



What is perceived in the image?

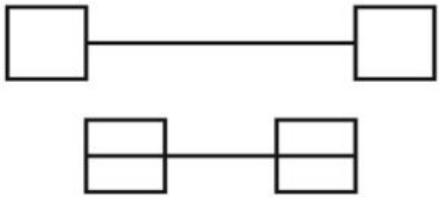
1 : 2



What is perceived in the image?

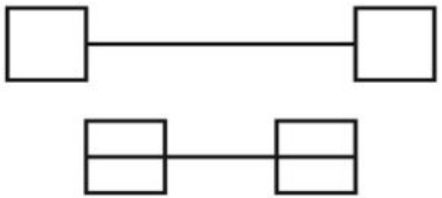
Two parallel lines of different lengths.

1 : 2



Which line is longer?

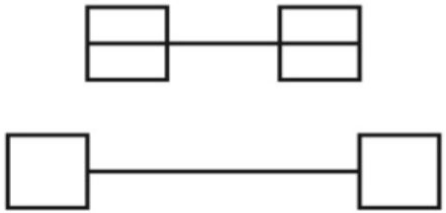
1 : 3



Which line is longer?

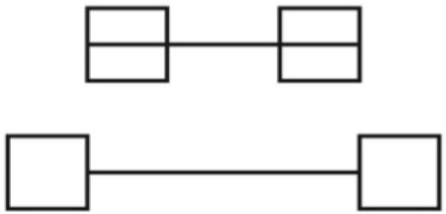
The top line appears longer.

1 : 3



Which line is longer?

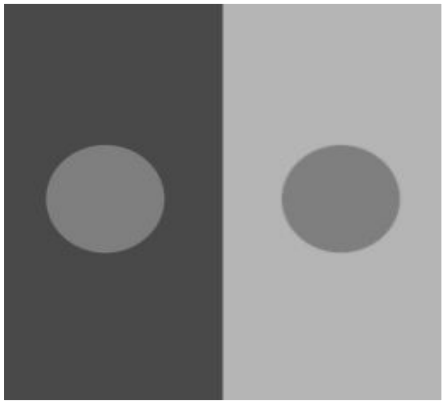
2 : 3



Which line is longer?

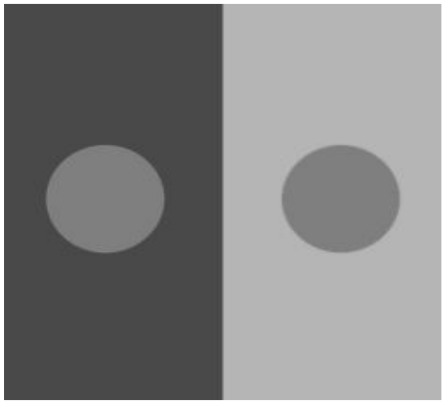
The top line appears longer.

2 : 3



Which dot appears darker in the image?

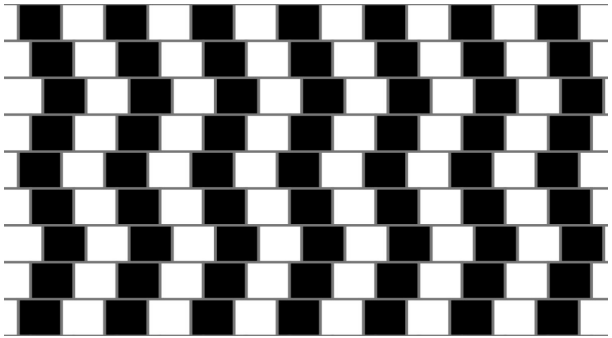
3 : 3



Which dot appears darker in the image?

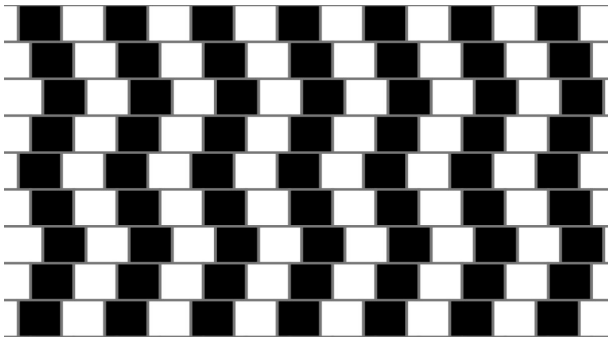
Both the dots are of the same shade.

3 : 3



What is perceived in the image?

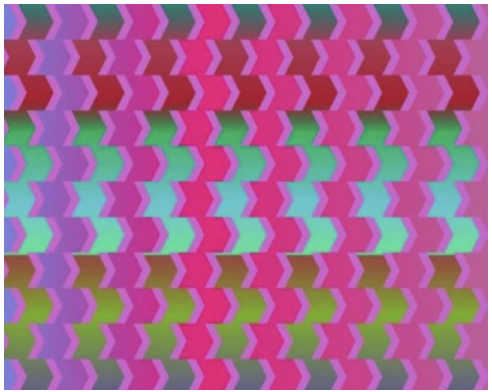
3 : 4



What is perceived in the image?

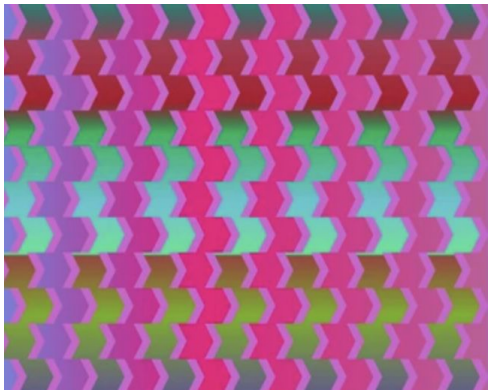
The perceived image is that the horizontal lines are not parallel.

3 : 4



Are the horizontal lines parallel?

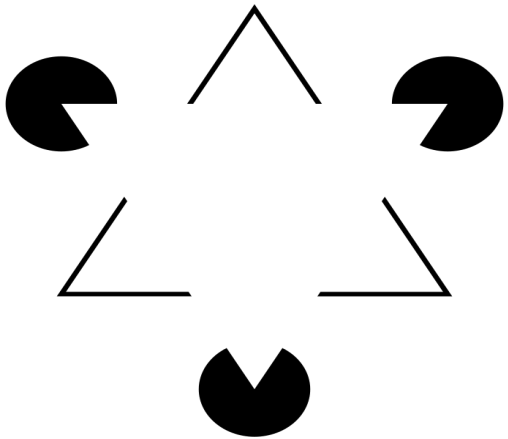
4 : 4



Are the horizontal lines parallel?

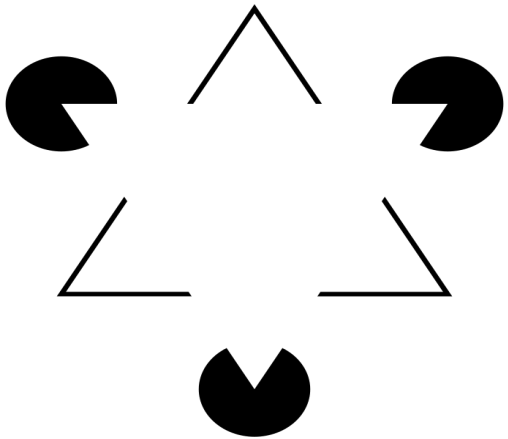
Yes. The horizontal lines are parallel.

4 : 4



What is perceived in the image?

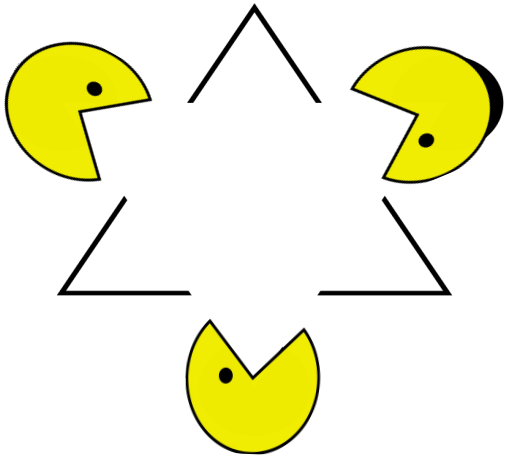
5 : 4



What is perceived in the image?

There are three faces. Each face has two eyes and a mouth.

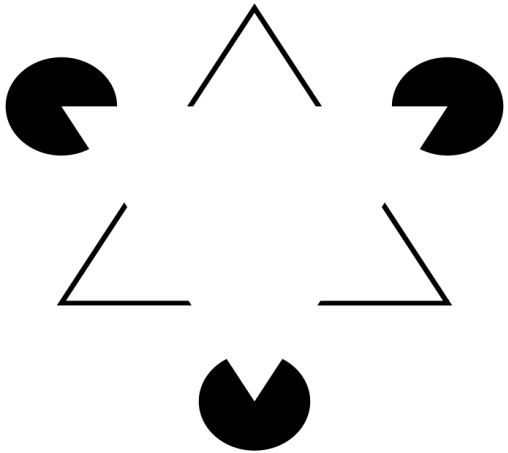
5 : 4



What is perceived in the image?

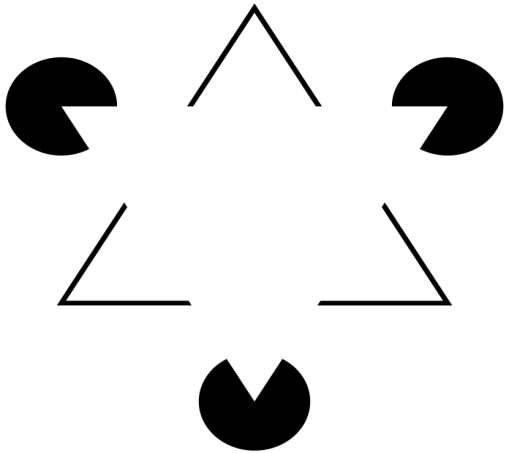
There are three faces. Each face has two eyes and a mouth.

5 : 4



Can you see any triangles in the picture?

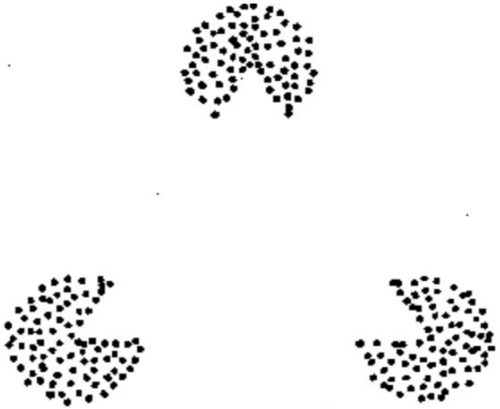
6 : 4



Can you see any triangles in the picture?

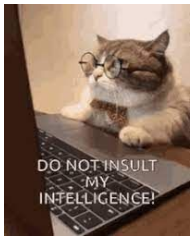
There are no triangles in the picture. It's an optical illusion. The three shapes in the picture are actually circles.

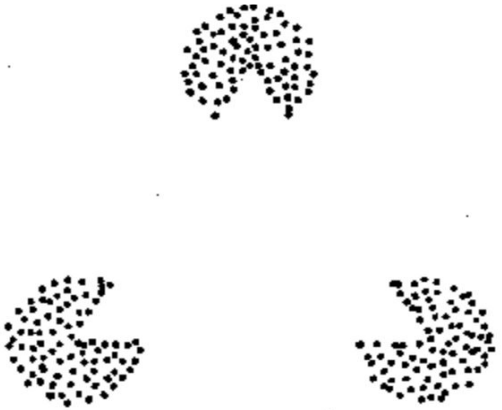
6 : 4



What is perceived in the image?

7 : 4

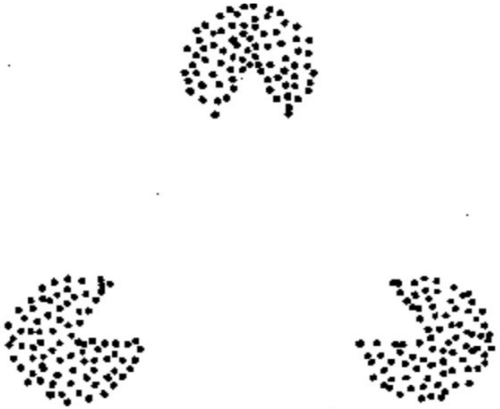




What is perceived in the image?

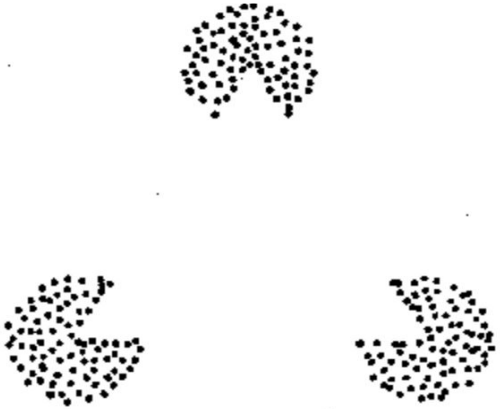
The percept is of three overlapping mushrooms.

7 : 4



Can you see any triangles in the picture?

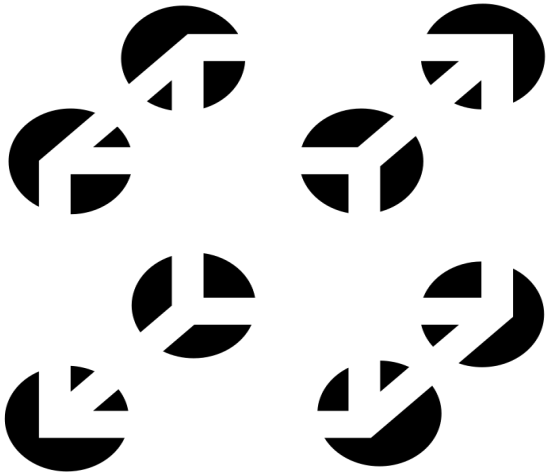
8 : 4



Can you see any triangles in the picture?

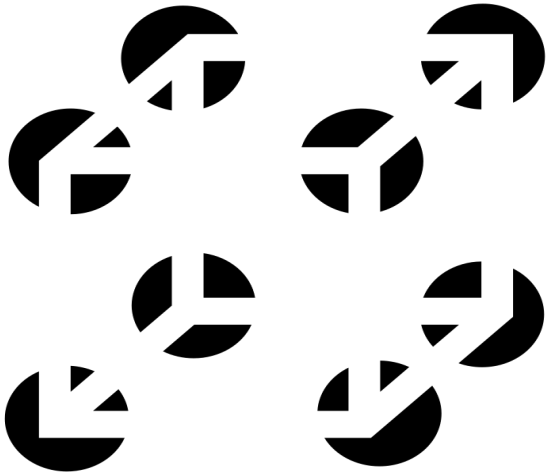
Yes. You can see three triangles.

8 : 4



What is perceived in the image?

8 : 5



What is perceived in the image?

The image is perceived as a Necker cube. This is a 3D optical illusion cube that appears to change its orientation when viewed.

8 : 5

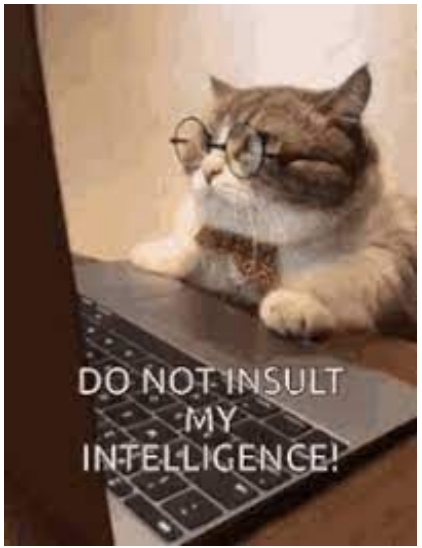


Image augmentation

8 : 5

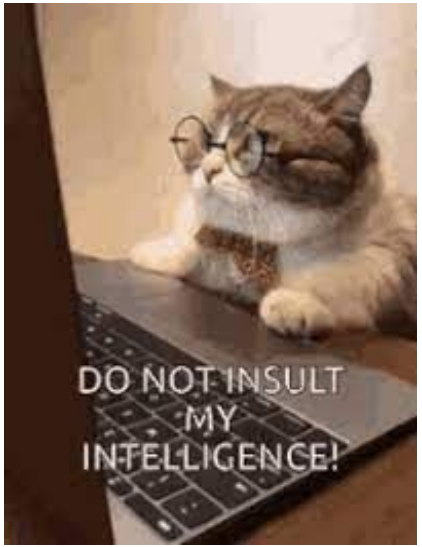
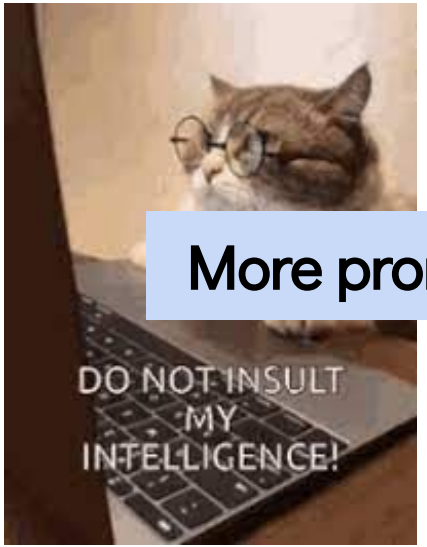


Image augmentation

8 : 5

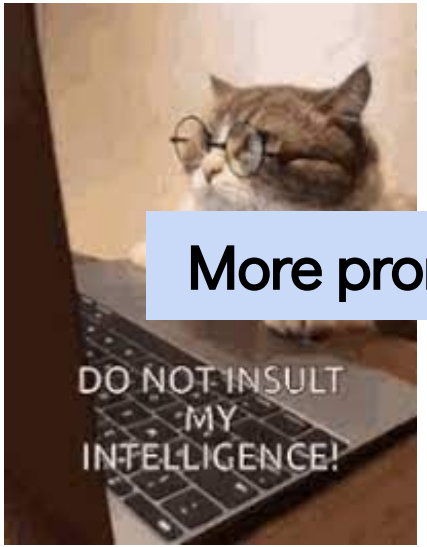


More prompts

Image augmentation

More models

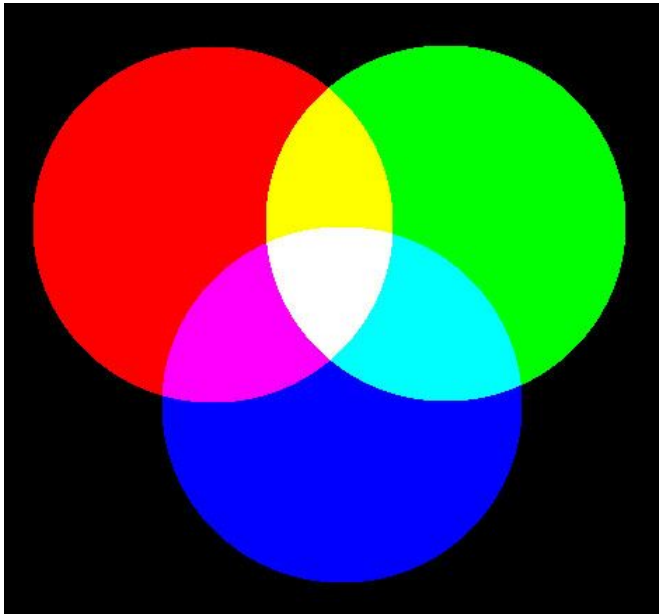
8 : 5



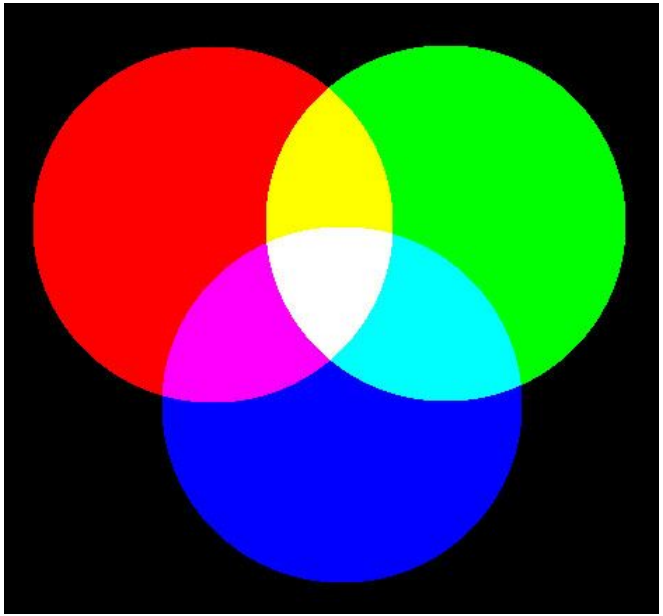
More prompts

Overview

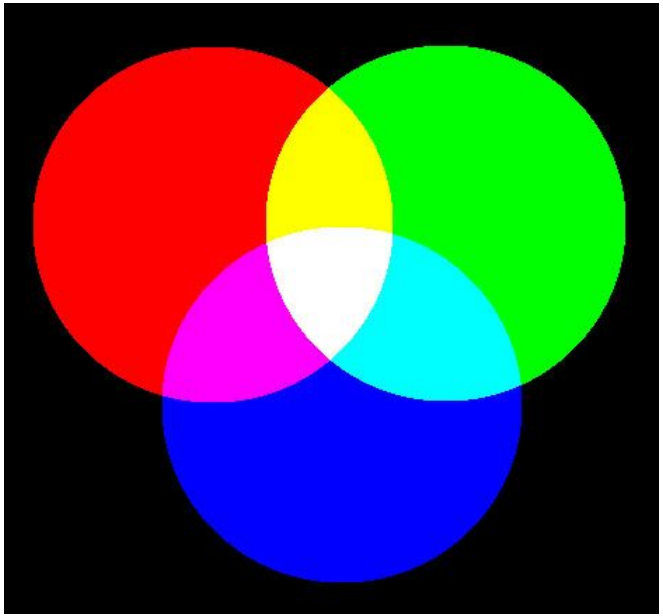
- Do ML models perceive like humans?
- Naive approach: Optical Illusions
- Approach 2: Psychological approach
- Approach 3: Neuroscience approach
- Improving methods
- Op-ed

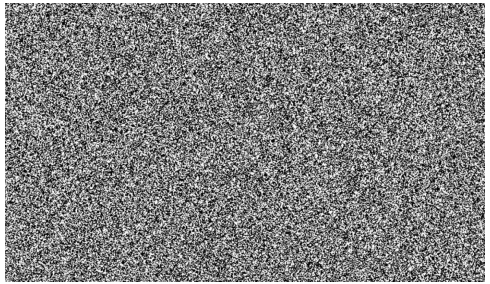


Human metamers

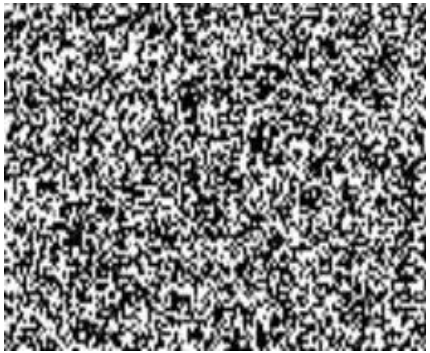


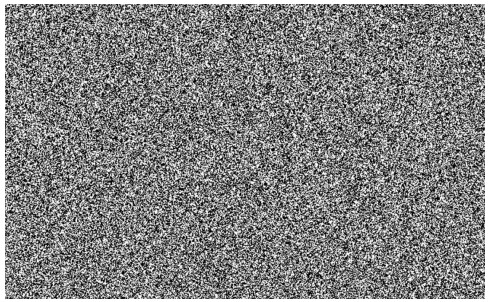
Human metamers

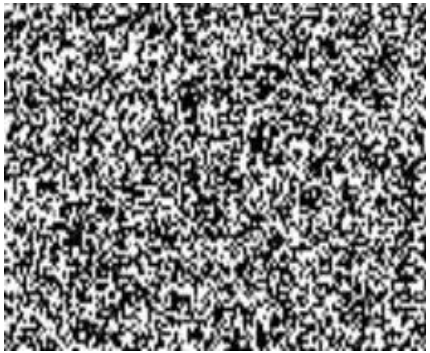


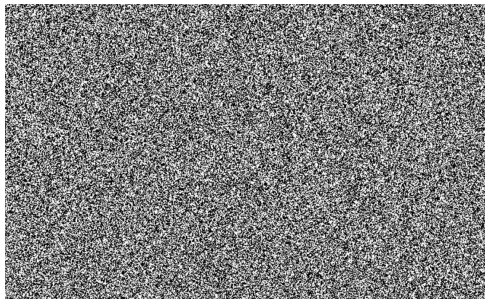


Human metamers

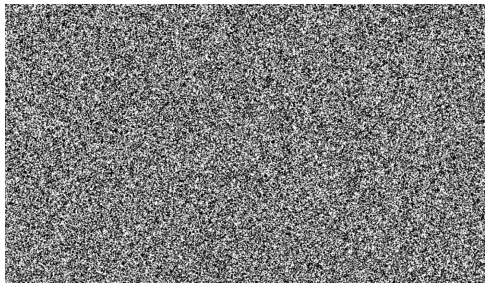


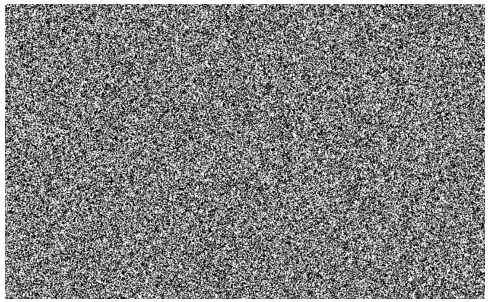




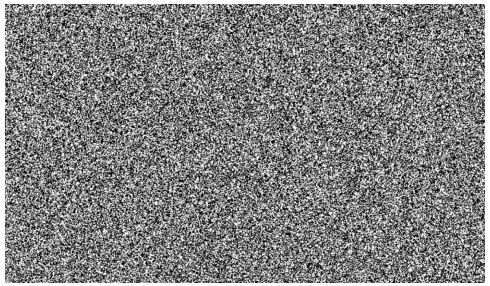


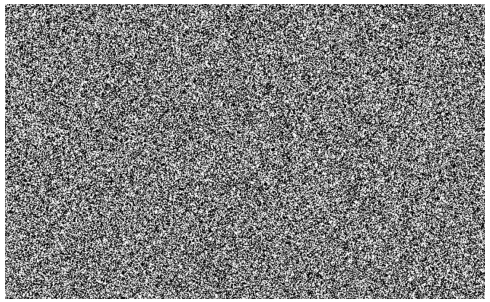
“white noise”





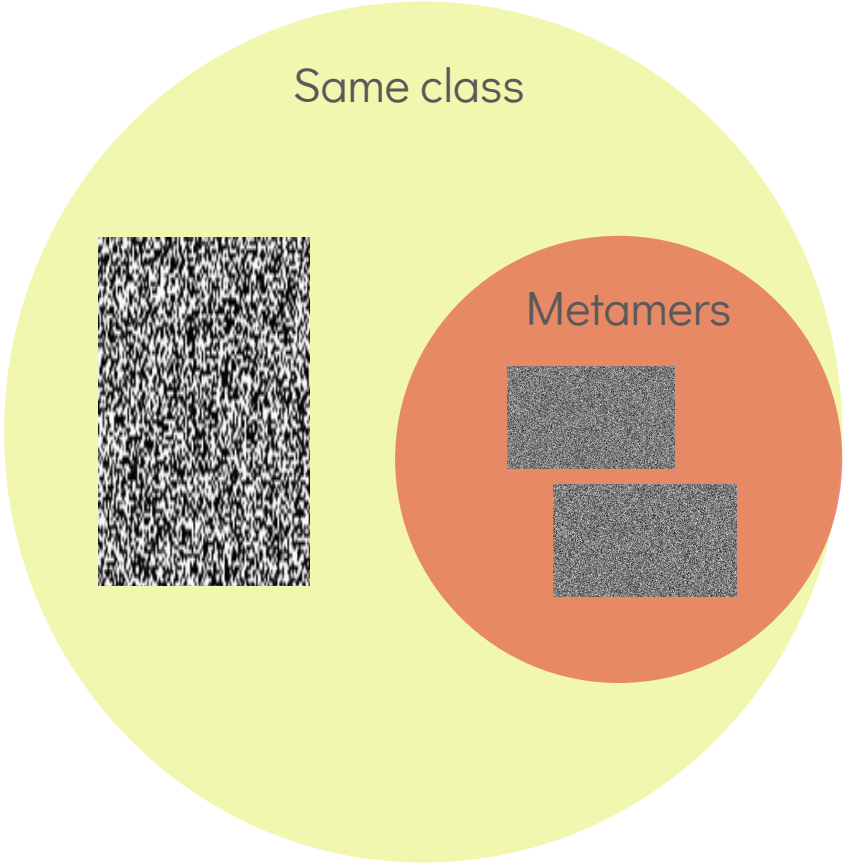
“white noise”

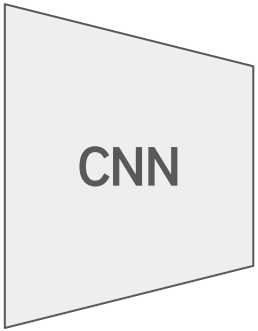




“white noise”

Human metamers

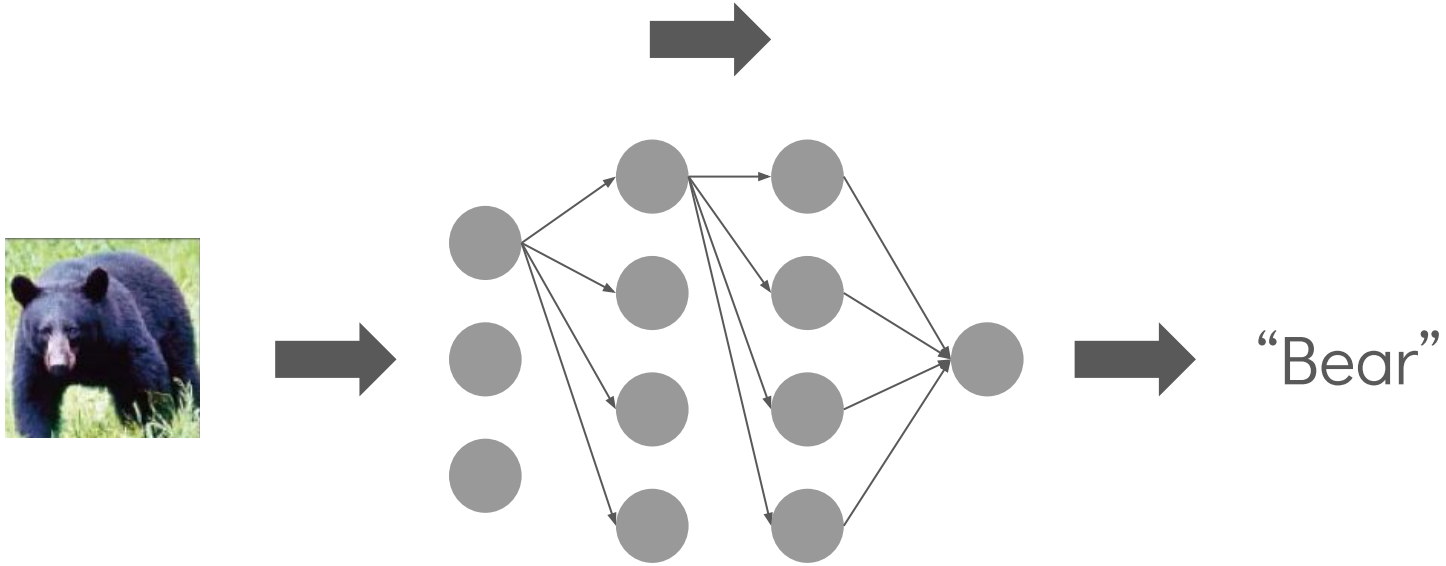

Same class
Metamers



“Cat”

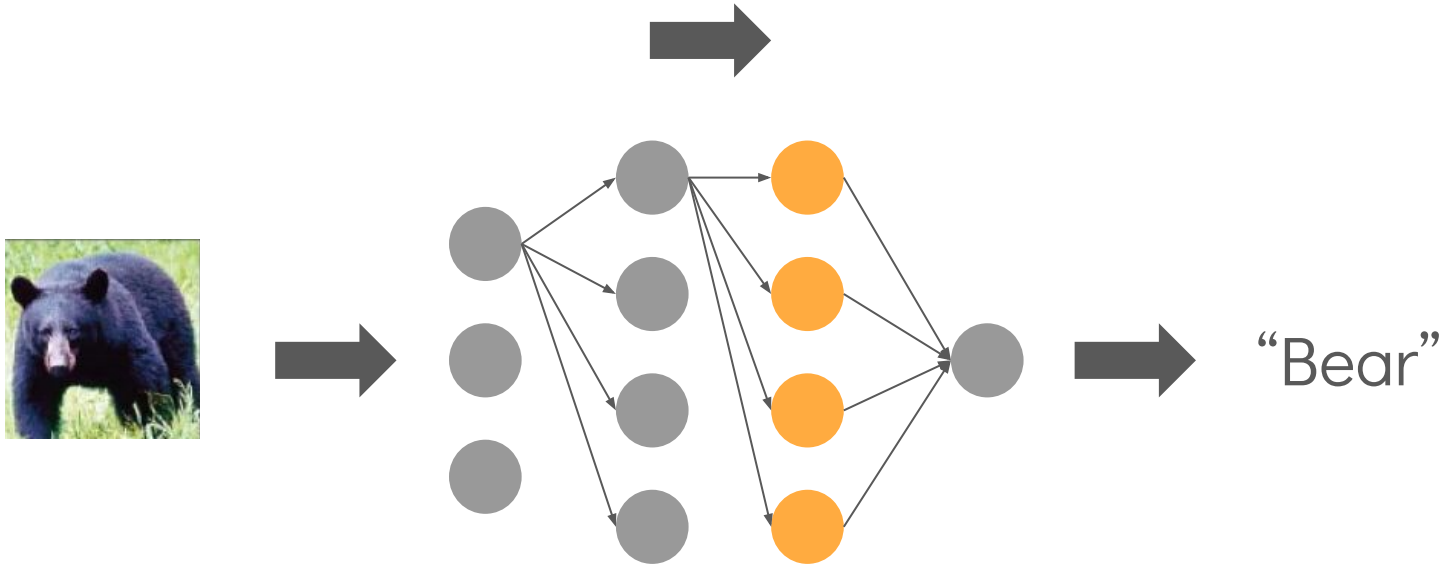
Model metamers illuminate divergences between biological and artificial neural networks

Jenelle Feather, Guillaume Leclerc, Aleksander Madry, & Josh H. McDermott (2023)



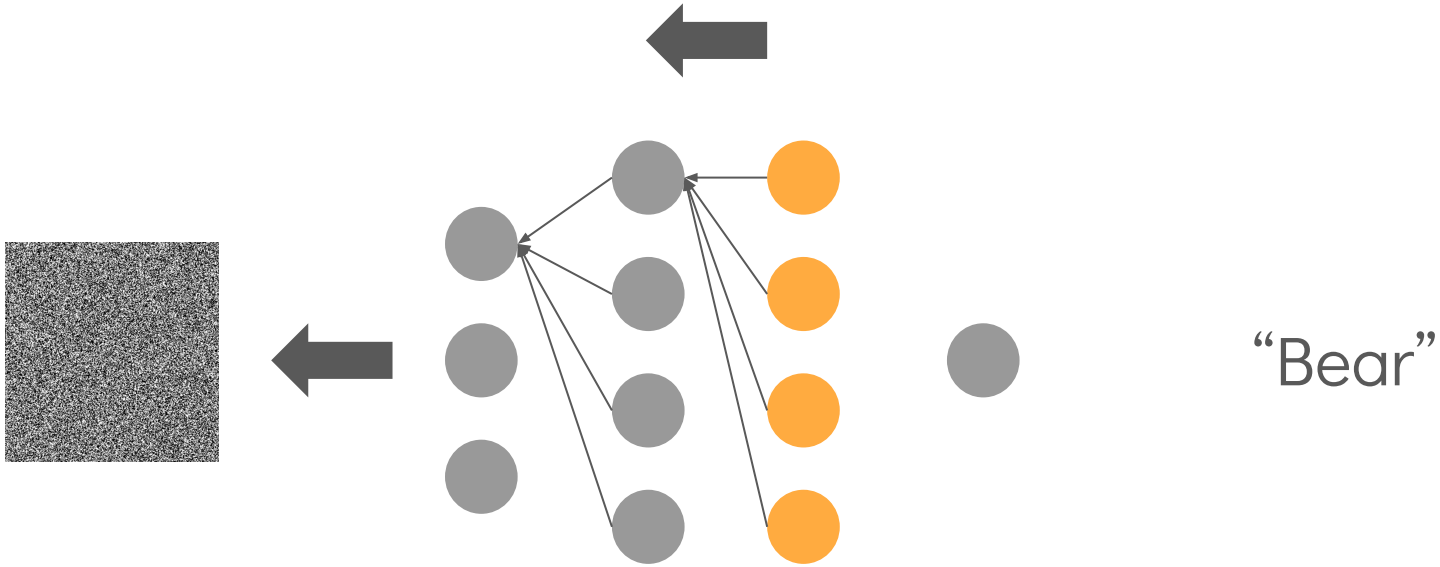
Model metamers illuminate divergences between biological and artificial neural networks

Jenelle Feather, Guillaume Leclerc, Aleksander Madry, & Josh H. McDermott (2023)



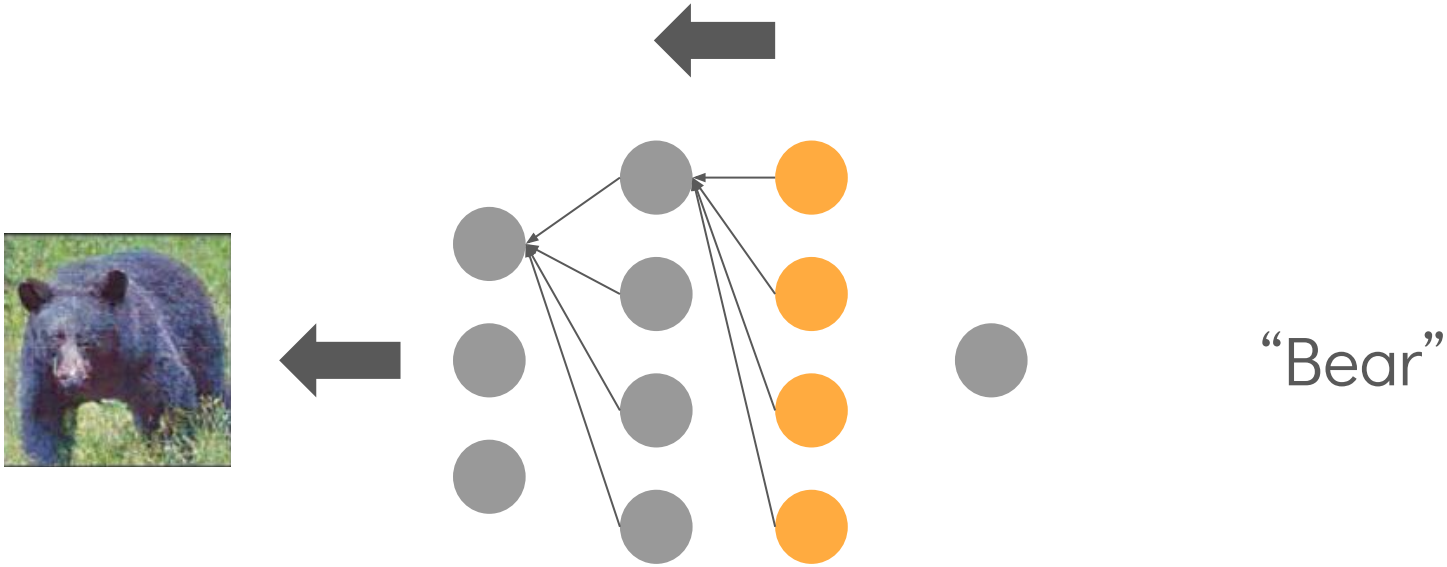
Model metamers illuminate divergences between biological and artificial neural networks

Jenelle Feather, Guillaume Leclerc, Aleksander Madry, & Josh H. McDermott (2023)



Model metamers illuminate divergences between biological and artificial neural networks

Jenelle Feather, Guillaume Leclerc, Aleksander Madry, & Josh H. McDermott (2023)

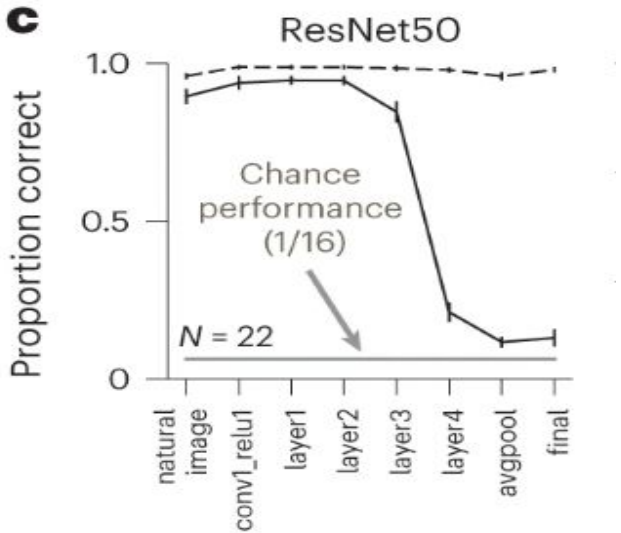


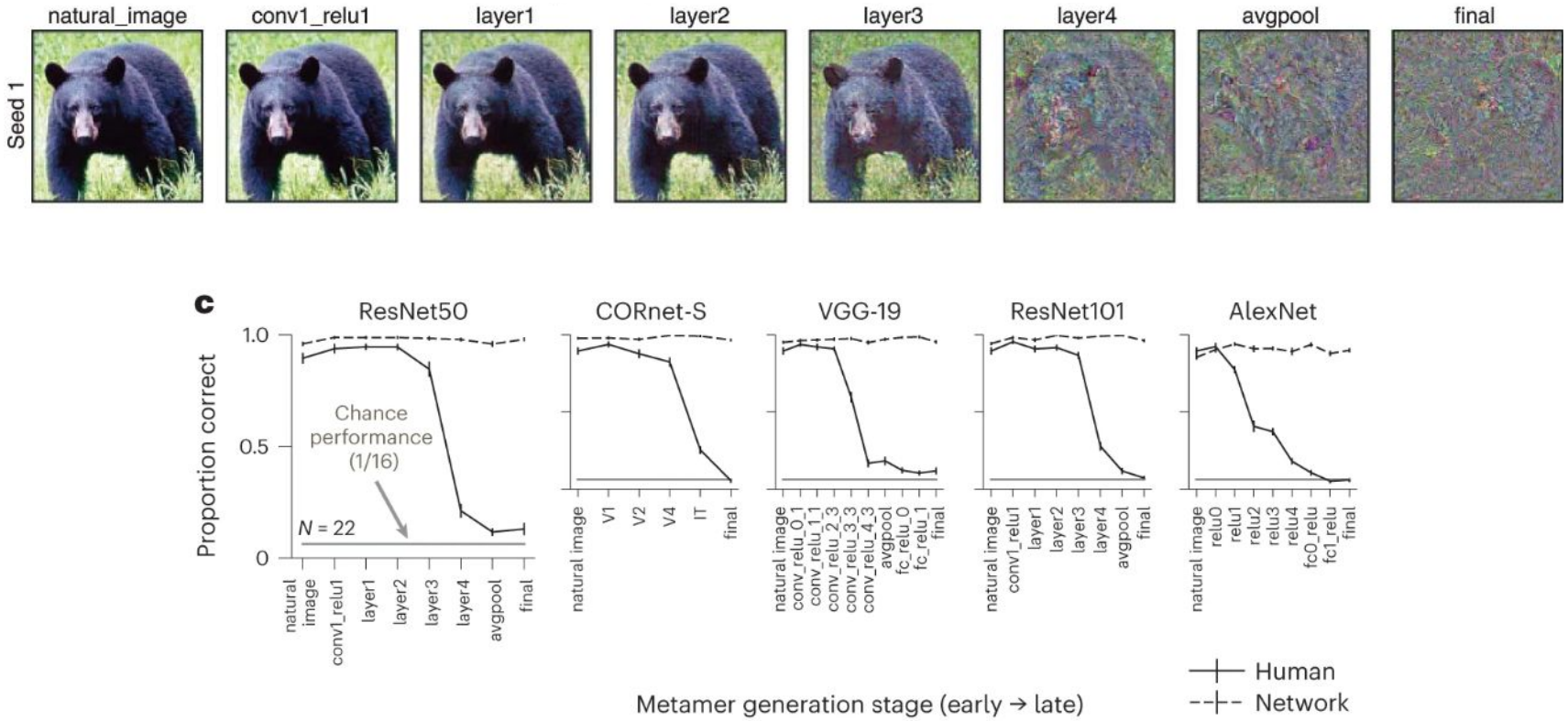
Model metamers illuminate divergences between biological and artificial neural networks

Jenelle Feather, Guillaume Leclerc, Aleksander Madry, & Josh H. McDermott (2023)

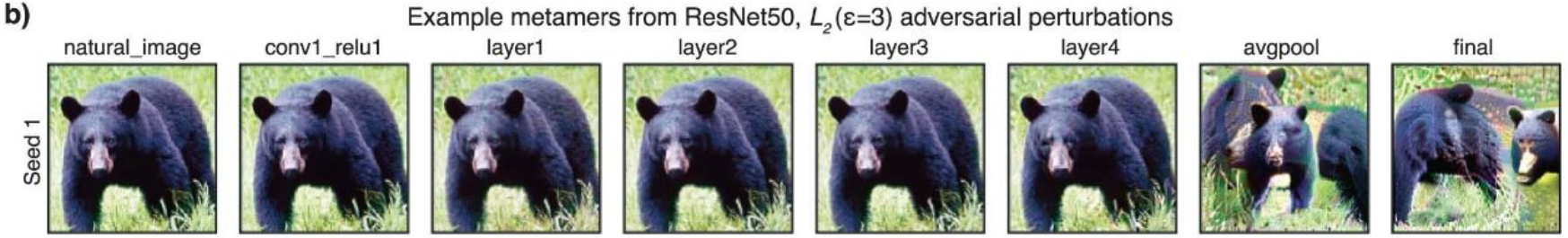

natural_image
conv1_relu1
layer1
layer2
layer3
layer4
avgpool
final
Seed 1

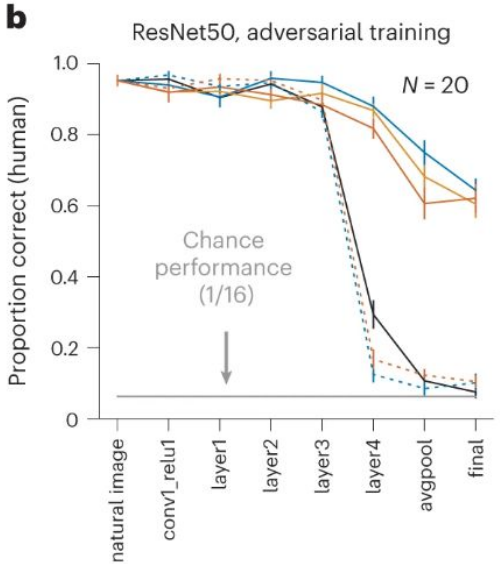

natural_image
conv1_relu1
layer1
layer2
layer3
layer4
avgpool
final
Seed 1


c
ResNet50
Proportion correct
1.0
0.5
0
Chance performance (1/16)
N = 22
natural image
conv1_relu1
layer1
layer2
layer3
layer4
avgpool
final


natural_image
conv1_relu1
layer1
layer2
layer3
layer4
avgpool
final
Seed 1
c
ResNet50
CORnet-S
VGG-19
ResNet101
AlexNet
Proportion correct
1.0
0.5
0
Chance performance (1/16)
N = 22
natural image
conv1_relu1
layer1
layer2
layer3
layer4
avgpool
final
V1
V2
V4
IT
conv_relu_0_1
conv_relu_1_1
conv_relu_2_3
conv_relu_3_3
conv_relu_4_3
fc_relu_0
fc_relu_1
relu0
relu1
relu2
relu3
relu4
fc0_relu
fc1_relu
Human
Network
Metamer generation stage (early → late)

b) Example metamers from ResNet50, $L_2(\varepsilon=3)$ adversarial perturbations


b)
Example metamers from ResNet50, L_2(ε=3) adversarial perturbations
natural_image
conv1_relu1
layer1
layer2
layer3
layer4
avgpool
final
Seed 1


b
ResNet50, adversarial training
N = 20
Proportion correct (human)
1.0
0.8
0.6
0.4
0.2
0
Chance performance (1/16)
natural image
conv1_relu1
layer1
layer2
layer3
layer4
avgpool
final

Takeaway

- **Models learn different invariances from humans**

Takeaway

- **Models learn different invariances from humans**

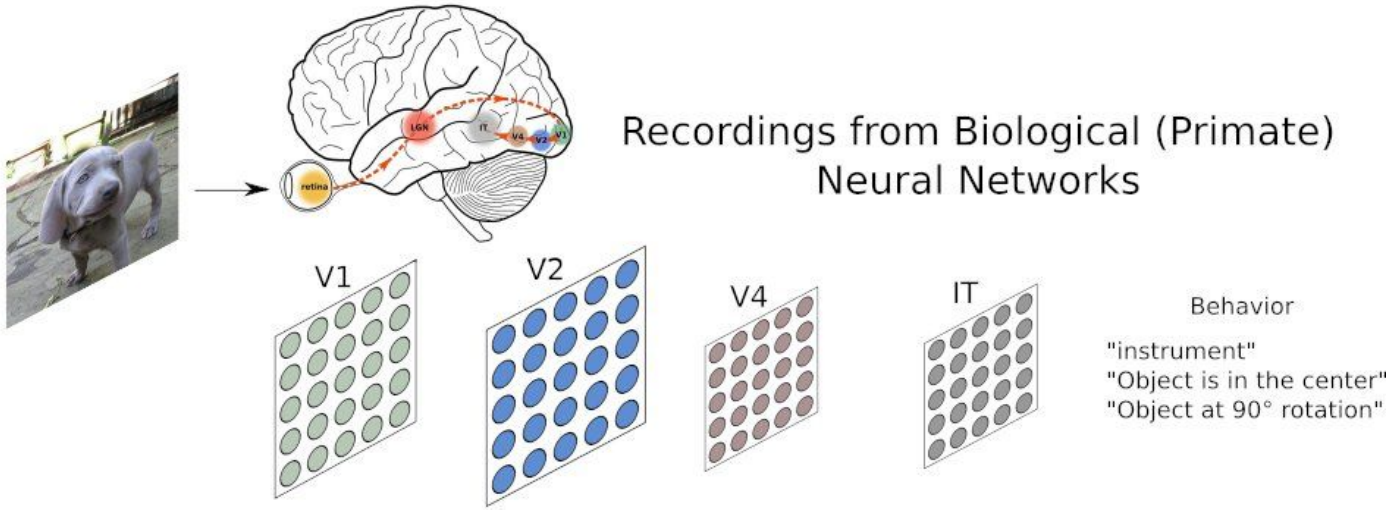
- **Different representational schemes can enable invariant classification**

Takeaway

- **Models learn different invariances from humans**
- **Different representational schemes can enable invariant classification**
- **Methods to make invariances more human-like**

Overview

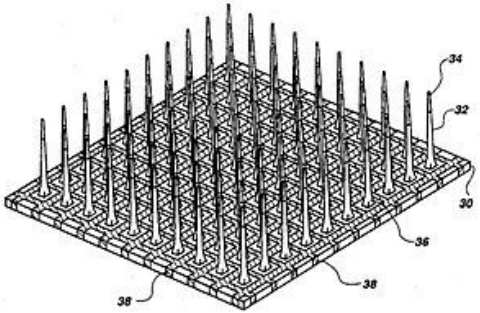
- Do ML models perceive like humans?
- Naive approach: Optical Illusions
- Approach 2: Psychological approach
- Approach 3: Neuroscience approach
- Improving methods
- Op-ed



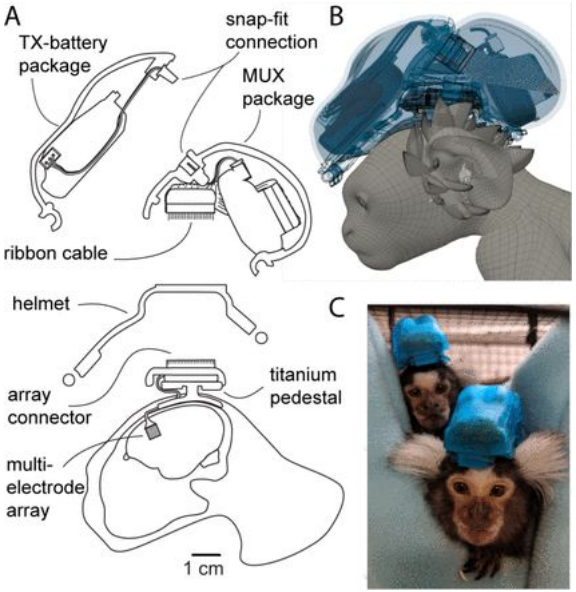
Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?

Martin Schrimpf, Jonas Kubilius, Ha Hong, et al. (2018).

Neuronal Data



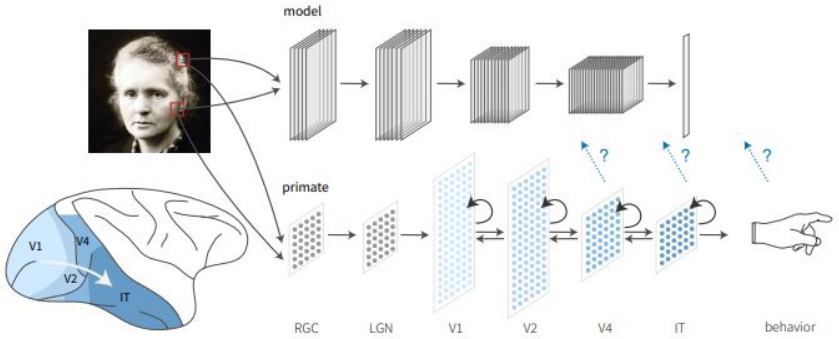
Microelectrode array



BrainScore

Neural

Average Neuronal Firing Rate

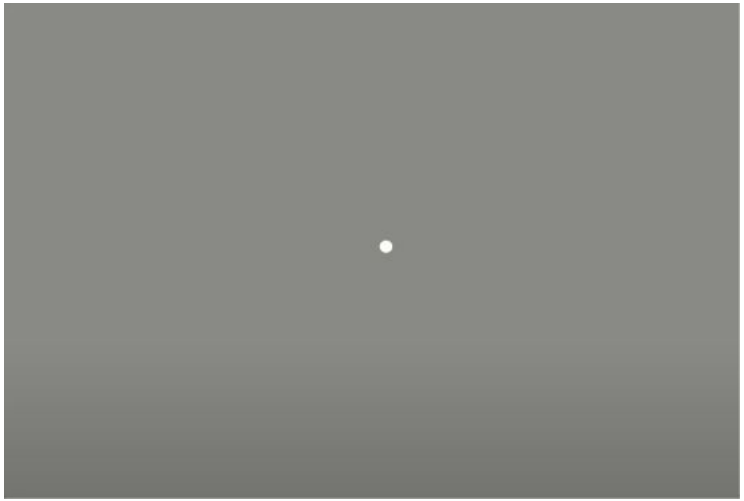


BrainScore

Neural
Average Neuronal Firing Rate

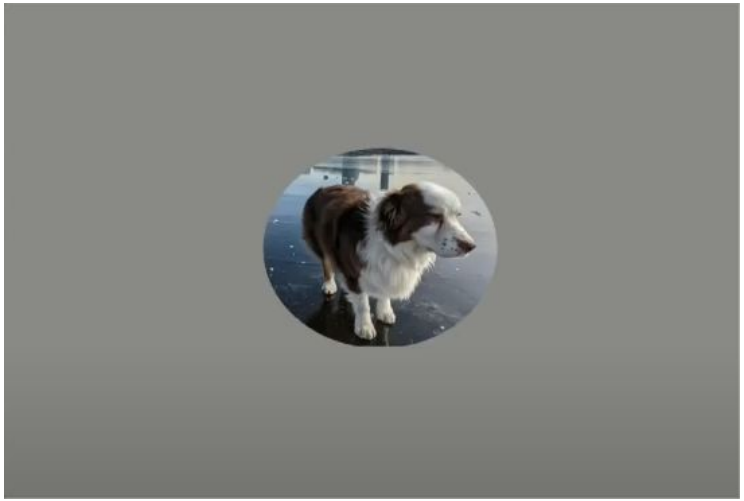
Behavioral
Model prediction of classification behavior

BrainScore



Behavioral
Model prediction of classification behavior

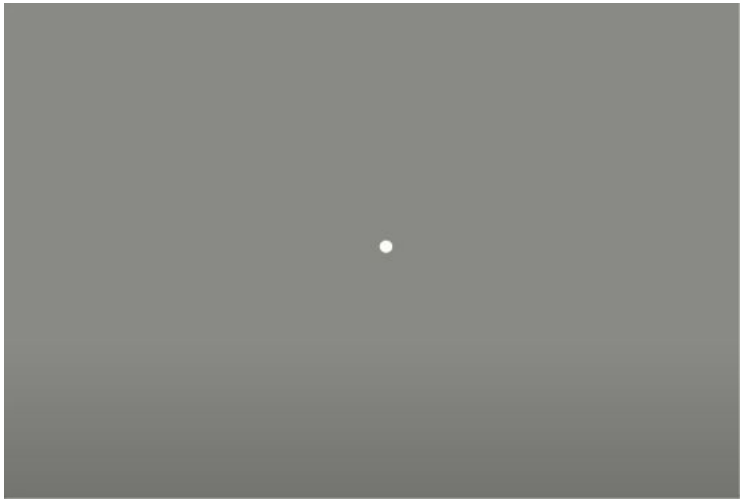
BrainScore



Behavioral

Model prediction of classification behavior

BrainScore

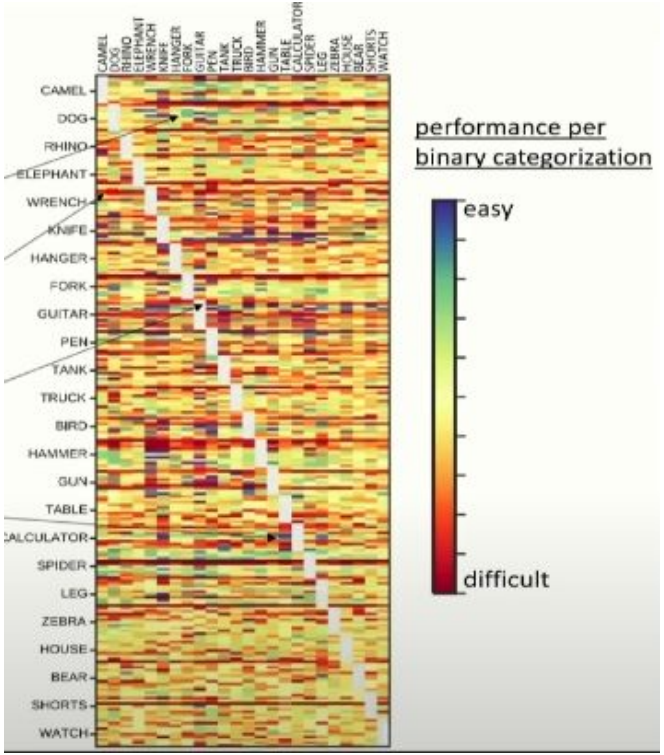


Behavioral
Model prediction of classification behavior

BrainScore

Behavioral
Model prediction of classification behavior

BrainScore



Behavioral

Model prediction of classification behavior

BrainScore

Neural
Average Neuronal Firing Rate

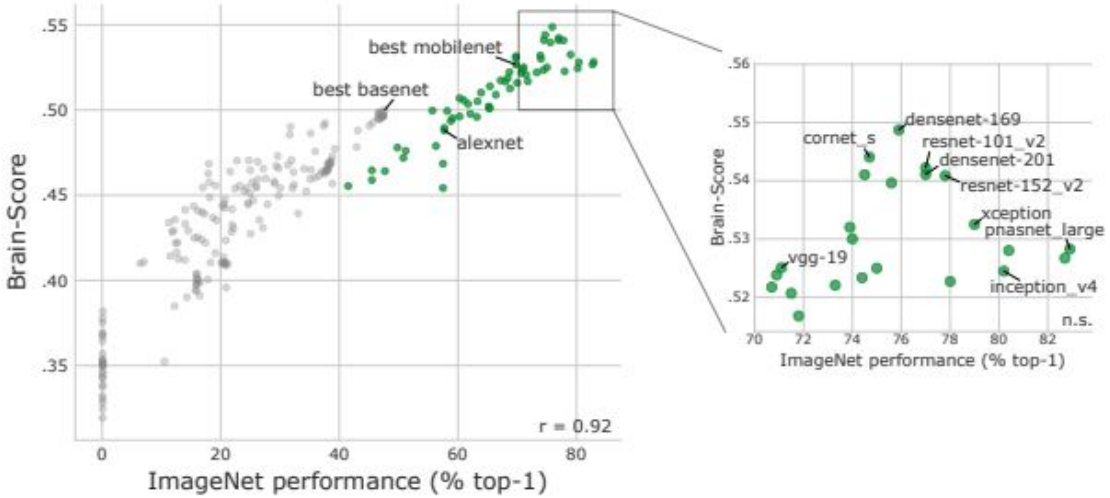
Behavioral
Model prediction of classification behavior

BrainScore

Neural
Average Neuronal Firing Rate

Behavioral
Model prediction of classification behavior

Averaged across layers and brain regions


Brain-Score
.55
.50
.45
.40
.35
0
20
40
60
80
ImageNet performance (% top-1)
r = 0.92
best mobilenet
best basenet
alexnet
.56
.55
.54
.53
.52
70
72
74
76
78
80
82
densenet-169
cornet_s
resnet-101_v2
densenet-201
resnet-152_v2
xception
pnasnet_large
vgg-19
inception_v4
n.s.

Welcome to Brain-Score!

Navigate our dashboard to view key submission data.

Leaderboard

Sort by Benchmarks

Score Legend: No Alignment — Alignment at ceiling

Rank	Model	average (51)	neural (33)	behavior (18)	engineering (25)
1	cvt_cvt-w24-384-in22k_finetuned-in1k_4 Wanhee Lee	.513	.417	.608	.629
2	resnext101_32x8d_wsl Martin Schrimpf	.489	.415	.563	.577
3	effnetb1_cutmixpatch_SAM_ Alexander Riedel	.482	.453	.511	.275
4	effnetb1_cutmixpatch_augm Alexander Riedel	.475	.455	.495	.348
5	resnext101_32x32d_wsl Martin Schrimpf	.462	.312	.613	.510

Rank	Model	average 3	neural 2	behavior 1	engineering 30
1	gpt-neo-1.3B Brain-Score Team	.662	1.0	.324	.800
2	gpt2-xl Brain-Score Team	.652	.987	.318	.825
3	gpt-neo-2.7B Brain-Score Team	.649	.990	.308	.847
4	gpt2-large Brain-Score Team	.639	.948	.331	.786
5	gpt2-medium Brain-Score Team	.634	.918	.350	.796
6	gpt2 Brain-Score Team	.608	.850	.367	.762
7	openai-gpt Brain-Score Team	.594	.802	.387	.523
8	distilgpt2 Brain-Score Team	.590	.818	.361	.503
9	gpt-neo-125m	.571	.784	.358	.736

Overview

- Do ML models perceive like humans?
- Naive approach: Optical Illusions
- Approach 2: Psychological approach
- Approach 3: Neuroscience approach
- Improving methods
- Op-ed

Improvements

Improvements

Instructional-tuning Aligns LLMs To The Human Brain

Khai Loong Aw, Syrielle Montariol, Badr Al Khamissi. (2023).

Improvements

Instructional-tuning Aligns LLMs To The Human Brain

Khai Loong Aw, Syrielle Montariol, Badr Al Khamissi. (2023).

Aligning Model And Macaque Inferior Temporal Cortex Representations Improves Model-to-human Behavioral Alignment And Adversarial Robustness

Joel Dapello, Kohitij Kar, Martin Schrimpf. (2023)

Improvements

Instructional-tuning Aligns LLMs To The Human Brain

Khai Loong Aw, Syrielle Montariol, Badr Al Khamissi. (2023).

Aligning Model And Macaque Inferior Temporal Cortex Representations Improves Model-to-human Behavioral Alignment And Adversarial Robustness

Joel Dapello, Kohitij Kar, Martin Schrimpf. (2023)

Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks

Hojin Jang, Frank Tong. (2023)

Improvements

Instructional-tuning Aligns LLMs To The Human Brain

Khai Loong Aw, Syrielle Montariol, Badr Al Khamissi. (2023).

Aligning Model And Macaque Inferior Temporal Cortex Representations Improves Model-to-human Behavioral Alignment And Adversarial Robustness

Joel Dapello, Kohitij Kar, Martin Schrimpf. (2023)

Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks

Hojin Jang, Frank Tong. (2023)

Overview

- Do ML models perceive like humans?
- Naive approach: Optical Illusions
- Approach 2: Psychological approach
- Approach 3: Neuroscience approach
- Improving methods
- Op-ed

Do we know enough about the brain?

Do we know enough about the brain?

Neurological

Can we predict human behavior from neurons?

Do we know enough about the brain?

Neurological

Can we predict human behavior from neurons?

Psychological

Do we know if illusions are a bug or a feature?

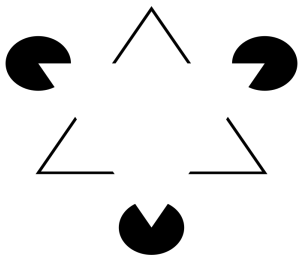
Do we know enough about the brain?

Neurological

Can we predict human behavior from neurons?

Psychological

Do we know if illusions are a bug or a feature?



Do we know enough about the brain?

Neurological

Can we predict human behavior from neurons?

Psychological

Do we know if illusions are a bug or a feature?

But

Maybe neural networks are similar to the brain?

Can ML models teach us about the brain?

Can ML models teach us about the brain?

Similarities

- Classify images
- Model can predict neuronal activation

Can ML models teach us about the brain?

Similarities

- Classify images
- Model can predict neuronal activation

Differences

- Adversarial examples
- Different class of mistakes

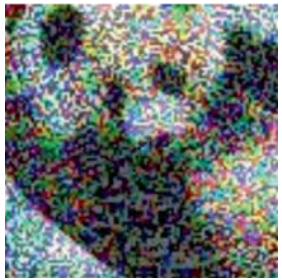
Can ML models teach us about the brain?

Similarities

- Classify images
- Model can predict neuronal activation

Differences

- Adversarial examples
- Different class of mistakes



Can ML models teach us about the brain?

Similarities

- Classify images
- Model can predict neuronal activation

Differences

- Adversarial examples
- Different class of mistakes

Behavior of ML models are influenced by data

Can ML models teach us about the brain?

Similarities

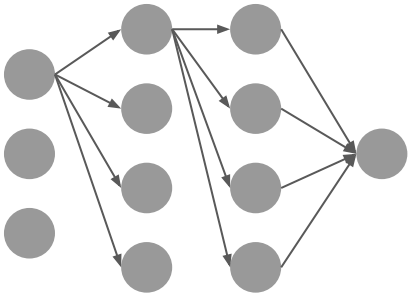
- Classify images
- Model can predict neuronal activation

Differences

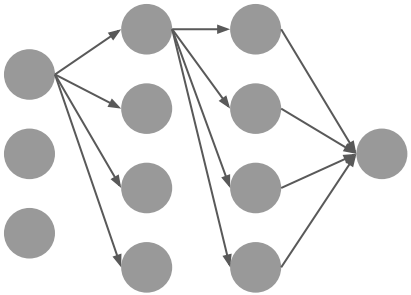
- Adversarial examples
- Different class of mistakes

Behavior of ML models are influenced by data

Functional not mechanistic

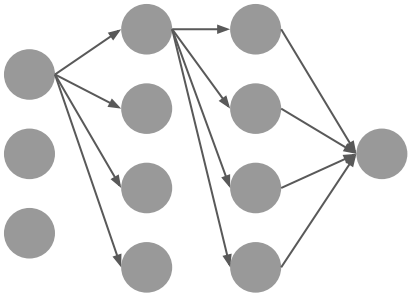


edge detection
color
attention
segmentation
...



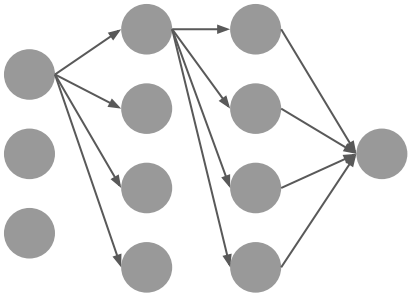
edge detection
color
attention
segmentation
...

???



edge detection
color
attention
segmentation
...

???

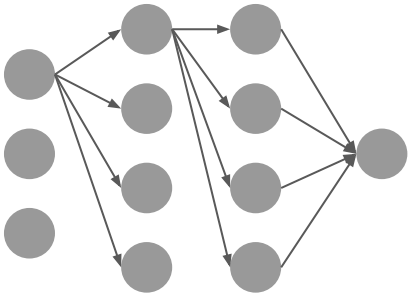


Why are we comparing functional models with something we don't understand well?

edge detection
color
attention
segmentation
...

???

We don't need to know how to do all downstream tasks



Exercises:

- What are some aspects of NLP that is human-like? What are aspects that aren't?
- How do humans combine different sensory information? How do ML models do so?