

AI Image Analysis



Image Classification

- What is in this image?
 1. Cat
 2. Car
 3. Castle
 4. Corn



Image Classification

- **What is in this image?**

1. Cat
2. Car
3. Castle
4. Corn



- **What features of the image led you to your answer?**

Image Features

- **Image features are geometric in nature**
 - Edges
 - Shapes

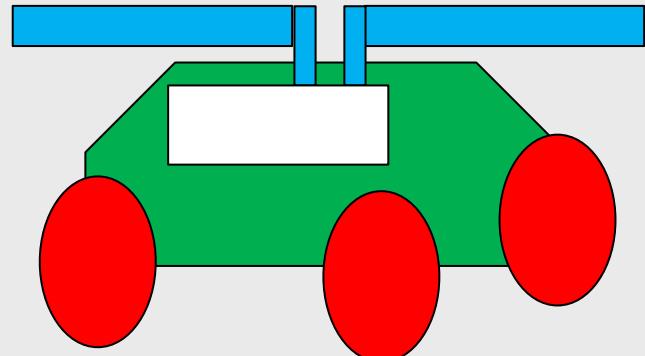
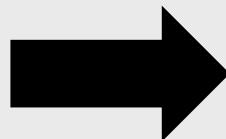


Image Features

- Absolute position not so important
 - If you move the car, it is still a car



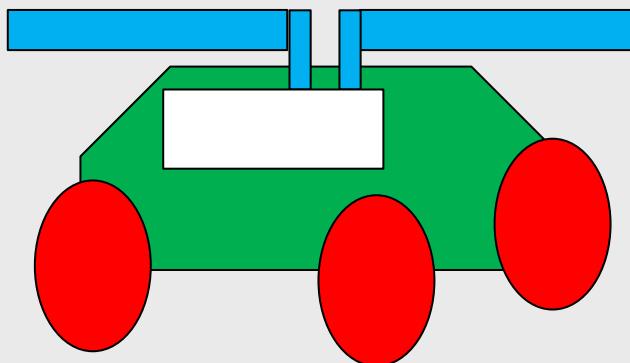
Car



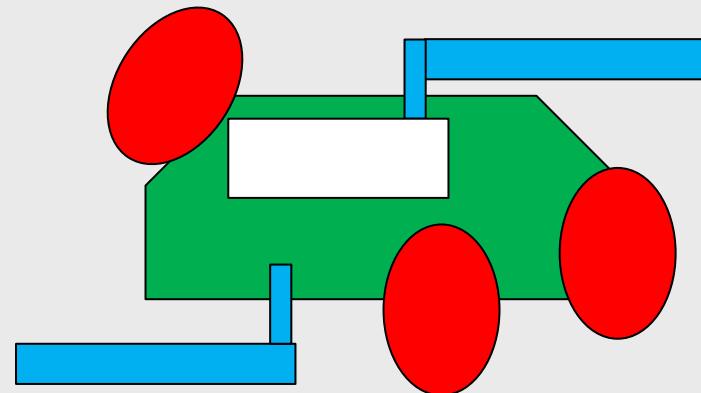
Car

Image Features

- Relative position of features is important
 - Wheels below car body
 - Headlights on front of body



Car



Not a Car

Learning Image Features

- **Image features**
 - Edges
 - Shapes built from edges
 - Combinations of shapes
 - Deeper hierarchy of combinations of shapes
 - Combinations of combinations of ... shapes
 - Translation invariant
 - Doesn't matter where in the image the object is
- **We need to learn what these features and their combinations are for different object categories**
- **Solution: Neural Networks**

Neural Networks and Image Analysis

- Neural networks began to be used for image analysis in the early 1990's
- Main architecture – **convolutional network**
- Convolution is an operation to find patterns in an image

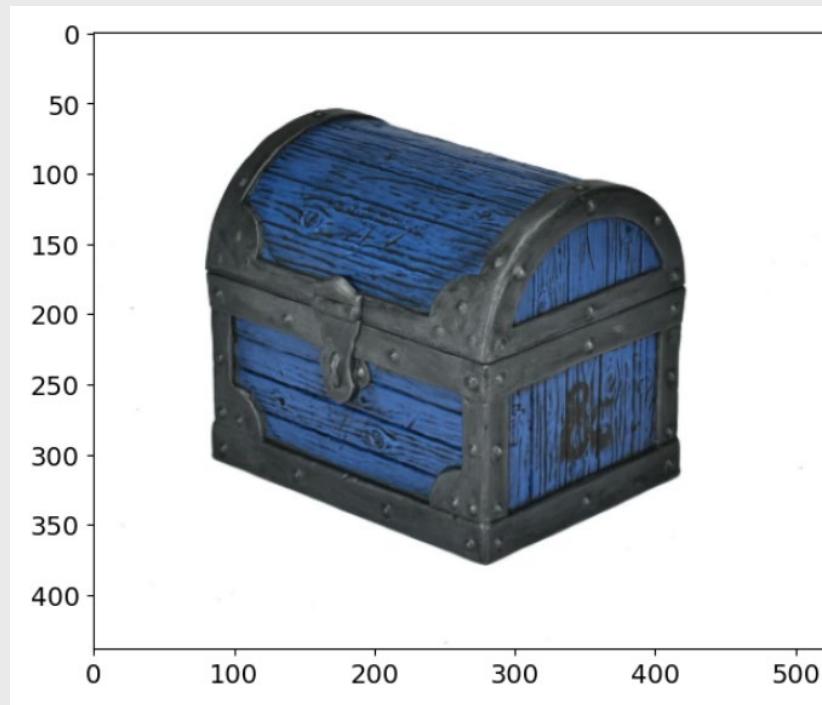
Convolution

- The image is a big matrix, the pattern is a small matrix
- If we convolve the two matrices, it will tell us if the pattern is in the image

Pattern for
vertical edges

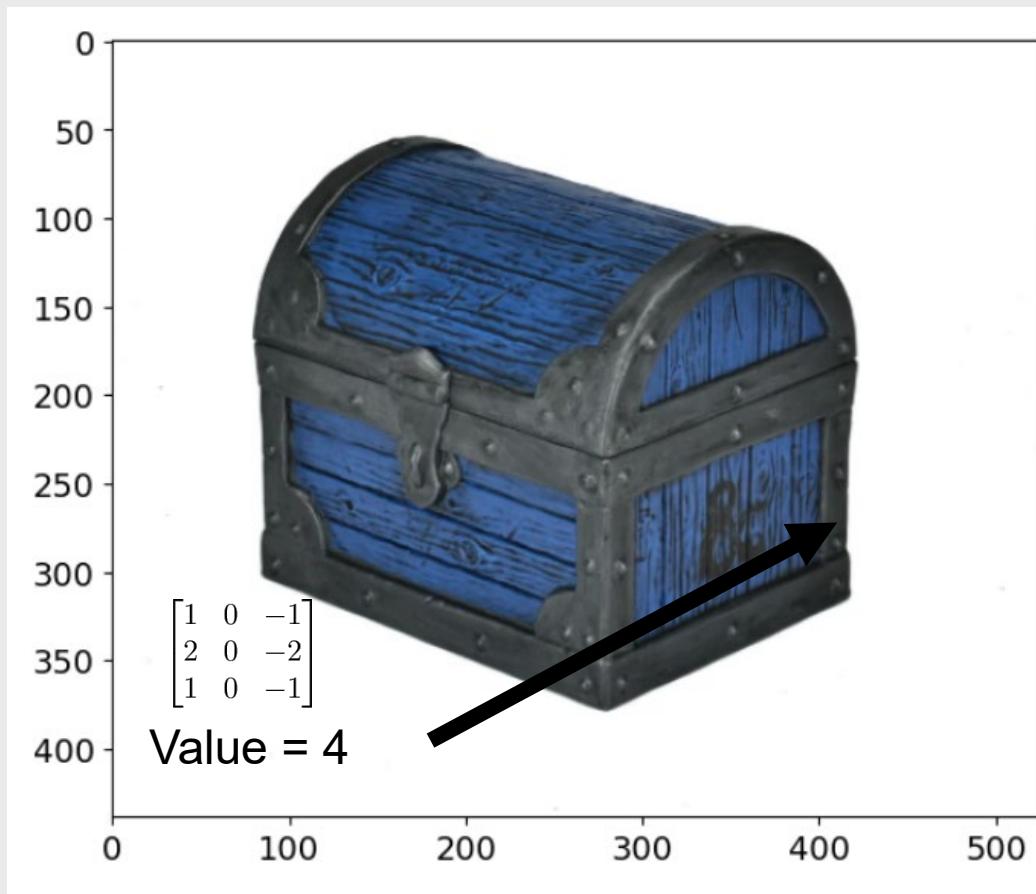
$$\begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

Image



Convolution

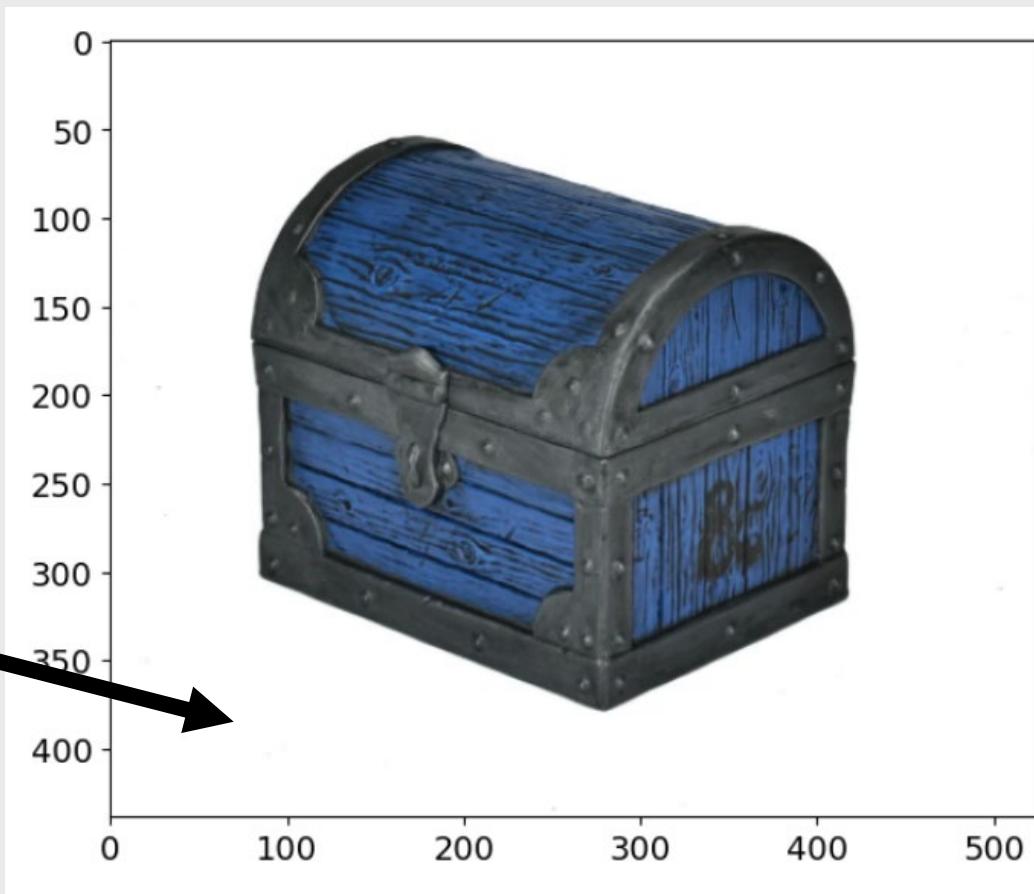
- The convolution at a pixel equals the similarity of the pattern and image centered at the pixel



$$\begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \quad \begin{bmatrix} -1 \\ -2 \\ -1 \end{bmatrix}$$

Convolutional Network

- The convolution at a pixel equals the similarity of the pattern and image centered at the pixel

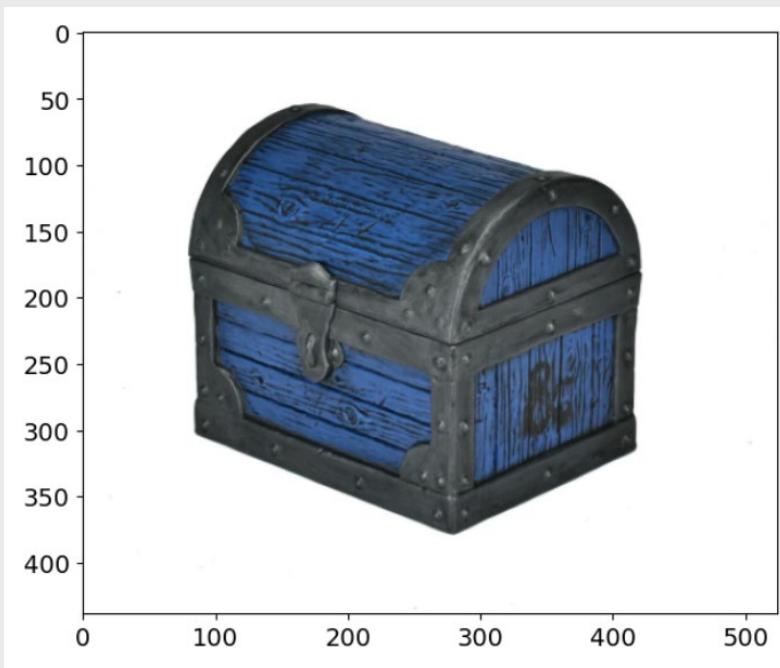


$$\begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

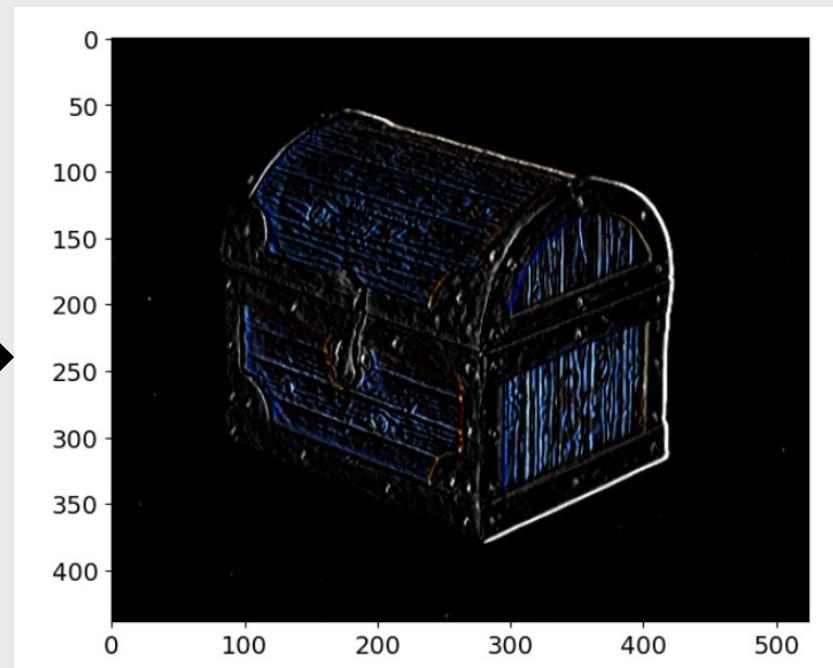
$$\begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

Convolutional Network

- Convolution with a pattern gives bright pixels where the image matches the pattern

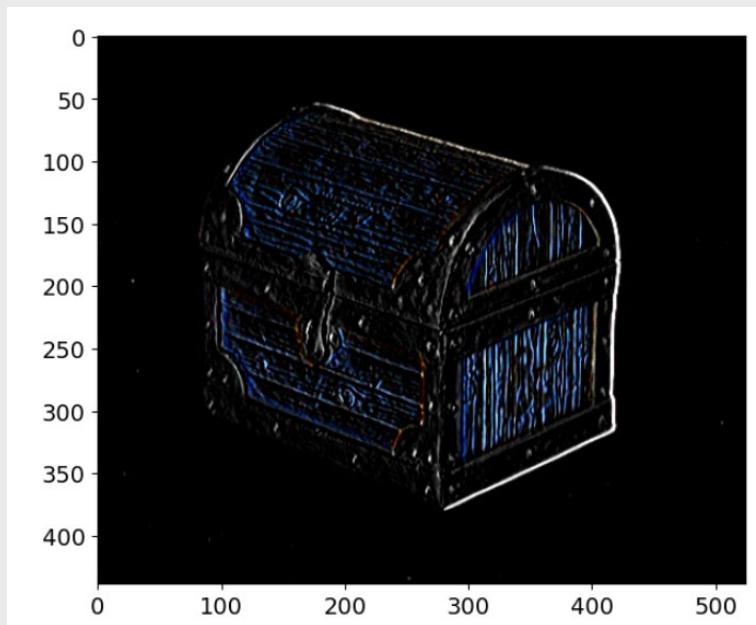


$$\begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$



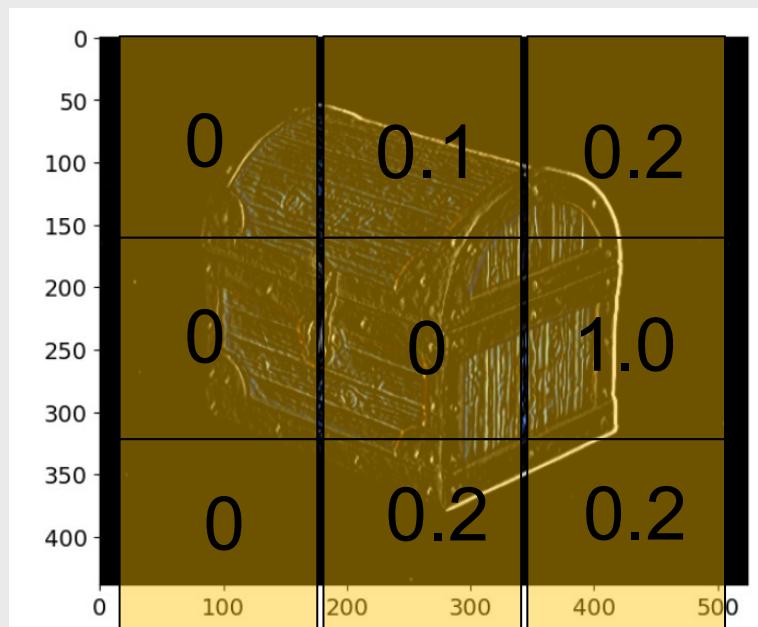
Max-Pooling

- After we convolve a pattern with the image, we have the same size image
- We don't need to know exactly where the feature is, so we can shrink the image



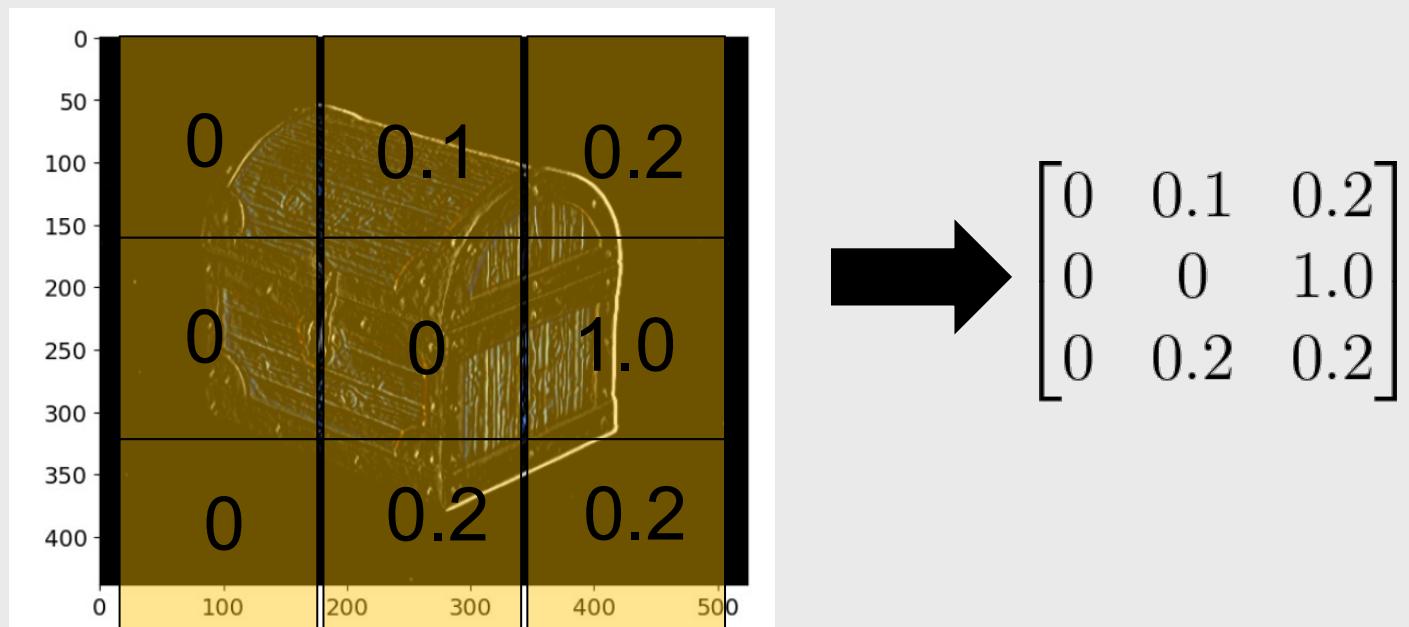
Max-Pooling

- After we convolve a pattern with the image, we have the same size image
- We don't need to know exactly where the feature is, so we can shrink the image
- Max-pooling lets us shrink the image with a grid of patches and keep the max value of the convolution in each patch



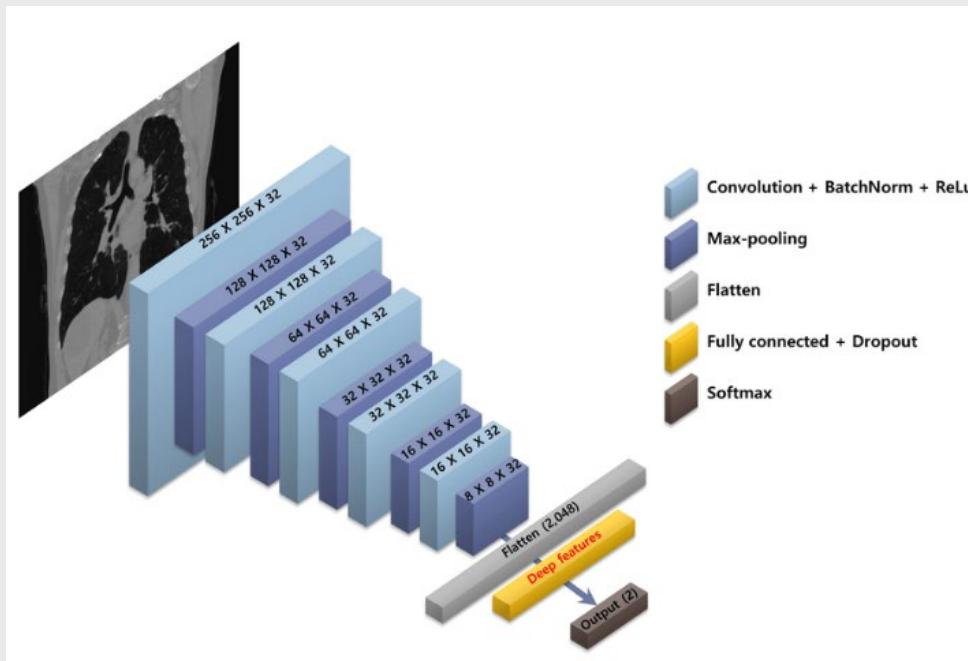
Max-Pooling

- After we convolve a pattern with the image, we have the same size image
- We don't need to know exactly where the feature is, so we can shrink the image
- Max-pooling lets us shrink the image with a grid of patches and keep the max value of the convolution in each patch



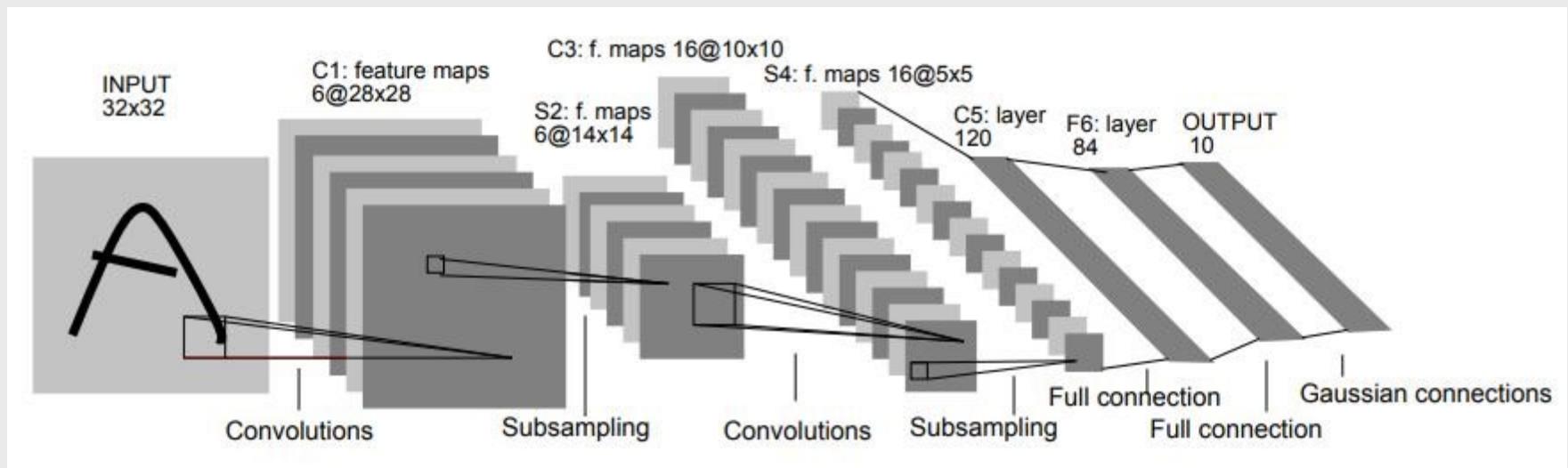
Convolutional Network

- A convolutional neural network is many layers of convolution and max pooling
 - Convolution – find patterns (also called features)
 - Max pooling – shrink the image and keep pattern information
- Learns many features to understand the image
- Deep networks let you learn combinations of features
- Last few fully connected layers classify the image based on the features



First Useful Convolutional Network

- 1989 – Yann LeCunn proposes convolutional networks for reading handwritten digits
- 1998 – demonstrated practical applications of convolutional network, called it LeNet



MNIST Database

- MNIST (Modified National Institute of Standards and Technology) database – 160,000 images of handwritten digits
- Created in 1994
- One of the first standardized image datasets for training and testing neural networks



Going Deeper

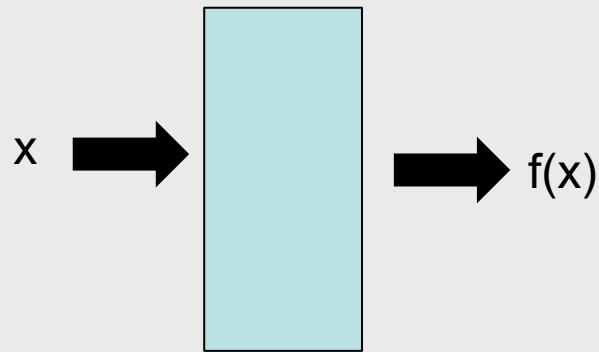
- People found that deep convolutional networks performed better
- People also found that deeper networks were harder to train
 - Deep here meant 10 – 20 layers

Going Deeper

- **People found that deep convolutional networks performed better**
- **People also found that deeper networks were harder to train**
 - Deep here meant 10 – 20 layers
- **ResNet is developed in 2015 to allow for training of deep networks**
 - ResNet had 152 layers!

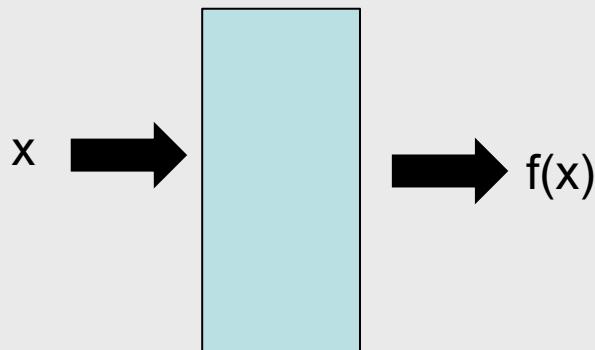
ResNet

- Normal layer takes in x and outputs $f(x)$

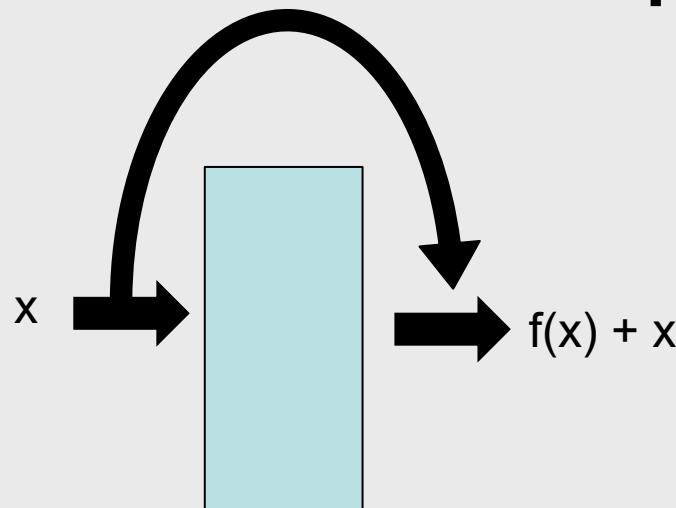


ResNet

- Normal layer takes in x and outputs $f(x)$

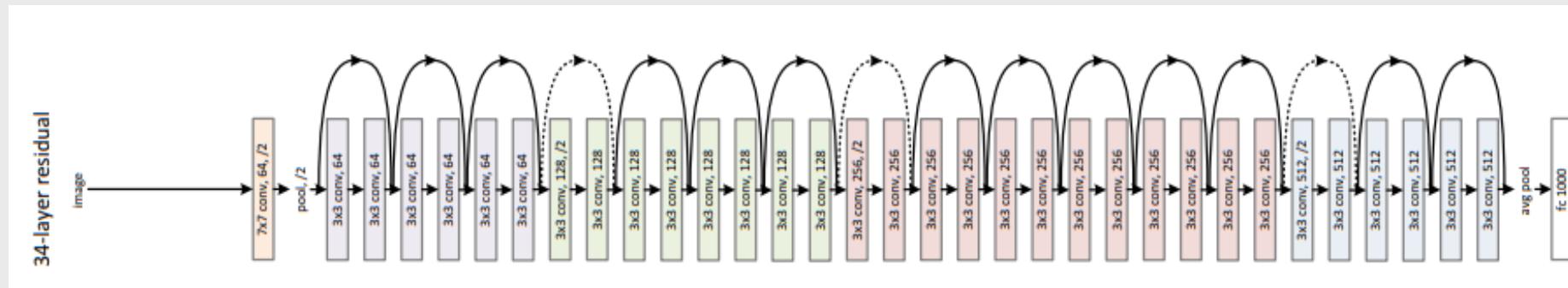


- ResNet layer takes in x and outputs $f(x) + x$



ResNet

- Outputting $f(x) + x$ lets the model see the impact of each parameter on output
 - This improves training process
 - Allows for very very deep networks



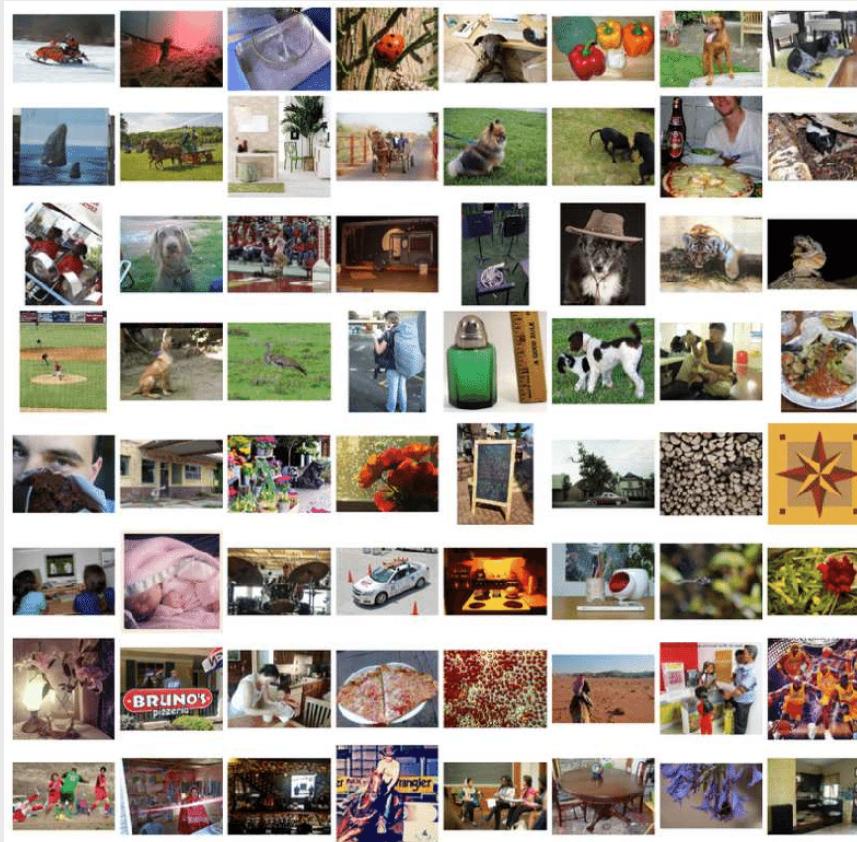
ImageNet

- **ImageNet is a dataset of labeled images used in computer vision**
 - **14 million images**
 - **20,000 categories**
- **Example categories**

898	water bottle
899	water jug
900	water tower
901	whiskey jug
902	whistle
903	wig
904	window screen
905	window shade
906	Windsor tie
907	wine bottle
908	wing
909	wok
910	wooden spoon
911	wool, woolen, woollen
912	worm fence, snake fence, snake-rail fence, Virginia fence

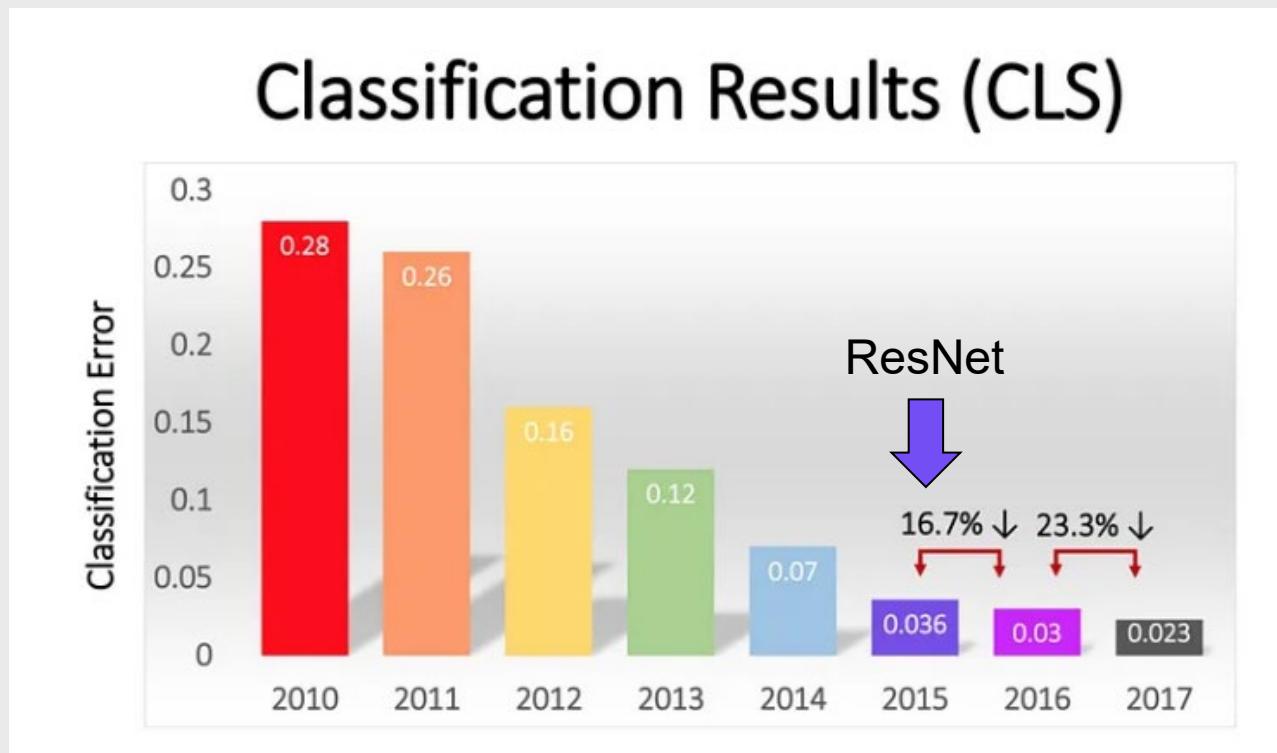
ImageNet Challenge

- **ImageNet Large Scale Visual Recognition Challenge**
 - **Started in 2010**
 - **Need to identify images in 1000 categories**



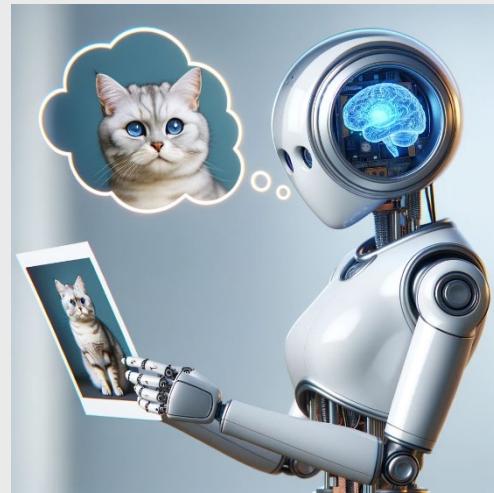
ImageNet Leaderboard

- ResNet provided a breakthrough improvement in ImageNet score



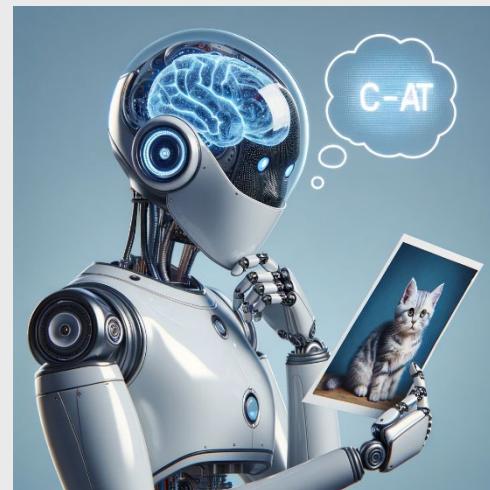
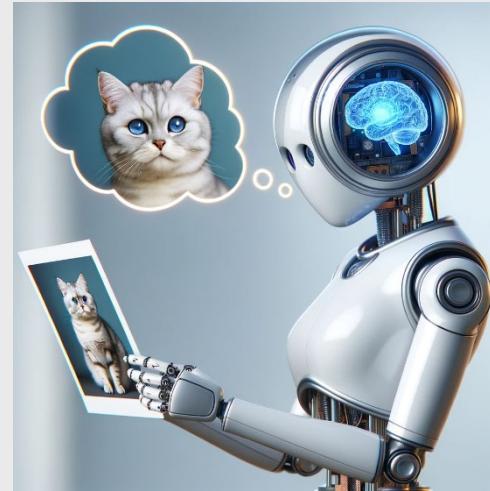
Mapping Image to Text

- ResNet convolutional models are good at building features about an image for classification



Mapping Image to Text

- ResNet convolutional models are good at building features about an image for classification
- What if we want to map an image to text that describes it?



Predictive Learning

- Predictive Learning
 - Identify image category
 - Traditional training method for neural network classifiers
- Learning Question: What is this?



Cat



Dog



Cat

Predictive Learning Challenges

- If categories are broad, predictive learning can be effective



Cat



Cat



Cat

Predictive Learning Challenges

- If categories are narrow, predictive learning can be ineffective



The image captures a close-up of a cat with a white and grey marked face, large expressive eyes, and a soft coat, set against a dark background, with a visible copyright symbol.



The image depicts an orange and white cat mid-meow or hiss, with one eye partially closed, against a blurred background.



The image shows an illustration of an orange, grumpy-looking cat standing upright with exaggerated features, including large eyes and a pout.

Contrastive Learning

- **Contrastive learning**
 - Determine if a pair of items are the same
 - Lets one learn to contrast categories
 - Useful if categories are very narrow

Contrastive Learning

- **Contrastive learning**
 - Determine if a pair of items are the same
 - Lets one learn to contrast categories
 - Useful if categories are very narrow
- **Learning Question: Are these the same?**



No



Yes

Contrastive Learning

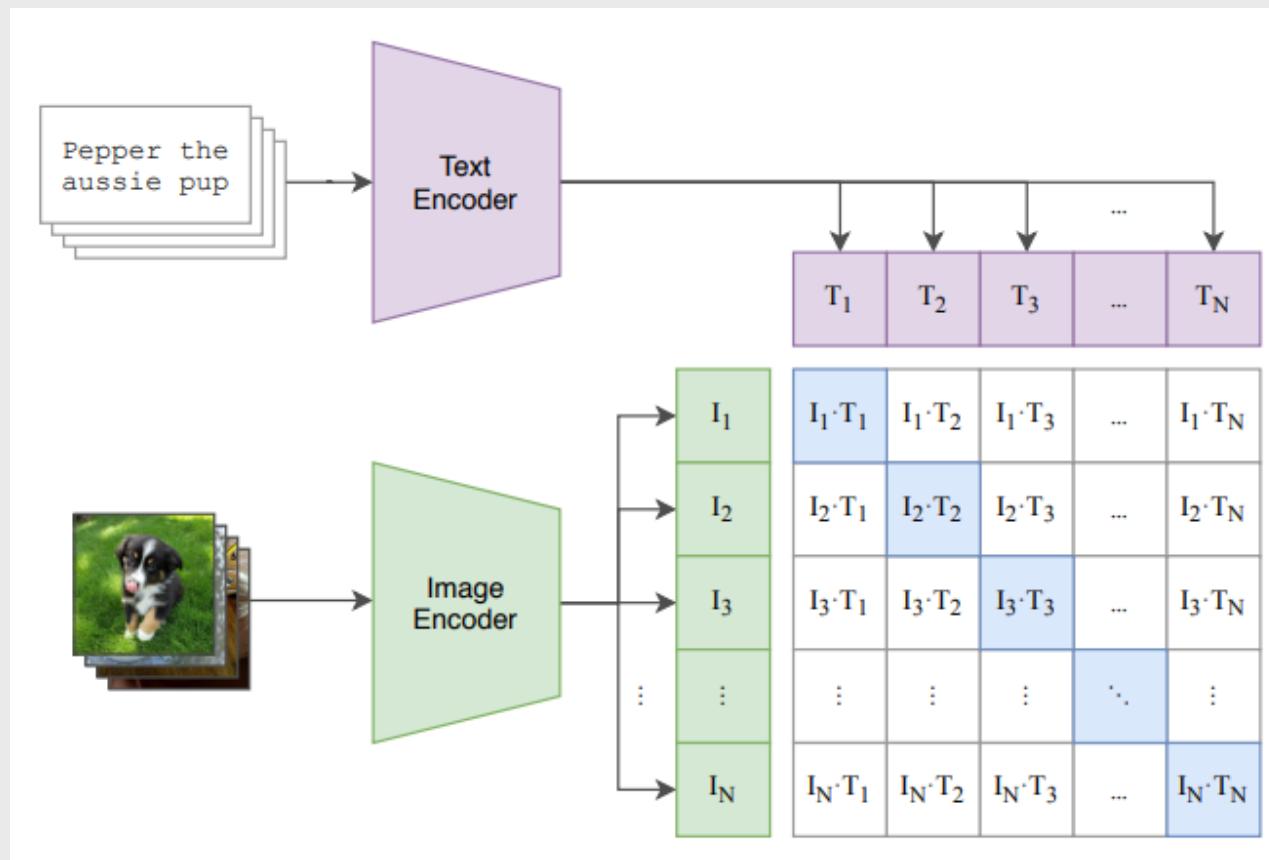


CLIP

- Contrastive Language-Image Pre-Training (CLIP) is a model developed by OpenAI in 2021 that maps images to text
- Combines transformers (text) with convolutional networks (images)
- Allows an AI to describe an image in words
- Key idea: **Contrastive Learning**

CLIP Training

- **Text encoder (transformer) and image encoder (ResNet) jointly trained so the embedding vectors they make point in the same direction**



Contrastive Learning in CLIP



The image captures a close-up of a cat with a white and grey marked face, large expressive eyes, and a soft coat, set against a dark background, with a visible copyright symbol.

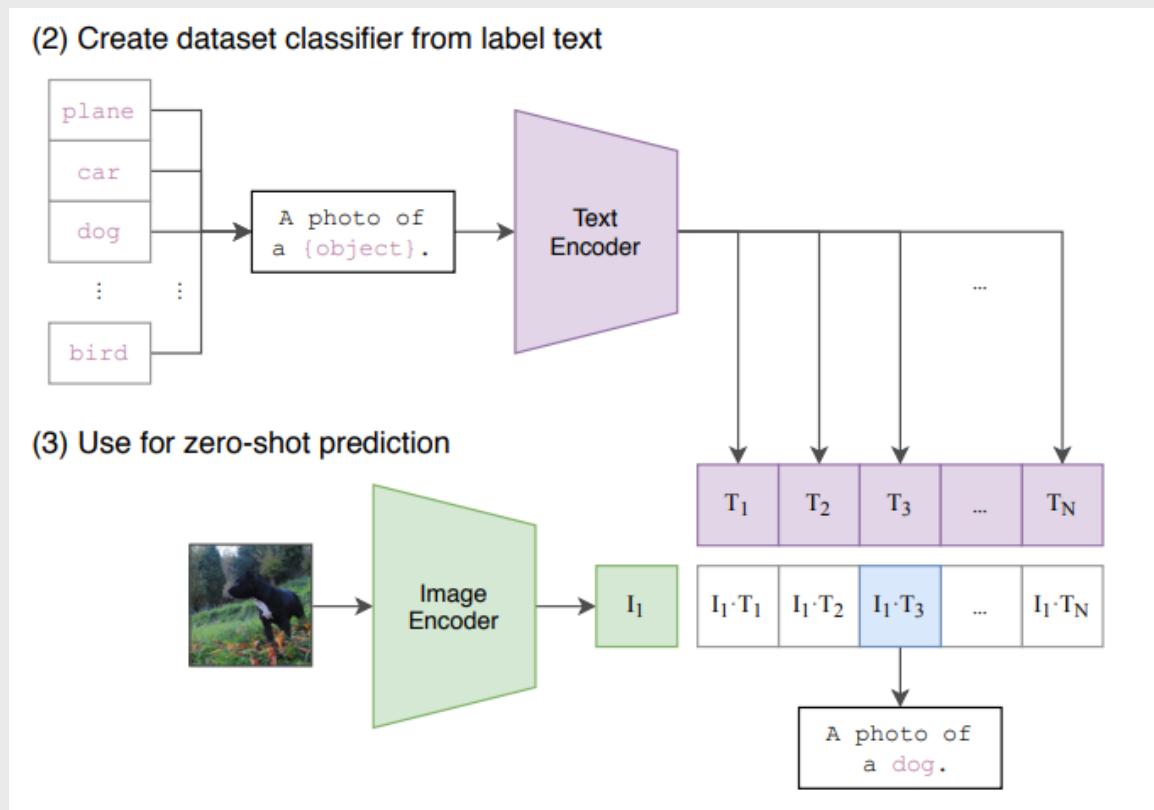
The image depicts an orange and white cat mid-meow or hiss, with one eye partially closed, against a blurred background.

The image shows an illustration of an orange, grumpy-looking cat standing upright with exaggerated features, including large eyes and a pout.

0.9	0.6	0.3
0.5	0.9	0.8
0.6	0.7	0.9

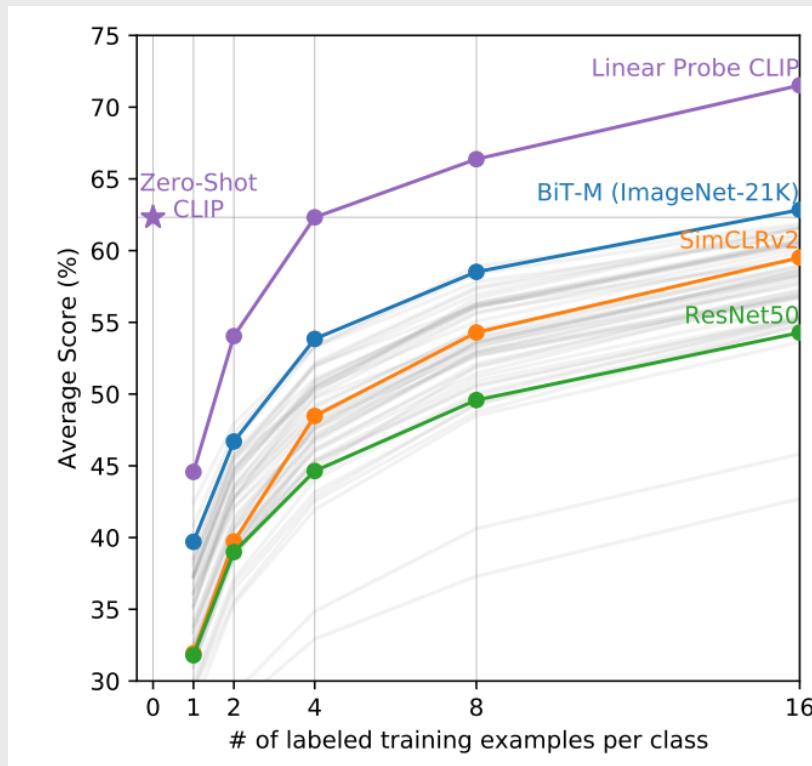
CLIP Prediction

- After trained, CLIP can guess the best caption for an image from a list of captions



CLIP Zero-Shot Classification

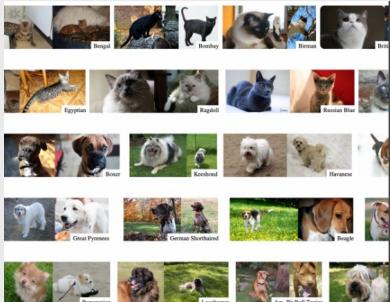
- CLIP performs zero-shot learning of images classes
 - “A photo of a {label}”
- Competitive with trained classifiers based on best ImageNet models



CLIP Prompt Engineering

- CLIP's performance improves with some simple prompt engineering

Oxford III Pets



A photo of a {label},
a type of pet.

Food 101



A photo of a {label},
a type of food.

FGCV Aircraft



A photo of a {label},
a type of aircraft.

Satellite Images



A satellite photo of a {label}.

GPT-4 With Vision



- **GPT-4o lets you generate a text from an image + text prompt**
 - Used to be called GPT-4V (Vision)
- **CLIP is likely the underlying model for the GPT-4o**

Vision

Learn how to use GPT-4 to understand images

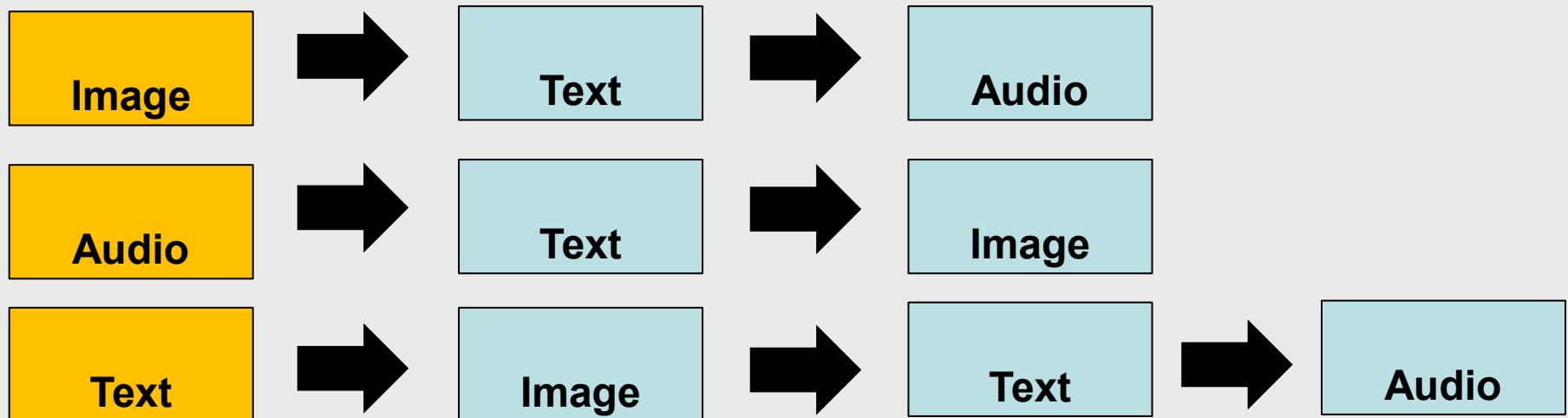
Introduction

GPT-4 with Vision, sometimes referred to as [GPT-4V](#) or `gpt-4-vision-preview` in the API, allows the model to take in images and answer questions about them. Historically, language model systems have been limited by taking in a single input modality, text. For many use cases, this constrained the areas where models like GPT-4 could be used.

GPT-4 with vision is currently available to all [developers who have access to GPT-4](#) via the `gpt-4-vision-preview` model and the Chat Completions API which has been updated to support image inputs. Note that the [Assistants API](#) does not currently support image inputs.

GPT-4o, DALLE-3, TTS, STT

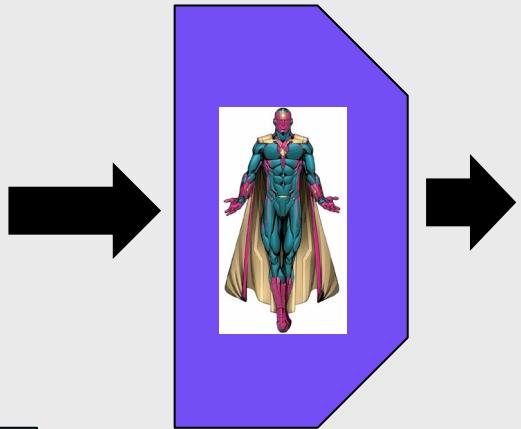
- We now have multiple generative AI technologies we can use
 - GPT-4o: Text to text
 - GPT-4o: Image to text
 - DALLE-3: Text to image
 - Audio: Text to speech
 - Whisper: Speech to text
- Many creative ways to combine them...



GPT-4o



please list each food item in this image.



In the image, I can see several food items stored in a refrigerator. From top to bottom, here are the visible items:

1. Eggs – a carton of large eggs is partially visible on the top shelf.
2. Strawberries – there's a plastic clamshell container of strawberries on the middle shelf.
3. Grapes – to the left of the strawberries, there are red grapes, some of which are falling out of an open bag.
4. Minced garlic – a jar of minced garlic labeled "GARLIC" is underneath the strawberries.
5. Potatoes – below the jar of garlic, there is a small bag of potatoes.
6. Cheese – on the middle shelf, in front of the potatoes, there appears to be a block of cheese.

I can also see the upper edges of some items in bins or drawers labeled "Fresh Keeper," which suggests the presence of fresh produce such as fruits or vegetables. However, the actual contents are obscured, and specifics cannot be determined from the visible portions.

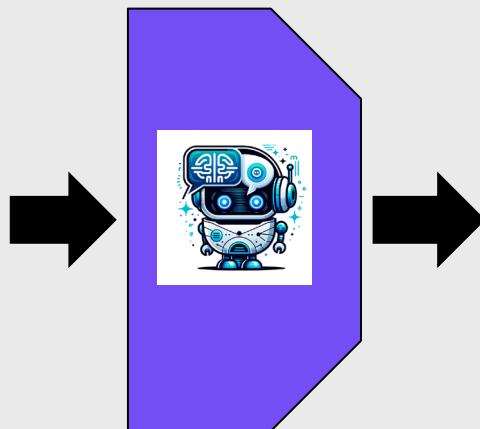
GPT-4o

make me a recipe with the following ingredients from my fridge:

In the image, I can see several food items stored in a refrigerator. From top to bottom, here are the visible items:

1. Eggs – a carton of large eggs is partially visible on the top shelf.
2. Strawberries – there's a plastic clamshell container of strawberries on the middle shelf.
3. Grapes – to the left of the strawberries, there are red grapes, some of which are falling out of an open bag.
4. Minced garlic – a jar of minced garlic labeled "GARLIC" is underneath the strawberries.
5. Potatoes – below the jar of garlic, there is a small bag of potatoes.
6. Cheese – on the middle shelf, in front of the potatoes, there appears to be a block of cheese.

I can also see the upper edges of some items in bins or drawers labeled "Fresh Keeper," which suggests the presence of fresh produce such as fruits or vegetables. However, the actual contents are obscured, and specifics cannot be determined from the visible portions.



Based on the ingredients you've listed, you can prepare a delicious dish called: "Oven-Baked Herbed Potato Rosti with Cheese and Garlic, Served with a Fresh Fruit Salad". Here's how you can make it:

'Ingredients:\n' +
'1. Potatoes - 3 large\n' +
'2. Cheese (Preferably Cheddar) - 1 cup, grated\n' +
'3. Minced garlic - 2 tablespoons\n' +
'4. Eggs - 2\n' +
'5. Red grapes - a handful\n' +
'6. Strawberries - a handful\n' +
'7. Apples - 1\n' +
'8. Carrots - 1 (for salad freshness and crunch)\n' +
'9. Salt and pepper to taste\n' +
'10. Assumed herbs - dried oregano, dried rosemary (if available)\n'
'Instructions:\n' +
'Oven-Baked Herbed Potato Rosti\n' +
'1. Preheat your oven to 200°C (400°F). \n' +
'2. Peel and grate the potatoes, then squeeze out the extra moisture. \n' +
'3. Mix the grated potatoes with the minced garlic, grated cheese, a teaspoon each of dried oregano and rosemary if available, salt, and pepper in a bowl. Add the eggs and mix everything together until well combined.\n' +
'4. Press the potato mixture into a greased baking dish or a cast-iron skillet and bake it in the oven for about 30-40 minutes until the top is golden brown and crispy.\n' +

GPT-4o

 **Rob Whitehead**  @RJFWhite · Nov 8, 2023 ...
Just integrated @OpenAI GPT4-Vision into MML to enable interoperable metaverse agents with sight.

Of course the first thing to make was ROASTMASTER 9000 that seriously judges you for your metaverse outfit choices...

@HermanNarula you may want to close your ears



ROASTMASTER 9000

... And with increasing strength RoastMaster 9000 is going to be able to judge your outfit even more accurately. The game has grown in popularity over the last few months and we've had many requests for a more advanced version of the game. We're happy to oblige and have now added a new feature that will allow you to judge other players' outfits even more accurately. This means that if you're wearing something that's not quite up to scratch, you'll be able to see it immediately and make sure you're not embarrassed by it.

... And with increasing strength RoastMaster 9000 is going to be able to judge your outfit even more accurately. The game has grown in popularity over the last few months and we've had many requests for a more advanced version of the game. We're happy to oblige and have now added a new feature that will allow you to judge other players' outfits even more accurately. This means that if you're wearing something that's not quite up to scratch, you'll be able to see it immediately and make sure you're not embarrassed by it.

5 13 48 9K

GPT-4o + DALLE-3

 Yukiya Okuda / THE GUILD 
@alumican_net ...

『今着てる服へのファッショナドバイザーアI』

GPT-4のVision APIとDALL·E3を組み合わせて作ってみた⚡本当にサクッとできたので、これからマルチモーダルなサービスがどんどん出てきそう
Translated from Japanese by Google

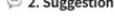
“Fashion advisor AI for the clothes you are currently wearing ”

I tried creating it by combining GPT-4's Vision API and DALL · E3 ⚡ It was really quick, so I think more and more multimodal services will come out in the future.

Was this translation accurate? Give us feedback so we can improve:  

Road to Fashionista

1. Choose a Image 

2. Suggestion 

現在の評価

- シンプルな色のシャツ
- ベージュのパンツ
- 休日のカジュアルなシーンに適したリラックスしたスタイル

アドバイス:

- カラーの調整 地味のシャツは汎用性が高いですが、もう少し色のアソシエイトを加えることで個性を出すことができます。例えば、明るい色のスカーフや、パターンのあるボタミンチーフを使うといいでしょ。
- パンツのシルエット 現在着いているパンツはリラックスフィットのようですが、もしスリムフィットを選ぶことで、より現代的なシルエットを作り出することができます。
- 履物の選択 靴が見えませんが、足元を意識することは全体のバランスを左右します。スニーカーやレザーシューズなど、シーンに合わせた選択をすると良いでしょ。
- アクセサリーの追加 時計やブレスレットなどのアクセサリーを加えることでコーディネートに奥行きを出せます。適度ではなく、さりげなく身につけるのがポイントです。
- レイヤーの追加 カジュアルなカーディガンやオーバンジャケットをレイヤーとして加えると、よりアッショウに深みが生まれます。
- 帽子のアドバイス ハットはカジュアルなシーンで最も重要なアクセサリー。多少少いサイズだと、よりスタイルリッチに見える可能性があります。

シンプルながらも、これらの変化を加えることで、より個性的なコーディネートにすることができるでしょう。

▶ 0:18 / 0:18 🔍 ⚙️ ↻ ↺ ↺ ↺

GPT-4o + Text To Speech

 **Gonzalo Espinoza Graham** 🎵 ✅
@geepytee

GPT-4V + TTS = AI Sports narrator ⚽⚽

Passed every frame of a football video to gpt-4-vision-preview, and with some simple prompting asked to generate a narration

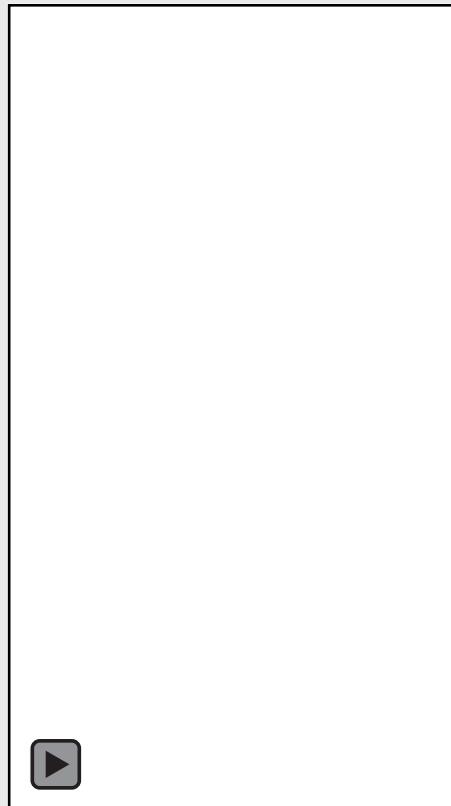
No edits, this is as it came out from the model (aka can be SO MUCH BETTER)



0:06 / 0:28 🔊 ⚽MessiMagic! ↻ ↺

HW Problem

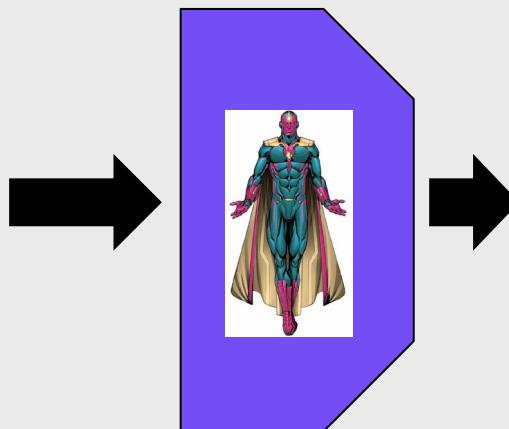
- **Make a narration of this video**
- **Bonus – clone a voice and have it narrate the video**



Richard Ogu
MAM Class of 2024

GPT-4o and Video

- We can use GPT-4o to analyze video
- Write code to extract frames from video, and feed them to GPT-4o



A man in a helmet riding a horse on green grass

Indexing Images

- We can make a folder of images searchable by indexing them with GPT-4o
- Have GPT-4o describe each image in detail, and save this to a csv file
- When you want to search for an image, load the csv file in the chat and it will find the image for you



	image_path	description
0	data/jotno/jotno crowd top.jpg	This image shows a large group of women sittin...
1	data/jotno/jotno cooking.png	The image depicts a woman and a young girl ins...
2	data/jotno/jotno nurse computer.jpeg	The image shows four individuals sitting at a ...
3	data/jotno/jotno eyes.jpeg	The image depicts a healthcare setting where a...
4	data/jotno/jotno teaching.jpeg	The image shows a group of people in a room wi...
5	data/jotno/jotno Yunus.jpeg	The image features two people standing side by...

Indexing Images

 a, image_descrip- tions.csv	 jotno cooking.png	 jotno crowd top.jpg	 jotno eyes.jpeg
 jotno kids.jpeg	 jotno logo.png	 jotno mela.jpeg	 jotno nurse computer.jpeg
 jotno table.jpeg	 jotno teaching.jpeg	 jotno Yunus.jpeg	 jotno_ad.mp4



	image_path	description
0	data/jotno/jotno crowd top.jpg	This image shows a large group of women sittin...
1	data/jotno/jotno cooking.png	The image depicts a woman and a young girl ins...
2	data/jotno/jotno nurse computer.jpeg	The image shows four individuals sitting at a ...
3	data/jotno/jotno eyes.jpeg	The image depicts a healthcare setting where a...
4	data/jotno/jotno teaching.jpeg	The image shows a group of people in a room wi...
5	data/jotno/jotno Yunus.jpeg	The image features two people standing side by...

GPT-4o Pricing

- GPT-4o charges by the token
- 512 x 512 image
 - 85 tokens at low resolution
 - 255 tokens at high resolution
- GPT-4o-mini: \$1.50 per 1 million input tokens
 - Low res: costs \$0.00085 -> \$0.01 per 1000 images
 - High res: costs \$0.00255 -> \$0.04 per 1000 images

Coding Session

- **Describe images with GPT-4o**
- **Find themes in image clusters**
- **Write a script from an Instagram reel video**
- **Create a narrated slide show from a set of images**
- **Try to learn about someone's personal business from their picture**