

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

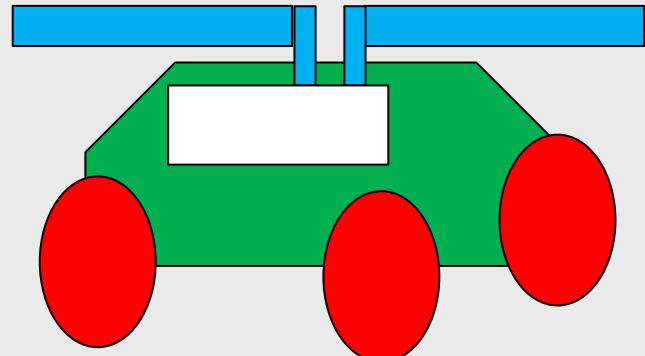
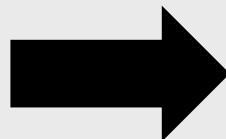


Image Features

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



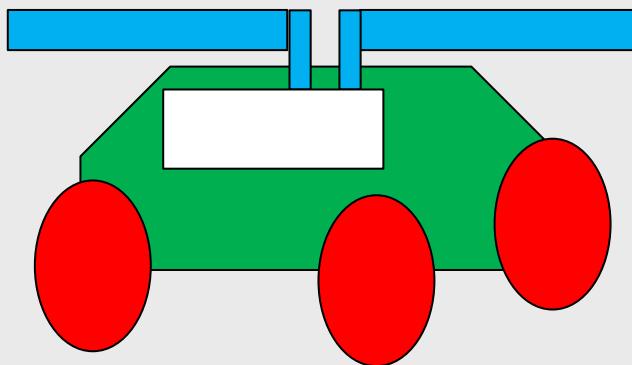
Car



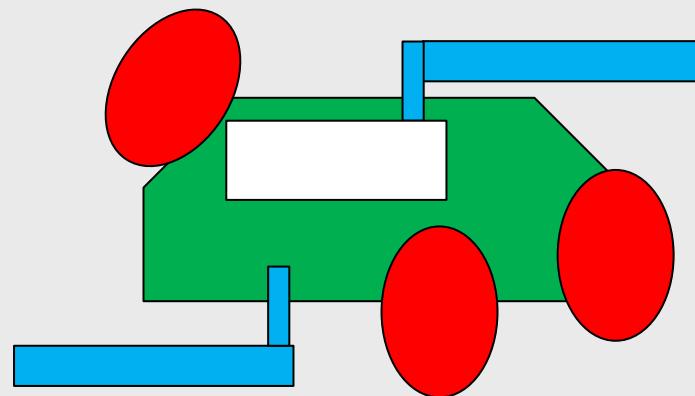
Car

Image Features

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



Car



Not a Car

Learning Image Features

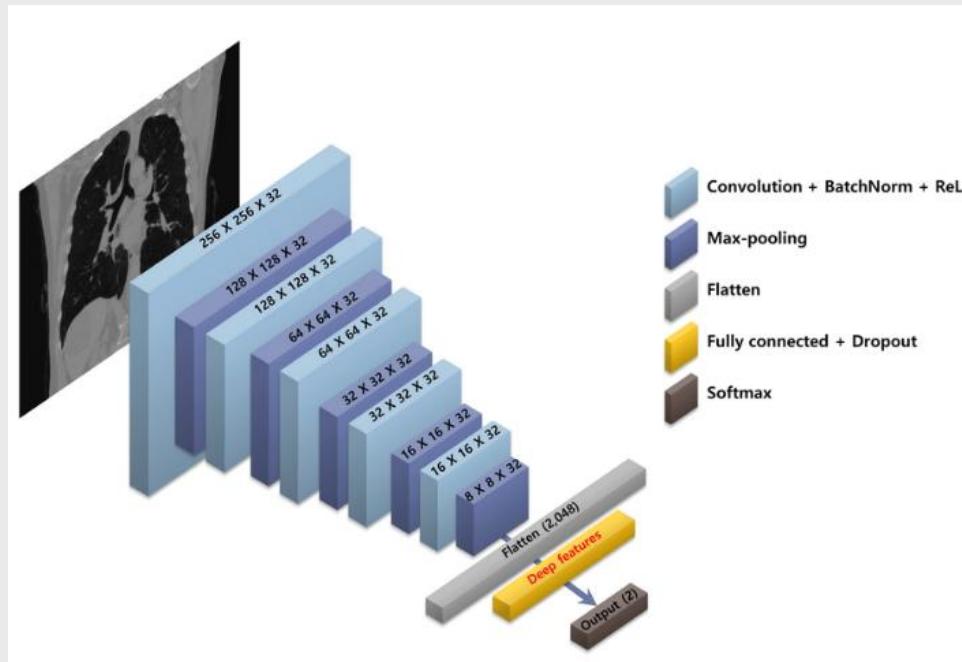
- **Image features**
 - Edges
 - Shapes built from edges
 - Combinations of shapes
 - Deeper hierarchy of combinations of shapes
 - Combinations of combinations of ... shapes
 - Translation invariant
 - Doesn't matter where in the image the object is
- **We need to learn what these features and their 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
- Convolutions are translation invariant
 - Is the pattern in the image? Don't care where

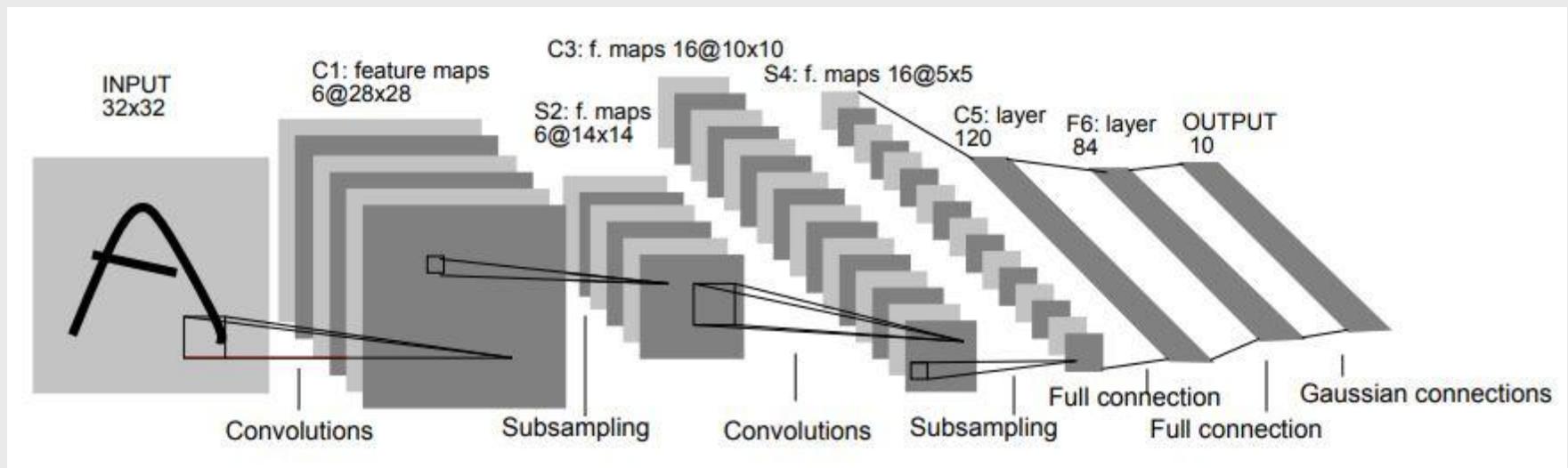
Convolutional Network

- 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

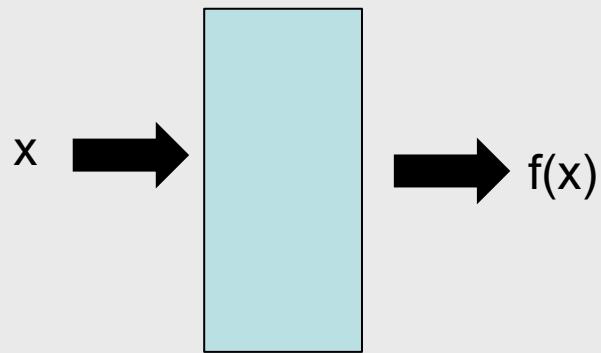
- People found that deep convolutional networks performed better
- People also found that deeper networks were harder to train
 - Deep here meant 10 – 20 layers

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!

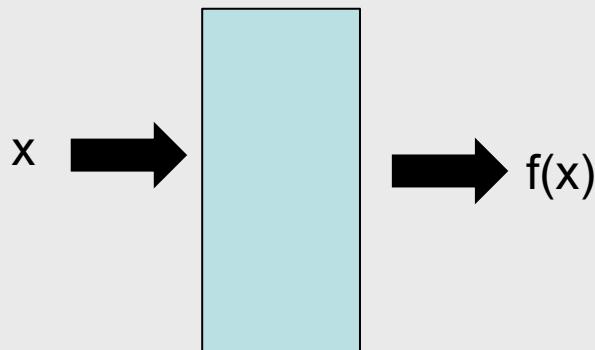
ResNet

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

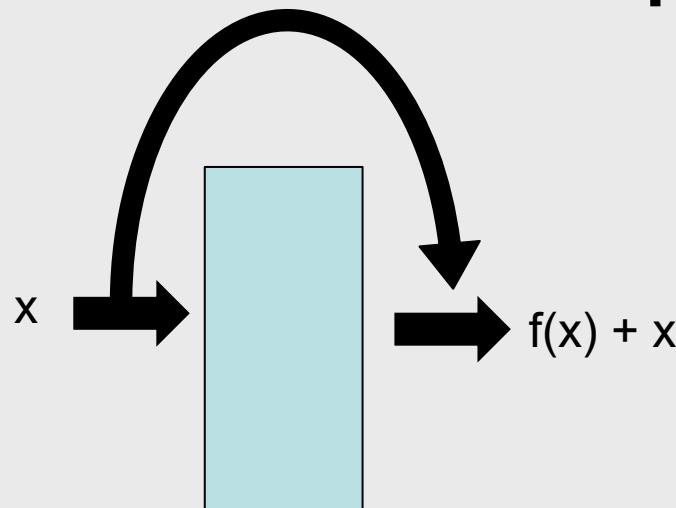


ResNet

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

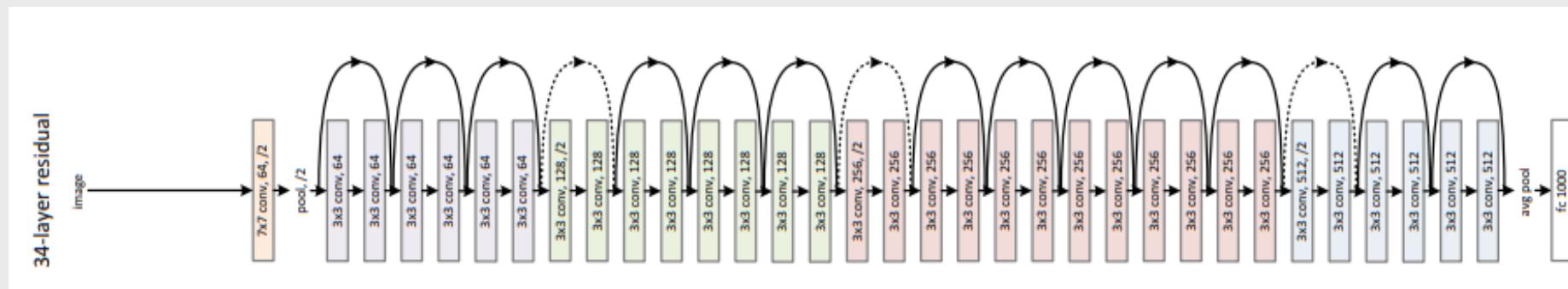


- ResNet layer takes in x and outputs $f(x) + x$



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



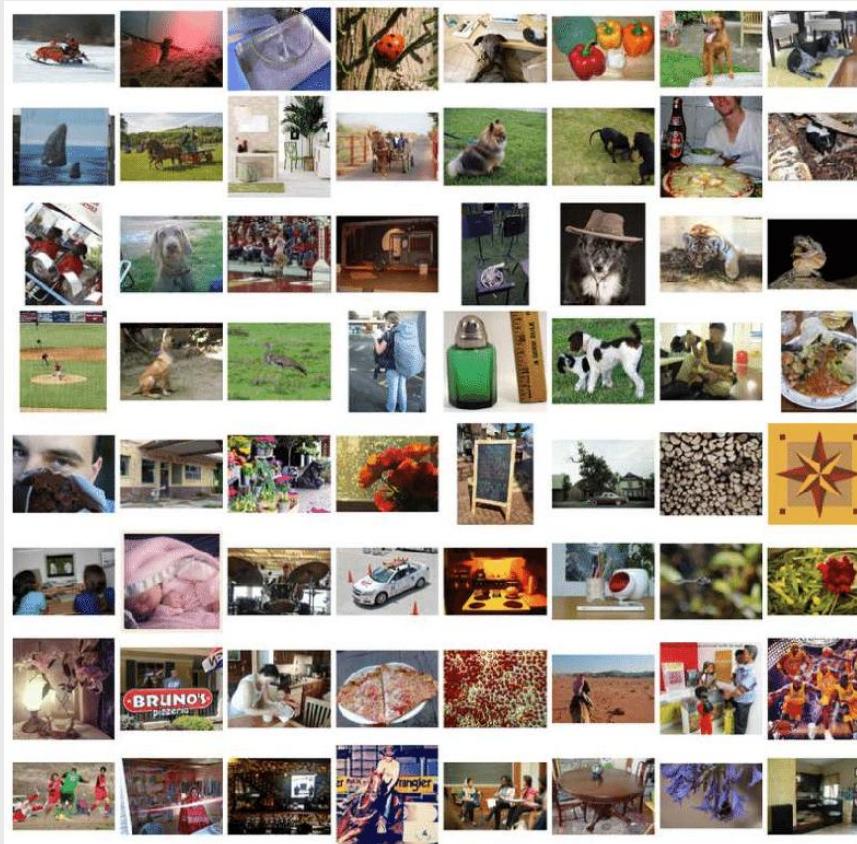
ImageNet

- **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

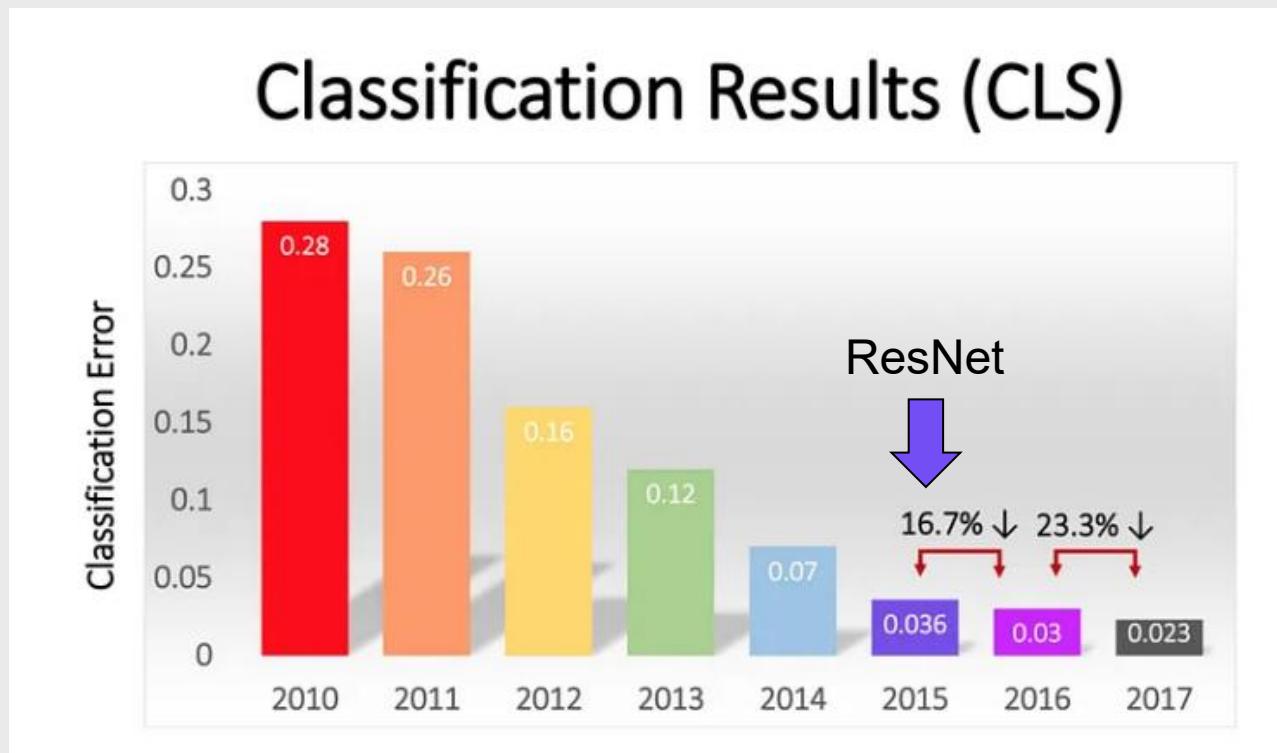
ImageNet Challenge

- 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



Transformers



Vision Transformers



Vision Transformers

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

**Alexey Dosovitskiy^{*,†}, Lucas Beyer^{*}, Alexander Kolesnikov^{*}, Dirk Weissenborn^{*},
Xiaohua Zhai^{*}, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby^{*,†}**

*equal technical contribution, †equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

Vision Transformers

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*,
Xiaohua Zhai* Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†

*equal technical contribution, †equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

EXCLUSIVE

ARTIFICIAL INTELLIGENCE

[Follow](#)

Meta Poaches Three OpenAI Researchers

Social-media giant has hired Lucas Beyer, Alexander Kolesnikov and Xiaohua Zhai for its superintelligence effort

By [Meghan Bobrowsky](#) [Follow](#)

June 25, 2025 10:28 pm ET



Lucas Beyer (bl16) ✅
@giffmana · [Follow](#)

hey all, couple quick notes:

- 1) yes, we will be joining Meta.
- 2) no, we did not get 100M sign-on, that's fake news.

Excited about what's ahead though, will share more in due time!

cc @_kolesnikov_ and @XiaohuaZhai.

1:15 AM · Jun 27, 2025



4.1K



[Reply](#)

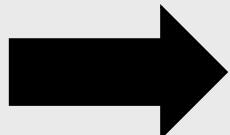
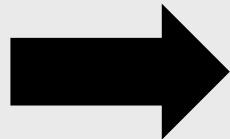


[Copy link](#)

[Read 386 replies](#)

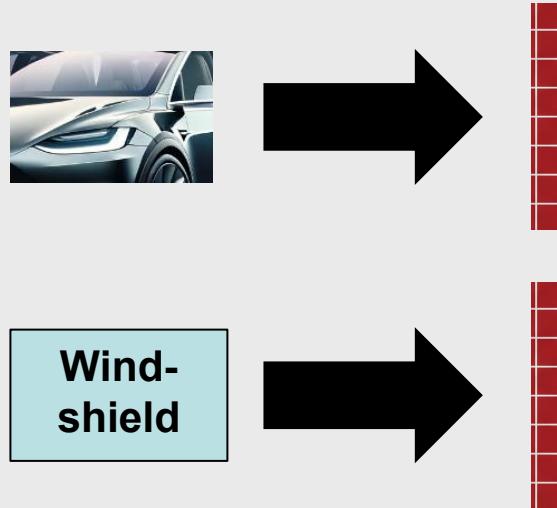
Vision Transformer Image Features

- Image features are square patches which are converted into numerical vectors

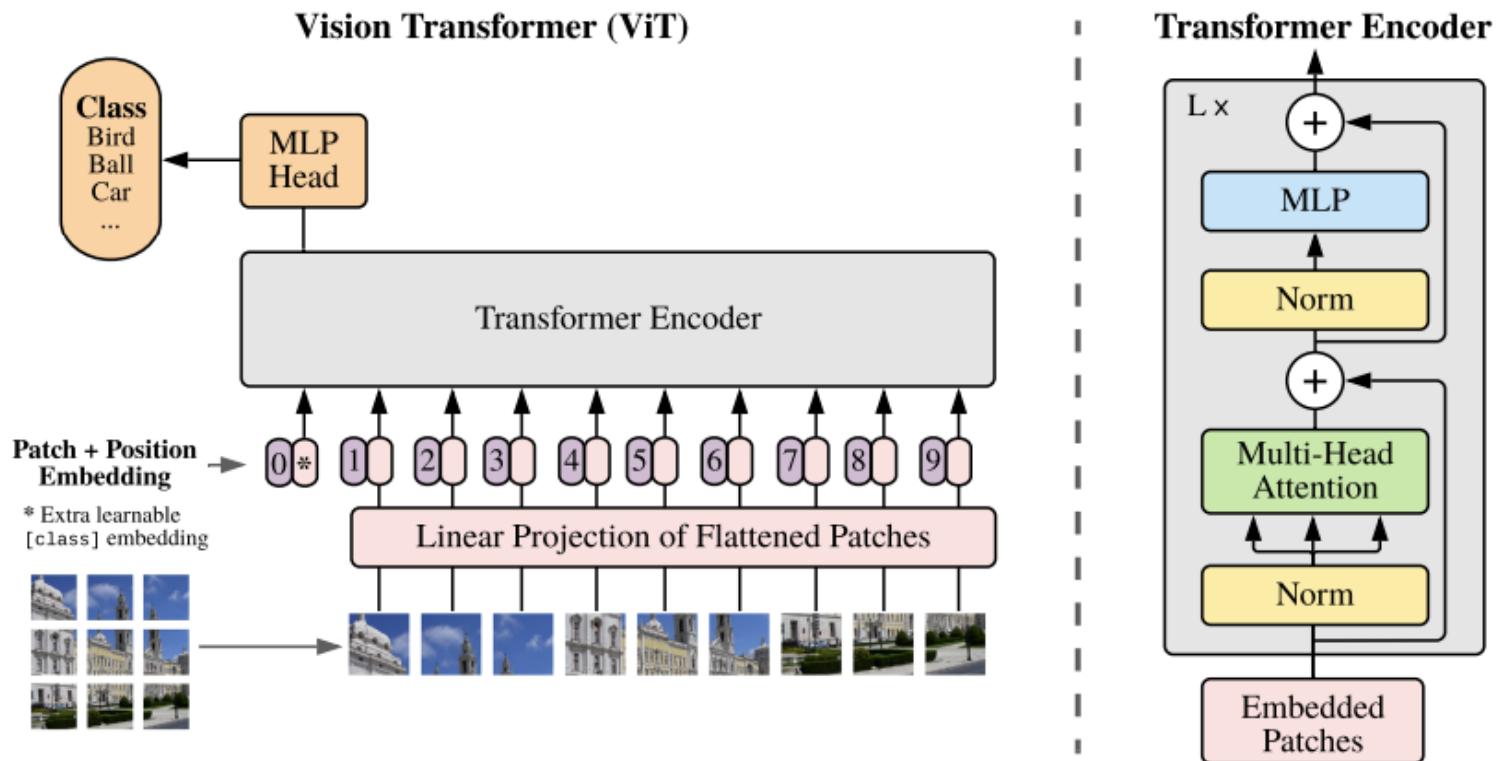


Text = Image

- **Transformers take vectors as inputs**
- **The input vector can come from an image or text**
- **This allows a transformer to handle images and text on the same footing**



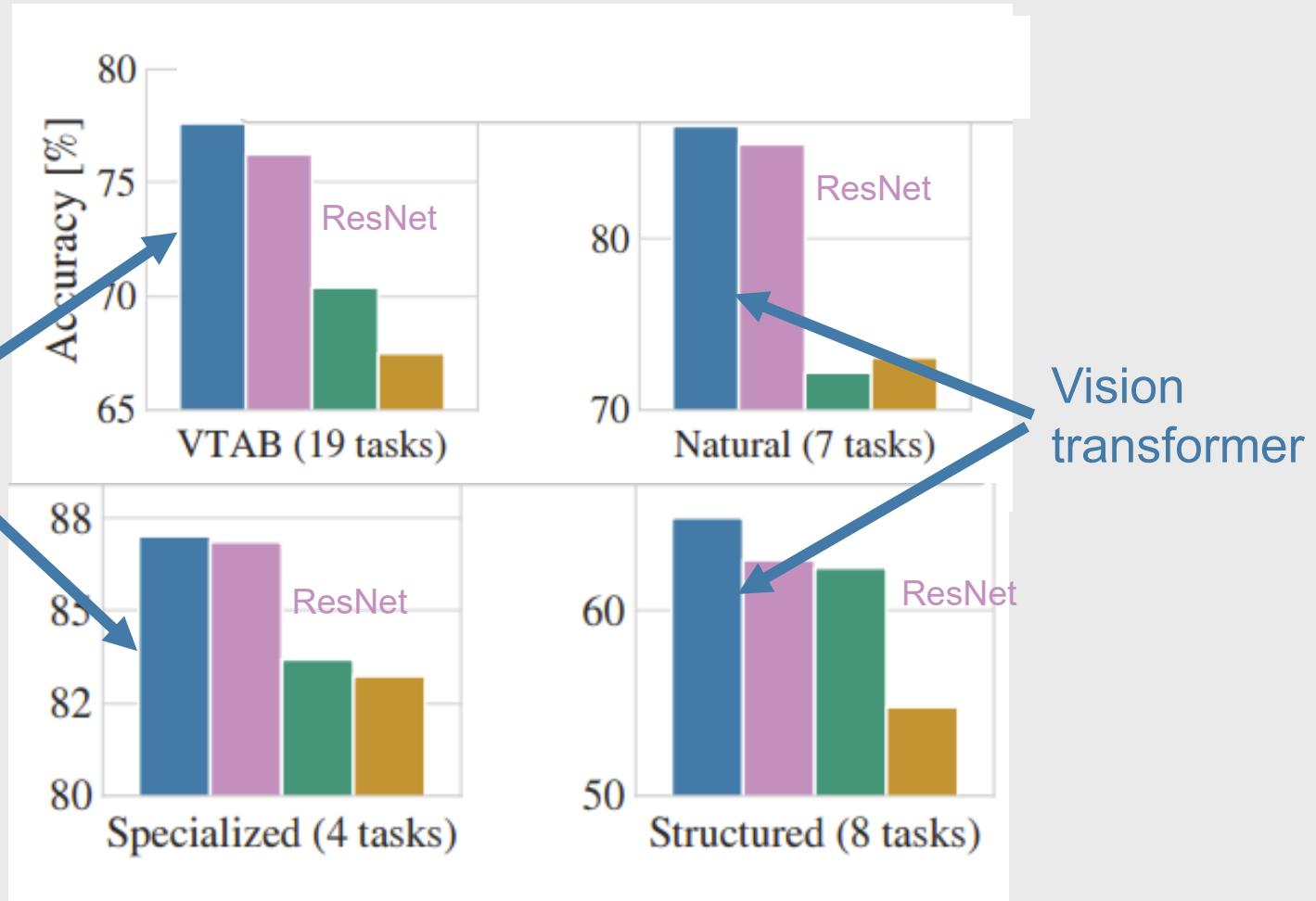
Vision Transformer Architecture



Vision Transformer Performance

Vision
transformer

Vision
transformer



ViT-H/14

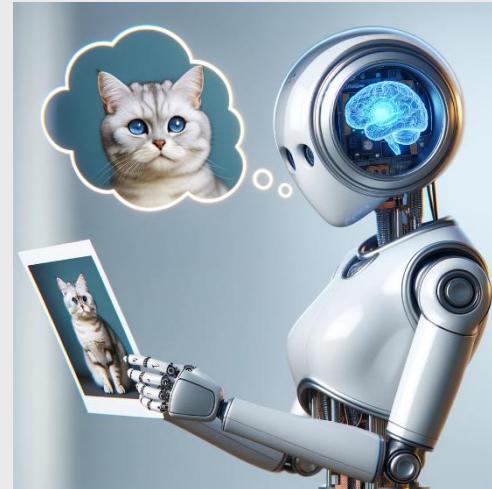
BiT-L (R152x4)

VIVI-Ex-100% (R50x3)

S4L (R50x1)

Mapping Image to Text

- ResNet and Vision Transformers are good for image classification
- What if we want to write text that describes an image?



Predictive Learning

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



Dog

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**
 - Multiple choice question
 - Lets one learn to contrast categories
 - Useful if categories are very narrow
- **Learning Question: Which image is A cute dog with a droopy face in front of a pink wall?**



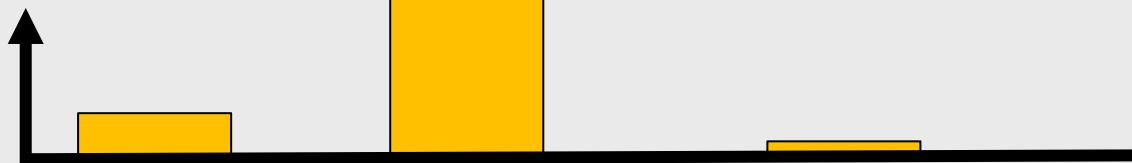
No



Yes



Hell no

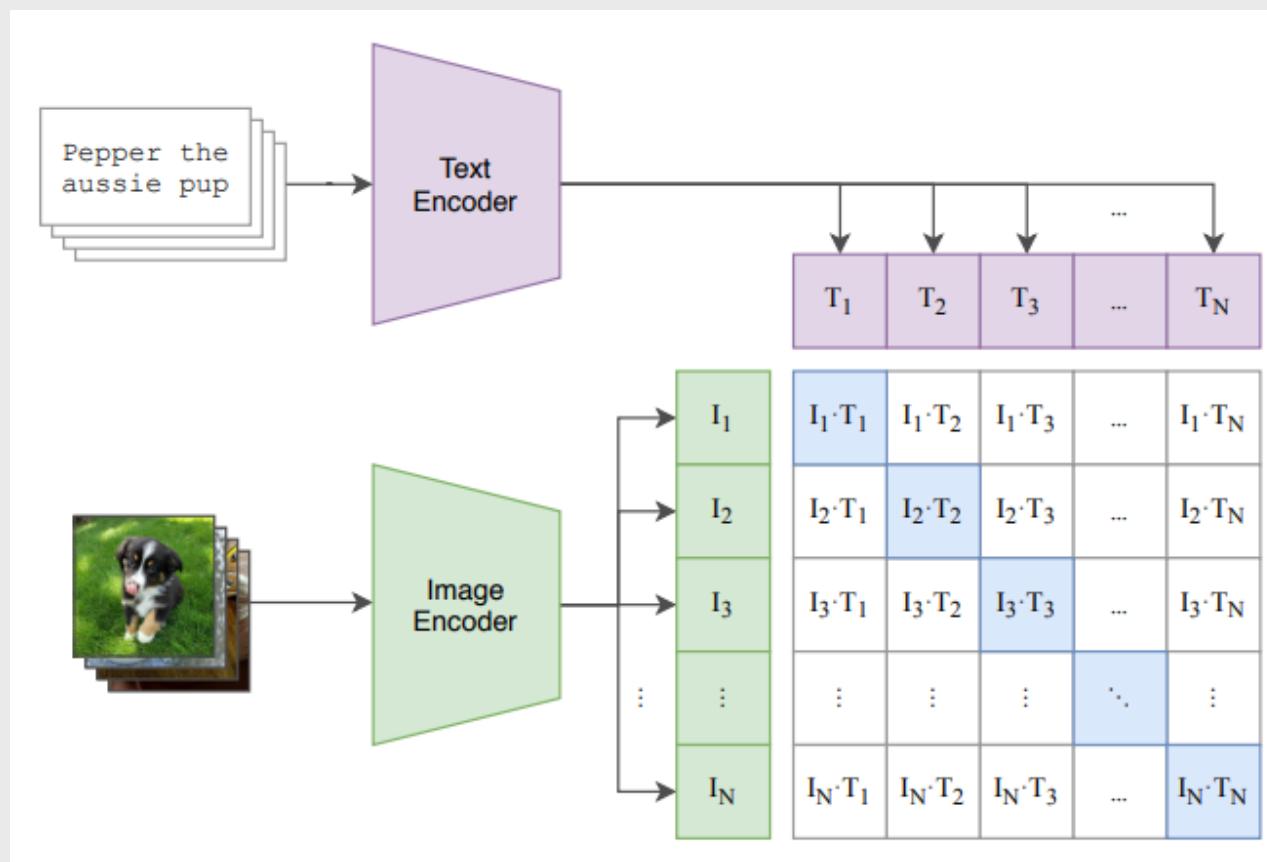


CLIP

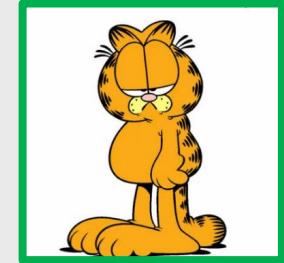
- Contrastive Language-Image Pre-Training (CLIP) is a model developed by OpenAI in 2021 that maps images to text
- Combines **text transformers** with **convolutional nets** or **vision transformers**
- Allows an AI to describe an image in words
- Key idea: **Contrastive Learning**

CLIP Training

- **Text encoder (transformer) and image encoder (vision transformer) 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.

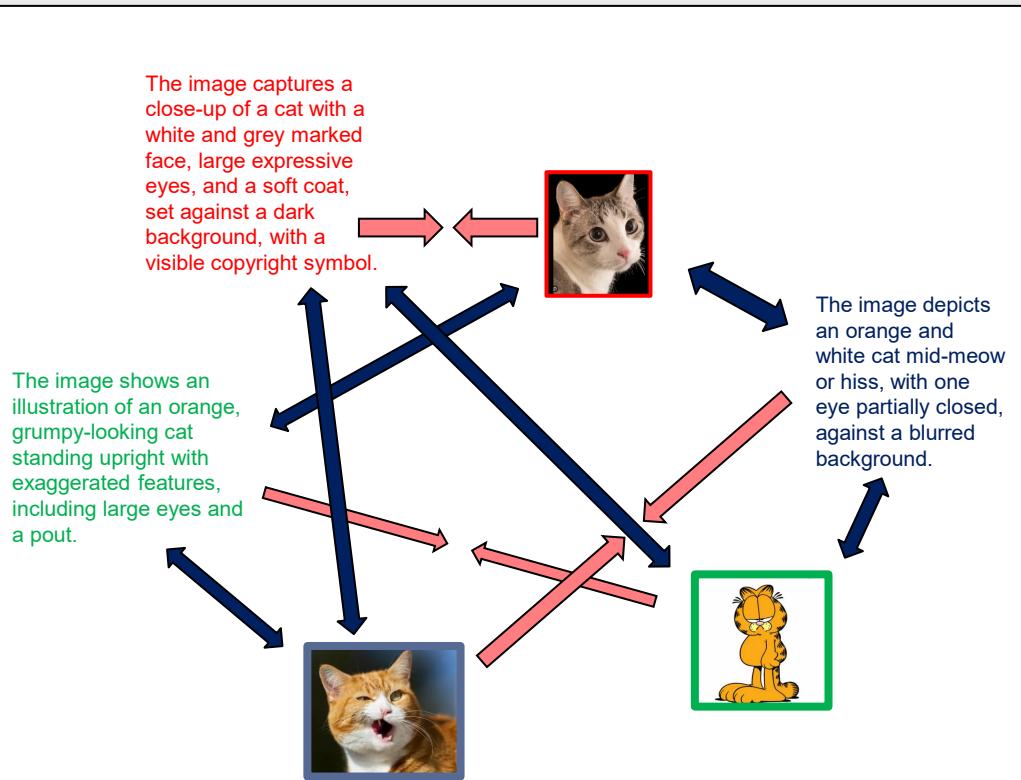
The image depicts an orange and white cat mid-meow or hiss, with one eye partially closed, against a blurred background.

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 Dynamics

- Training with the CLIP objective involves pushing and pulling
- Pull together matching pairs
- Push apart non-matching pairs



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

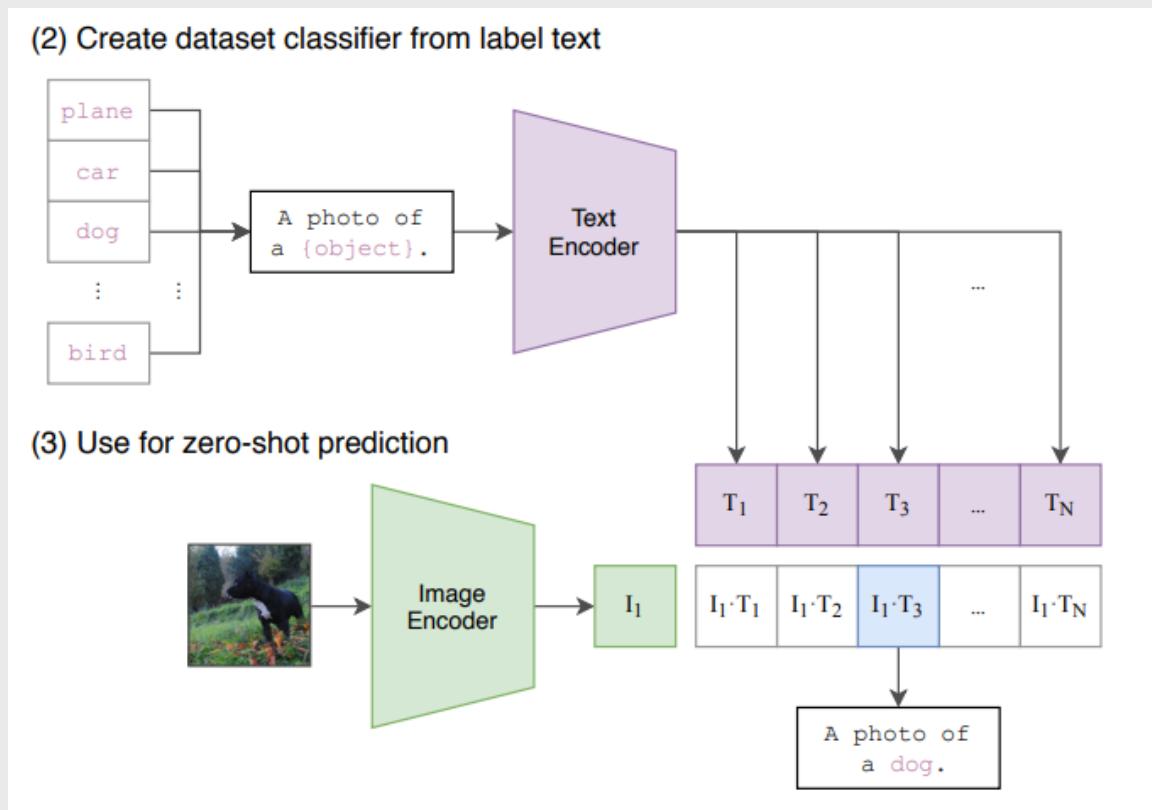


The image depicts an orange and white cat mid-meow or hiss, with one eye partially closed, against a blurred background.



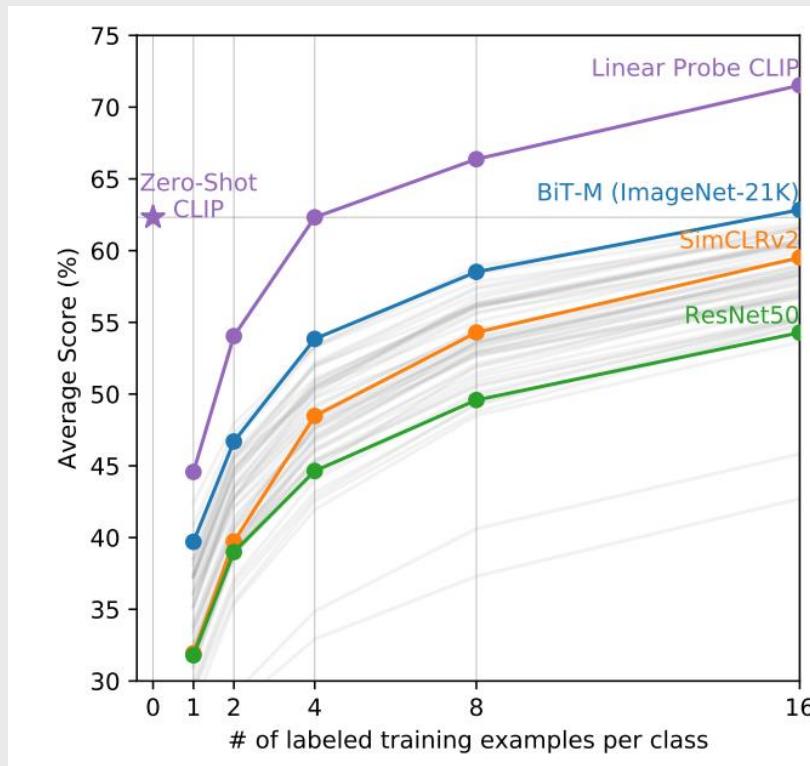
CLIP Prediction

- 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





Multimodal AI Models

- Original multimodal model was OpenAI's GPT-4V (Vision)

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.

Multimodal AI Model API

- To use multimodal models with the API, we simply give it an image and text input
- Image processing – we will need to encode the image to be in the right format (AI does this for us)

```
def encode_image(image_path):
    """Encodes a local image file to a base64 string."""
    with open(image_path, "rb") as image_file:
        return base64.b64encode(image_file.read()).decode('utf-8')
```

Input Image Count

- GPT 5.2 can take approximately 100 input images
- We just need to make a list of the processed image files and give this list to the API

content_payload
is the list of
processed
images



```
# List of pizza images to analyze
pizza_files = ["images/pizza1.png", "images/pizza2.png", "images/pizza3.png"]

for img_path in pizza_files:
    b64 = encode_image(img_path)
    content_payload.append({
        "type": "input_image",
        "image_url": f"data:image/png;base64,{b64}"
    })

# Call GPT-5.2 with medium reasoning for accuracy
response = client.responses.create(
    model="gpt-5.2",
    input=[{"role": "user", "content": content_payload}],
    reasoning={"effort": "medium"}
)
```

OpenAI API Pricing

- **Multimodal AI models are cheap 😊**
- **100 images with GPT-5.2 is ~ \$0.56**

GPT-5 mini		
Item	Unit	Cost (\$)
Input token	1000 tokens	0.00025
Output token	1000 tokens	0.00200
Image	3200 tokens	0.00080

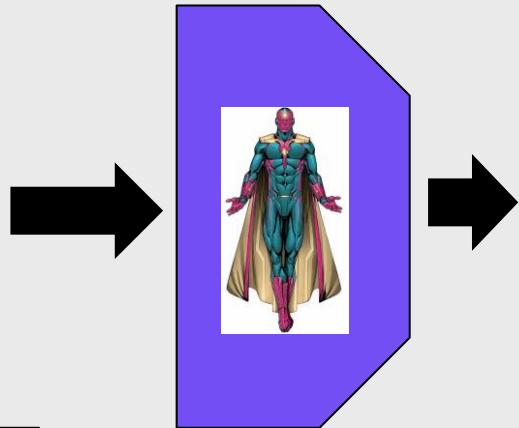
GPT-5.2		
Item	Unit	Cost (\$)
Input token	1000 tokens	0.00175
Output token	1000 tokens	0.01400
Image	3200 tokens	0.00560

GPT-5.2 pro		
Item	Unit	Cost (\$)
Input token	1000 tokens	0.02100
Output token	1000 tokens	0.16800
Image	3200 tokens	0.06720

What's in My Fridge?



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.

Recipe Generator



please list each food item in this image.

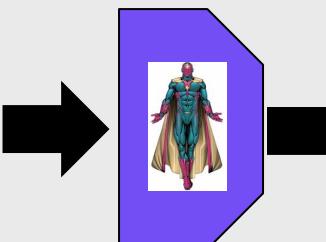


make me a recipe with the following ingredients from my fridge:

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:
'1. Potatoes - 3 large
'2. Cheese (Preferably Cheddar) - 1 cup, grated
'3. Minced garlic - 2 tablespoons
'4. Eggs - 2
'5. Red grapes - a handful
'6. Strawberries - a handful
'7. Apples - 1
'8. Carrots - 1 (for salad freshness and crunch)
'9. Salt and pepper to taste
'10. Assumed herbs - dried oregano, dried rosemary (if available)
'Instructions:
'Oven-Baked Herbed Potato Rosti
'1. Preheat your oven to 200°C (400°F).
'
'2. Peel and grate the potatoes, then squeeze out the extra moisture.
'
'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.
'
'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.'

Image Caption Generator



Uploaded Car

Caption

Redefine your journey amidst the blending hues of luxury and thrill. Unveil the road story with every turn.



Hashtags

#FavouriteFourRings #audispot254 #RoadRedefine #SleekRed #LuxuryThrill

🎵 TikTok Sound Ideas

- "Top Gear Thrill Ride" – A high-speed, adrenaline-laced soundtrack for fast-paced rides.

Game Dialogue Generator

 **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

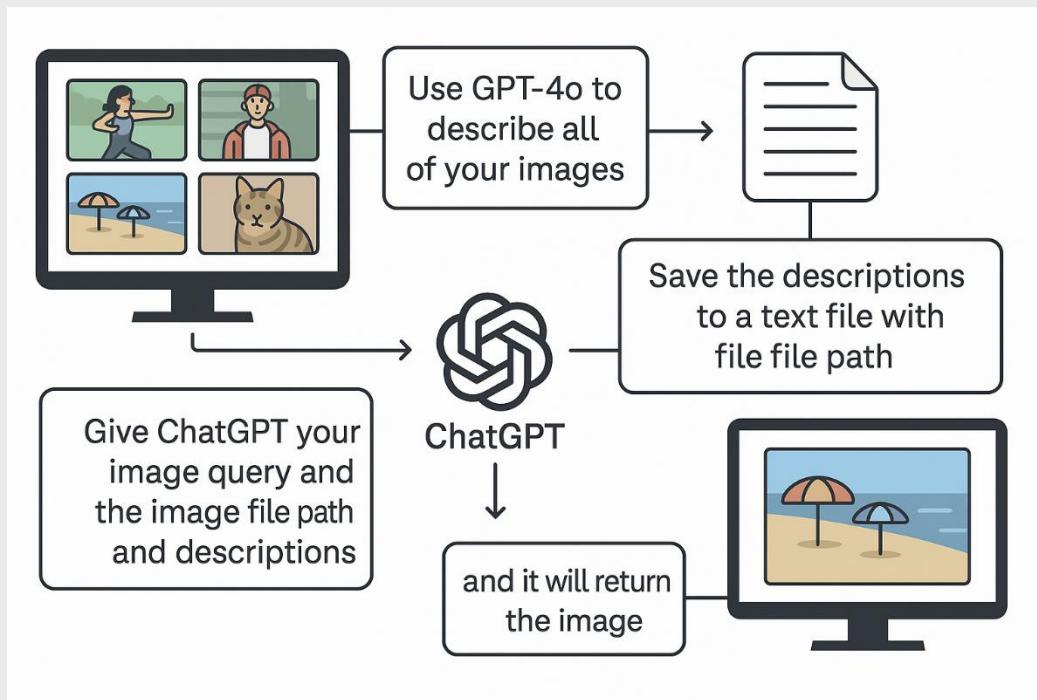


5 13 48 9K

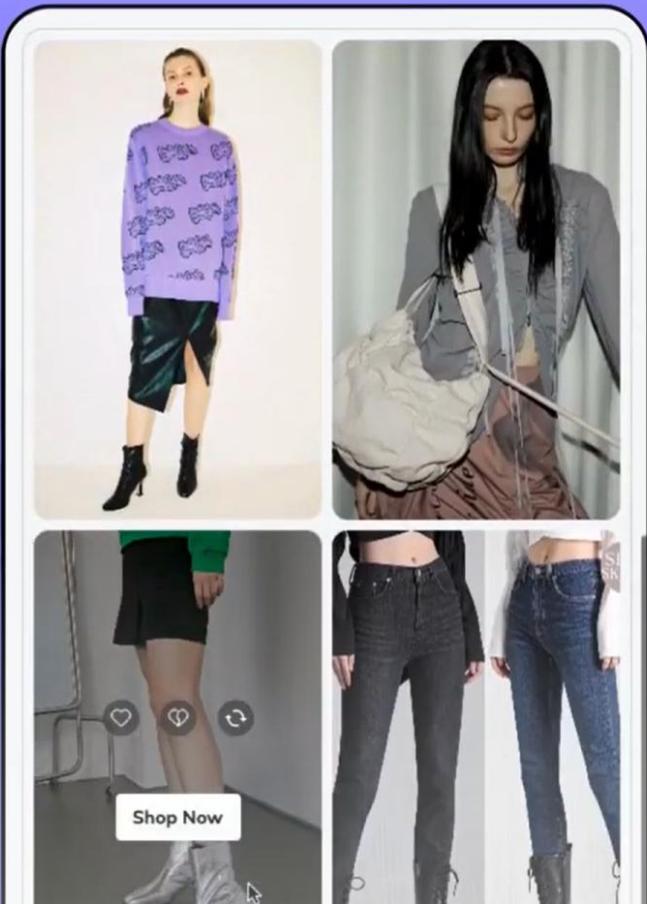
 

Searchable Image Database

- Use AI to describe all your images
- Save the description to a text file with the file path
- Give AI your image query and the image file paths and descriptions and it will return the image



AI Fashion Stylist



AI STYLIST

**“Show me the fall
trendy looks.”**

yesplz.ai

Checking Inventory

count how many bags
of ice in this image



Checking Inventory



count how many bags of ice in this image

Based on the images provided, here is a count of the bags of ice in the scene:

Total Count

There are **16** bags of ice visible in the image.

Breakdown by Location

- On the Cart: There is a stack of 8 bags being transported on the blue hand truck/cart.
- On the Floor: There are 7 bags scattered on the floor near the entrance of the ice storage unit.
- In Transit: One worker is seen carrying 1 bag from the storage unit toward the cart.

Multimodal AI App Complexity

- **Level 1: Perception – What is this?**
 - e.g. What's in my fridge?
- **Level 2: Reasoning – Why does this matter?**
 - e.g. Give me a recipe I can make with what's in my fridge
- **Level 3: Structure – Turn this into data**
 - e.g. Convert the fridge items into a spreadsheet
- **Level 4: Agency – Do something about it**
 - e.g. AI sees I'm running low on eggs and places an order



Coding Session

- In our coding session we will build an image captioning web app using AI and Streamlit
- We will add in some other cool features

